Package 'DOAGDC'

May 2, 2019

Title A Package to Download, Organize and Analyze Genomic Data Commons (GDC) Data

Version 0.99.0

Date 2019-05-01

Description

In order to download, organize and analyze GDC data we have developed DOAGDC. It implements and uses many well-
known R packages like DESeq2, edgeR, WGCNA, and mclust.
This is a free open-source software that implements academic research
by the authors and co-workers. If you use it, please support
the project by citing the appropriate journal article listed in
citation(``DOAGDC'').

Depends R (¿= 3.4)

Imports stats,
utils,
RCurl,
curl,
R.utils,
readr,
jsonlite,
stringr,
stringi,
tools,
httr,
data.table,
mclust,
MineICA,
RColorBrewer,
devtools,
methods,
pheatmap,
ggplot2,
gplots,
EBSeq,
edgeR,
DESeq2,
gage,
BiocParallel,
IHW,
VennDiagram,

dendextend,
amap,
GO.db,
annotate,
org.Hs.eg.db,
DO.db,
reactome.db,
goseq,
clusterProfiler,
DOSE,
ReactomePA,
igraph,
pathview,
mgcv,
XML,
SummarizedExperiment,
xlsx,
survival,
ggthemes,
grid,
reshape,
extrafont,
scales,
cgdsr,
AnnotationDbi,
dynamicTreeCut,
forcats,
reshape2,
dplyr

Suggests rmarkdown (¿= 0.9),
knitr (¿= 1.12),
WGCNA,
yarrr,
survminer,
ggbiplot,
gageData,
testthat

License GPL-3 — file LICENSE

URL https://github.com/Facottons/DOAGDC

BugReports https://github.com/Facottons/DOAGDC/issues

Encoding UTF-8

LazyData true

VignetteBuilder knitr

RoxygenNote 6.1.1

Roxygen list(markdown = TRUE)

R topics documented:

`annotation_table` *UCSC refseq table*

Description

This dataset was exported from UCSC Genome Browser available from `http://research.stlouisfed.org/fred2`.

Usage

```
annotation_table
```

Format

A data frame with 16929 rows and 6 variables

References

`http://www.stat.berkeley.edu/~brill/Papers/jspifinal.pdf`

beta2mValues *Convert beta values to mvalues*

Description

Convert beta values to mvalues

Usage

```
beta2mValues(probeName, env, saveData = FALSE)
```

Arguments

probeName	A character string containing the probe name desired from the selected 'Name' probes.
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).
saveData	Logical value where TRUE indicates that the concatenate and filtered matrix should be saved in local storage. The default is FALSE.

Examples

```
## Not run:
beta2mValues(probeName = "cg16672562", saveData = TRUE, env = "env name without quotes")

## End(Not run)
```

clinical_terms *Search for clinical terms for all available tumors*

Description

Search for clinical terms for all available tumors

Usage

```
clinical_terms(Name, workDir = "~/Desktop", tumor = "all", dataType,
  group.number = 2, term, dataBase = "legacy", term_keyword = NULL,
  tumor_with_term = 1, p_cutoff = 0.05, FDR_cutoff = 0.05,
  confidence.level = 0.95, Width = 7, Height = 7, Res = 300,
  Unit = "in", image_format = "svg", env, cexAxixX = 16,
  cexAxixY = 16)
```

Arguments

`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`workDir`	A character string specifying the path to work directory.
`tumor`	A character vector indicating which tumors should be used in the analysis. The default is "all".
`dataType`	A character string indicating which dataType will be used to group the patients. Only needed when the environment was not created yet and the user does not have the `term` argument.
`group.number`	Numerical value indicating how many groups should be generated. This argument is advisable if subcategory number is larger than 5. The default is 2.
`term`	A character string containing the clinical term to be used.
`dataBase`	A character string specifying `"GDC"` for GDC Data Portal or `"legacy"` for GDC Legacy Archive.
`term_keyword`	A character string containing a possible fragment of the term. Used when the specific term is unknown.
`tumor_with_term`	A numerical value indicating the minimum number of tumors containing a specific term with the keyword used. The default is 1.
`p_cutoff`	Numerical value indicating the maximum value for p-values. The default is `0.05`.
`FDR_cutoff`	Numerical value indicating the maximum value for FDR values. The default is `0.05`.
`confidence.level`	A numerical value containing the confidence interval to be used. The default is 0.95.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

`cexAxixX, cexAxixY`
: A numerical value giving the amount by which labels in X and Y axis respectively should be modified relative to the default size. The default value is 16 for both arguments.

Examples

```
## Not run:
#searching for terms having the keyword inserted in common 'tumor_with_term = 3'
clinical_terms(tumor = c("BRCA", "UCS", "OV"), term_keyword = "year", tumor_with_term = 3)

#searching for all terms available with at least three tumors in common 'tumor_with_term = 3'
clinical_terms(tumor = c("BRCA", "UCS", "OV"), term_keyword = "", tumor_with_term = 3)

#using the analysis with a specified term
clinical_terms("UCS", "isoform", term = "vital_status")

## End(Not run)
```

`concatenate_files` *Concatenate GDC files into a single matrix and prepar the data*

Description

`concatenate_files` is a function designed to concatenate GDC files into a single matrix, where the columns stand for patients code and rows stand for data names.

Usage

```
concatenate_files(dataType, normalization = TRUE, Name, dataBase,
  HTSeq = NULL, workDir = "~/Desktop", tumor, workflowType,
  tumorData = TRUE, cutoffBetaNA = 0.25, cutoffBetasd = 0.005,
  onlyFilter = FALSE, Platform = "", use_hg19_mirbase20 = FALSE, env,
  saveData = FALSE)
```

Arguments

`dataType`
: Type of data. It could be `"methylation", "mutation", "clinical_supplement", "bios`

- Only present in "Legacy" database:`"protein", "Exon quantification", "miRNA gen quantification", "miRNA isoform quantification", "isoform", "image", "cl`
- Only present in "GDC" database:`"miRNA Expression Quantification", "Isoform`

`normalization`
: Logical value where `TRUE` specify the desire to work with normalized files only. When FALSE, in the second run, do not forget to set env argument. This argument is only applyable to gene and isoform expression data from GDC Legacy Archive. The default is `TRUE`.

`Name`
: A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.

`dataBase`
: A character string specifying `"GDC"` for GDC Data Portal or `"legacy"` for GDC Legacy Archive.

`HTSeq` A character string indicating which HTSeq workflow data should be downloaded (only applied to "GDC" gene expression): "Counts", "FPKM", or "FPKM-UQ".

`workDir` A character string specifying the path to work directory.

`tumor` A character string contaning one of the 33 tumors available in the TCGA project. For instance, the `"BRCA"` stands for breast cancer.

`workflowType` A character string specifying the workflow type for mutation data in "gdc". Where:

- "varscan" stands for VarScan2 Variant Aggregation and Masking
- "mutect" stands for MuTect2 Variant Aggregation and Masking
- "muse" stands for MuSE Variant Aggregation and Masking
- "somaticsniper" stands for SomaticSniper Variant Aggregation and Masking
- "all" means to concatenate all workflows into a single matrix.

`tumorData` Logical value where `TRUE` specifies the desire to work with tumor tissue files only. When set to FALSE, it creates two matrices, one containing tumor data and other containing data from not-tumor tissue. The default is `TRUE`.

`cutoffBetaNA` Numerical value indicating the maximum threshold percentage (in decimal form) to tolerate and to remove rows containing NA for beta values (methylation data). The default is 0.25.

`cutoffBetasd` Numerical value indicating the standard deviation threshold of beta values (methylation data). It keeps only rows that have standard deviation of beta values higher than the threshold. The default is `0.005`.

`onlyFilter` Logical value where `TRUE` indicates that the matrix is already concatenate and the function should choose a different `Name`, without concatenate all the files again. The default is FALSE.

`Platform` A character string indicating the platform name for methylation, exon quantificaton, miRNA, and mutation data.

- For mutation and exon quantificaton data:`"Illumina GA", "Illumina HiSeq" or "al`
- For methylation data`"Illumina Human Methylation 450", "Illumina Human Meth`
- For miRNA data:`"Illumina GA", "Illumina HiSeq", "H-miRNA_8x15K" (for GBM` `"all"`.

The default for all dataType cited is `"all"` (when downloading data).

`use_hg19_mirbase20`
Logical value where `TRUE` indicates that only hg19.mirbase20 should be used. This parameter is needed when using `dataBase = "legacy"` and one of the available miRNA `dataType` in "legacy" ("miRNA gene quantification" and "miRNA isoform quantification"). The default is FALSE.

`env` A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

`saveData` Logical value where `TRUE` indicates that the concatenate and filtered matrix should be saved in local storage. The default is FALSE.

Value

A matrix with data names in row and patients code in column.

Examples

```
## Not run:
# Concatenating isoform expression data into a single matrix
concatenate_files(dataType = "isoform", Name = "uc002peh.2", dataBase = "legacy", tumor = "BRCA")

## End(Not run)
```

---

co_expression *Co-Expression Analyses*

Description

Powered by WGCNA

Usage

```
co_expression(Data = NULL, Name, dataBase, workDir, tumor,
  normalization = TRUE, tumorData = TRUE, traitData = NULL,
  networkType = "unsigned", minModuleSize = 15, max_softpower = 20,
  nthreads = 1, MEDissThres = 0.25, Width = 2000, Height = 1500,
  Res = 300, Unit = "px", image_format = "png", env,
  saveCheckpoints = FALSE, loadCheckpoint = NULL,
  pearsonCutoff = 0.5)
```

Arguments

`Data`	Used for external non-log expression data. Matrix or data frame. This `Data` must have patients/sample code as `colnames` and genes as `rownames`.
`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`dataBase`	A character string specifying `"GDC"` for GDC Data Portal or `"legacy"` for GDC Legacy Archive.
`workDir`	A character string specifying the path to work directory.
`tumor`	A character string contaning one of the 33 tumors available in the TCGA project. For instance, the `"BRCA"` stands for breast cancer.
`normalization`	Logical value where `TRUE` specify the desire to work with normalized files only. When FALSE, in the second run, do not forget to set env argument. This argument is only applyable to gene and isoform expression data from GDC Legacy Archive. The default is `TRUE`.
`tumorData`	Logical value where `TRUE` specifies the desire to work with tumor tissue files only. When set to FALSE, it creates two matrices, one containing tumor data and other containing data from not-tumor tissue. The default is `TRUE`.

`traitData` A character string where "default" indicates that the trait data to be used is provide by the output of `check_clinical_terms` function. Use "custom" for data inserted by the user after the `table2DOAGDC` function and with the table object named as `trait_data`. This object must have patients/sample code as `rownames` and the trait categories as `colnames`. By default trait data is not used.

`networkType` A character string indicating which network type should be used. WGCNA allows: "unsigned", "signed", and "signed hybrid". The default is "unsigned".

`minModuleSize` Numerical value specifying the minimum cluster size. The default is 15.

`max_softpower` Numerical value indicating the maximum soft thresholding power for which the scale free topology fit indices are to be calculated. The default is 20.

`nthreads` Numerical value indicating how many threads to allow. The number of threads should not be more than the number of actual processors/cores. The default is 1.

`MEDissThres` Numerical value specifying the maximum dendrogram cut height for module merging qualified by dissimilarity (i.e., 1-correlation). The default is 0.25.

`Width` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Height` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Res` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Unit` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`image_format` A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".

`env` A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

`saveCheckpoints` Logical value where TRUE indicates that an external representation of analysis' objects will be saved to file in disk. The default is FALSE.

`loadCheckpoint` Logical value where TRUE indicates that the saved checkpoint will be loaded and the analysis is going to continue from that point. The default is FALSE.

`pearsonCutoff` Numerical value specifying the minimum Pearson correlation value. The default is 0.5.

Examples

```
## Not run:
co_expression(Data, workDir, tumor, env = "env name without quotes")

## End(Not run)
```

`cross_co_expression` *Cross co-expression gene list against differential expression genes list*

Description

Cross co-expression gene list against differential expression genes list

Usage

```
cross_co_expression(pairName = "G2_over_G1", Width = 2000,
  Height = 2000, Res = 300, Unit = "px", Colors = c("green",
  "blue", "red"), VennDiagram_imagetype = "png", workDir, env)
```

Arguments

`pairName`	A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Colors`	A character vector indicating the colors to be used in the venn diagram. The default is c('green', 'blue', "red").
`VennDiagram_imagetype`	A character string indicating the image_format (e.g. "tiff", "png" or "svg"). The default is "png".
`workDir`	A character string specifying the path to work directory.
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Examples

```
## Not run:
cross_co_expression(pairName = "G2_over_G1", env = "env name without quotes")

## End(Not run)
```

`dea_DESeq2` *Run DESeq2 gene Differential Expression Analysis (DEA).*

Description

Run DESeq2 gene Differential Expression Analysis (DEA).

Usage

```
dea_DESeq2(Name, coreNumber = 2, test = "Default Test", groupGen,
  clinical_pair, FC_cutoff = 2, workDir, tumor, FDR_cutoff = 0.05,
  Width = 2000, Height = 1500, Res = 300, Unit = "px",
  image_format = "png", env, cooksCutoff = FALSE)
```

Arguments

`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`coreNumber`	A numeric value indicating how many CPU cores should be used in the analysis. The default value is 2.
`test`	A character string indicating which test should be used: `"LRT", "wald" or "Default Test"` The default is `"Default Test"`.
`groupGen`	A character string indicating the groups generation function: • "mclust" - `groups_identification_mclust()`; • "CoxHR" - `groups_identification_coxHR()`; • "clinical" - `groups_identification_clinical()`.
`clinical_pair`	A character string containing one of the group pairs selected after statistical analysis runned in `clinical_terms()` function.
`FC_cutoff`	Numerical value indicating the maximum value for Fold Change (FC). The default is 2.
`workDir`	A character string specifying the path to work directory.
`tumor`	A character string contaning one of the 33 tumors available in the TCGA project. For instance, the `"BRCA"` stands for breast cancer.
`FDR_cutoff`	Numerical value indicating the maximum value for FDR values. The default is `0.05`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

image_format	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).
cooksCutoff	Cooks distance remove outliers from the analysis; More details in DESeq2 results page. The default is FALSE.

Value

A matrix with DE genes in row and statistical values in columns.

Examples

```
## Not run:
#considering concatenate_files and groups_identification already runned
dea_DESeq2(Name = "HIF3A", test = "LRT", env = "env name without quotes")

## End(Not run)
```

---

dea_EBSeq *Run EBSeq gene Differential Expression Analysis (DEA).*

Description

Run EBSeq gene Differential Expression Analysis (DEA).

Usage

```
dea_EBSeq(Name, workDir, env, tumor, groupGen, clinical_pair,
  pairName = "G2_over_G1", rounds = 7, normType = "QuantileNorm",
  EBTest_Qtrm = 0.75, EBTest_QtrmCut = 10, p_cutoff = 0.05,
  FDR_cutoff = 0.05, FC_cutoff = 2, Width = 2000, Height = 1500,
  Res = 300, Unit = "px", image_format = "png",
  Bullard_quantile = 0.75)
```

Arguments

Name	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
workDir	A character string specifying the path to work directory.
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

`tumor` A character string contaning one of the 33 tumors available in the TCGA project. For instance, the `"BRCA"` stands for breast cancer.

`groupGen` A character string indicating the groups generation function:

- "mclust" - `groups_identification_mclust()`;
- "CoxHR" - `groups_identification_coxHR()`;
- "clinical" - `groups_identification_clinical()`.

`clinical_pair` A character string containing one of the group pairs selected after statistical analysis runned in `clinical_terms()` function.

`pairName` A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.

`rounds` Numerical value indicating the number of iterations. It is recommended to check the Alpha and Beta convergence plots in output and adjust this value until the hyper-parameter estimations converged. The default is `7`.

`normType` A character string indicating which EBSeq normalization factors type should be used in the analysis "QuantileNorm" or "MedianNorm". The default is `"all"`.

`EBTest_Qtrm, EBTest_QtrmCut`
Numerical value. It is removed from the analysis genes with EBTest_Qtrm th quantile ¡ = EBTest_QtrmCut. More details in EBSeq EBTest page. The default is `EBTest_Qtrm = 0.75` and `EBTest_QtrmCut = 10`.

`p_cutoff` Numerical value indicating the maximum value for p-values. The default is `0.05`.

`FDR_cutoff` Numerical value indicating the maximum value for FDR values. The default is `0.05`.

`FC_cutoff` Numerical value indicating the maximum value for Fold Change (FC). The default is `2`.

`Width` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Height` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Res` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Unit` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`image_format` A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".

`Bullard_quantile`
Numerical value indicating the quantile for the Bullard's normalization. The default is `0.75`.

Value

A matrix with DE genes in row and statistical values in columns.

Examples

```
## Not run:
#considering concatenate_files and groups_identification already runned
dea_EBSeq(pairName = "G2_over_G1", rounds = 7, Name = "HIF3A", env = "env_name")

## End(Not run)
```

`dea_edgeR` *Run edgeR gene Differential Expression Analysis (DEA).*

Description

Run edgeR gene Differential Expression Analysis (DEA).

Usage

```
dea_edgeR(Name, Method = "exacttest", clinical_pair, groupGen,
  FC_cutoff = 2, workDir, env, FDR_cutoff = 0.05, Width = 2000,
  Height = 1500, Res = 300, Unit = "px", image_format = "png")
```

Arguments

`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`Method`	A character string indicating which method should be used: `"exacttest"` or `"glmlrt"`. The default is `"exacttest"`.
`clinical_pair`	A character string containing one of the group pairs selected after statistical analysis runned in `clinical_terms()` function.
`groupGen`	A character string indicating the groups generation function: • "mclust" - `groups_identification_mclust()`; • "CoxHR" - `groups_identification_coxHR()`; • "clinical" - `groups_identification_clinical()`.
`FC_cutoff`	Numerical value indicating the maximum value for Fold Change (FC). The default is 2.
`workDir`	A character string specifying the path to work directory.
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).
`FDR_cutoff`	Numerical value indicating the maximum value for FDR values. The default is `0.05`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

Height	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Res	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Unit	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
image_format	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".

Value

A matrix with DE genes in row and statistical values in columns.

Examples

```
## Not run:
#considering concatenate_files and groups_identification already runned
dea_edgeR(Name = "HIF3A", env = "env name without quotes")

## End(Not run)
```

---

download_gdc *Download data from GDC Data Portal and GDC Legacy Archive*

Description

download_gdc is a function designed to download methylation, mutation, clinical data, protein expression, MAGETAB, gene expression, isoform expression, miRNA expression and clinical images data from GDC Data Portal and GDC Legacy Archive.

Usage

```
download_gdc(dataType = "gene", tumor, dataBase = "legacy",
  HTSeq = "", workDir = "~/Desktop", all.files = FALSE,
  Platform = "all")
```

Arguments

dataType	Type of data. It could be "methylation", "mutation", "clinical_supplement", "bios • Only present in "Legacy" database:"protein", "Exon quantification", "miRNA gen quantification", "miRNA isoform quantification", "isoform", "image", "cl • Only present in "GDC" database:"miRNA Expression Quantification", "Isoform
tumor	A character string contaning one of the 33 tumors available in the TCGA project. For instance, the "BRCA" stands for breast cancer.
dataBase	A character string specifying "GDC" for GDC Data Portal or "legacy" for GDC Legacy Archive.
HTSeq	A character string indicating which HTSeq workflow data should be downloaded: "Counts", "FPKM", "FPKM-UQ" or "all". The default is "all".
workDir	A character string specifying the path to work directory.

`all.files` A logical value. Set `FALSE` to avoid the download of not used data to reduce download size, e.g. quantification files. The default is `FALSE`.

`Platform` A character string indicating the platform name for methylation, exon quantificaton, miRNA, and mutation data.

- For mutation and exon quantificaton data:`"Illumina GA", "Illumina HiSeq" or "al`
- For methylation data`"Illumina Human Methylation 450", "Illumina Human Meth`
- For miRNA data:`"Illumina GA", "Illumina HiSeq", "H-miRNA_8x15K" (for GBM "all"`.

The default for all dataType cited is `"all"` (when downloading data).

Examples

```
## Not run:
# Downloading gene expression data from GDC Data Portal
download_gdc(dataType = "gene", "BRCA", "GDC", workDir = "~/Desktop", HTSeq = "counts")

# Downloading mutation data from GDC Legacy Archive
download_gdc(dataType = "mutation", "BRCA", "GDC", Platform = "Illumina HiSeq")

## End(Not run)
```

`DO_React_enrich` *DESEASE-ONTOLOGY and REACTOME ENRICHMENT*

Description

DESEASE-ONTOLOGY and REACTOME ENRICHMENT

Usage

```
DO_React_enrich(p_cutoff = 0.05, FDR_cutoff = 0.05, Width = 8,
  Height = 4, Res = 300, Unit = "in", image_format = "png",
  Tool = "edgeR", pairName = "G2_over_G1", env)
```

Arguments

`p_cutoff` Numerical value indicating the maximum value for p-values. The default is `0.05`.

`FDR_cutoff` Numerical value indicating the maximum value for FDR values. The default is `0.05`.

`Width` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Height` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Res` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Unit` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`image_format` A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".

`Tool` A character string indicating which differential expression analysis tool was last used.

`pairName` A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.

`env` A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Value

Enriched terms.

Examples

```
## Not run:
DO_React_enrich(Tool = "edgeR", env = "env name without quotes")

## End(Not run)
```

`draw_heatmap` *Draw a heatmap*

Description

Draw a heatmap

Usage

```
draw_heatmap(Tool, FC_cutoff = 2, Name, Method = "euclidean",
  pairName = "G2_over_G1", RawValues = FALSE, Width = 6,
  Height = 6, Res = 300, Unit = "in", image_format = "svg", env,
  ScaleMethod = "row", outerMargins = c(0, 0, 0, 0), cexCol = 0,
  cexRow = 0, degree = 45, labRow = NULL, labCol = NULL)
```

Arguments

`Tool` A character string indicating which differential expression analysis tool was last used.

`FC_cutoff` Numerical value indicating the maximum value for Fold Change (FC). The default is 2.

`Name` A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.

`Method` The agglomeration method to be used: `"euclidean"`, `"maximum"`, `"manhattan"`, `"canber`
The default is `"euclidean"`. More details about `Method` argument in `amap` package Dist `method` argument.

`pairName` A character string indicating the pair name to be used. When there are only two groups the default is `"G2_over_G1"`

`RawValues` A logical value. If `"TRUE"` the expression values are going to be converted to Z-Score before draw the heat map.

`Width` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Height` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Res` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`Unit` Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.

`image_format` A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".

`env` A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

`ScaleMethod` A character string indicating which method of scale should be used: `"none"`, `"row"`, `"column"`. The default is `"row"`. More details about `ScaleMethod` argument in heatmap `scale` argument.

`outerMargins` A numerical vector of the form c(bottom, left, top, right) giving the outer margins measured in lines of text. The default is no outer margins, i.e `"c(0,0,0,0)"`. Note: This argument is only used when `"labRow"` and `"labCol"` are set.

`cexCol, cexRow`
A numerical value giving the amount by which `"labRow"` and `"labCol"` should be modified relative to the default size.

`degree` The `"labCol"` rotation in degrees. The default value is 45 degrees.

`labRow, labCol`
A logical value. If `"TRUE"` it is displayed row names (`"labRow"`) and col names (`"labCol"`). The default is `"FALSE"` for both, in order to kepp the plot clean.

Examples

```
## Not run:
draw_heatmap("EBSeq", Name = "HIF3A", env = "env name without quotes")

## End(Not run)
```

GOnto *Perform Gene-Ontology Pathways Enrichment*

Description

Perform Gene-Ontology Pathways Enrichment

Usage

```
GOnto(condition, use_genes_without_cat = TRUE, FDR_cutoff = 0.05,
  Width = 2000, Height = 2000, Res = 300, Unit = "px",
  image_format = "png", Tool, ID = "GeneID", pairName = "G2_over_G1",
  env)
```

Arguments

`condition`	A character string containing which condition should be used: "Upregulated", "Downregulated" or "all".
`use_genes_without_cat`	Logical value where `FALSE` indicate that genes outside the category being tested will be ignored in the calculation of p-values. The default is "FALSE".
`FDR_cutoff`	Numerical value indicating the maximum value for FDR values. The default is `0.05`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`Tool`	A character string indicating which differential expression analysis tool was last used.
`ID`	A character string indicating which ID should be used: "HUGO", "GeneSymbol", "ensembl" , "refGene" or "GeneID". The default is `"GeneID"`.
`pairName`	A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Value

Enriched terms.

Examples

```
## Not run:
GOnto(condition = "Upregulated", Tool = "edgeR", env = "env name without quotes")

## End(Not run)
```

`groups_identification_coxHR` *Separate patients in groups*

Description

`groups_identification_cox` is a function designed to separate patients in groups, based in clinical data, and coxHR.

Usage

```
groups_identification_coxHR(Name, dataType, Width = 3000,
  Height = 2000, Res = 400, Unit = "px", image_format = "png",
  saveData = TRUE, env, tumor, dataBase, workDir)
```

Arguments

`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`dataType`	Type of data. It could be `"methylation", "mutation", "clinical_supplement", "bios` • Only present in "Legacy" database:`"protein", "Exon quantification", "miRNA gen quantification", "miRNA isoform quantification", "isoform", "image", "cl` • Only present in "GDC" database:`"miRNA Expression Quantification", "Isoform`
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`saveData`	Logical value where `TRUE` indicates that the concatenate and filtered matrix should be saved in local storage. The default is FALSE.

env — A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria:

- 'tumor_dataBase_dataType_tumor_data' or
- 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

tumor — A character string contaning one of the 33 tumors available in the TCGA project. For instance, the "BRCA" stands for breast cancer.

dataBase — A character string specifying "GDC" for GDC Data Portal or "legacy" for GDC Legacy Archive.

workDir — A character string specifying the path to work directory.

Examples

```
## Not run:
groups_identification_coxHR("HIF3A", "gene", tumor = "UCS", dataBase = "legacy")

## End(Not run)
```

---

groups_identification_mclust

Separate patients in groups

Description

groups_identification_mclust is a function designed to separate patients in groups, powered by mclust.

Usage

```
groups_identification_mclust(dataType, group.number = "",
  modelName = NULL, uncertaintyCutoff = 0.05, reRunPlots = FALSE,
  nBreaks = 55, Width = 2000, Height = 1500, Res = 300,
  Unit = "px", image_format = "png", saveData = TRUE, env, tumor,
  dataBase, workDir = "~/Desktop", Name)
```

Arguments

dataType — Type of data. It could be "methylation", "mutation", "clinical_supplement", "bios

- Only present in "Legacy" database:"protein", "Exon quantification", "miRNA gen quantification", "miRNA isoform quantification", "isoform", "image", "cl
- Only present in "GDC" database:"miRNA Expression Quantification", "Isoform

group.number — Numerical value indicating how many groups should be generated.

modelName — A character string indicating which mclust modelName will be used. For more details please check mclustModelNames help file.

uncertaintyCutoff — Numerical value indicating which uncertainty value for the separation should be tolerated. Patients over this threshold will be removed from the analysis.

`reRunPlots`	Logical value where TRUE indicate that the function should run the step of group generation using the `uncertaintyCutoff` parameter for filtering the data. The default is `FALSE`.
`nBreaks`	Numerical value giving the number of cells for the `hist` bars. As default `nBreaks = 55`.
`Width, Height, Res, Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`saveData`	Logical value where `TRUE` indicates that the concatenate and filtered matrix should be saved in local storage. The default is FALSE.
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).
`tumor`	A character string contaning one of the 33 tumors available in the TCGA project. For instance, the `"BRCA"` stands for breast cancer.
`dataBase`	A character string specifying `"GDC"` for GDC Data Portal or `"legacy"` for GDC Legacy Archive.
`workDir`	A character string specifying the path to work directory.
`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.

Examples

```
## Not run:
#separating isoform uc002peh.2 expression data patients in two groups
groups_identification_mclust("isoform", 2, Name = "uc002peh.2", modelName = "E")

## End(Not run)
```

`GSEA` *Gene Set Enrichment Analysis*

Description

Gene Set Enrichment Analysis

Usage

```
GSEA(FDR_cutoff = 0.05, Width = 10, Height = 3, Res = 500,
  Unit = "in", image_format = "png", Tool = "edgeR", ID = "GeneID",
  pairName = "G2_over_G1", env)
```

Arguments

FDR_cutoff	Numerical value indicating the maximum value for FDR values. The default is 0.05.
Width	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Height	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Res	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Unit	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
image_format	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
Tool	A character string indicating which differential expression analysis tool was last used.
ID	A character string indicating which ID should be used: "HUGO", "GeneSymbol", "ensembl" , "refGene" or "GeneID". The default is "GeneID".
pairName	A character string indicating which condition name should be used. When there are only two groups the default is "G2_over_G1".
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Value

Enriched terms.

Examples

```
## Not run:
GSEA(Tool = "edgeR", env = "env name without quotes")

## End(Not run)
```

KEEG_enrich *Perform KEEG Pathways Enrichment*

Description

Perform KEEG Pathways Enrichment

Usage

```
KEEG_enrich(Tool = "edgeR", FDR_cutoff = 0.05, ID = "GeneID",
  pairName = "G2_over_G1", Width = 8, Height = 4, Res = 300,
  Unit = "in", image_format = "png", env)
```

Arguments

`Tool`	A character string indicating which differential expression analysis tool was last used.
`FDR_cutoff`	Numerical value indicating the maximum value for FDR values. The default is `0.05`.
`ID`	A character string indicating which ID should be used: "HUGO", "GeneSymbol", "ensembl" , "refGene" or "GeneID". The default is `"GeneID"`.
`pairName`	A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Value

Enriched terms.

Examples

```
## Not run:
KEEG_enrich(Tool = "edgeR", env = "env name without quotes")

## End(Not run)
```

`PCA_Analysis` *Generate plot for Principal Component Analysis*

Description

Generate plot for Principal Component Analysis

Usage

```
PCA_Analysis(Tool, Name, workDir, pairName = "G2_over_G1", Width = 4,
  Height = 2, Res = 300, Unit = "in", image_format = "png", env)
```

Arguments

`Tool`	A character string indicating which differential expression analysis tool was last used.
`Name`	A character string indicating the desired values to be used in next analysis. For instance, "HIF3A" in the legacy gene expression matrix, "mir-1307" in the miRNA quantification matrix, or "HER2" in the protein quantification matrix.
`workDir`	A character string specifying the path to work directory.
`pairName`	A character string indicating which condition name should be used. When there are only two groups the default is `"G2_over_G1"`.
`Width`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Height`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Res`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`Unit`	Graphical parameters. See par for more details. As default `Width = 2000, Height = 1500, 300 and Unit = "px"`.
`image_format`	A character string indicating which image_format will be used. It could be "png" or "svg". The only unit available in "svg" is inches ('in'). The default is "png".
`env`	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Examples

```
## Not run:
PCA_Analysis("EBSeq", "gene", "HIF3A", pairName = "G2_over_G1", env = "env name without quotes")

## End(Not run)
```

`table2DOAGDC` *From global to DOAGDC environment*

Description

From global to DOAGDC environment

Usage

```
table2DOAGDC(table, dataType, normalization = TRUE, dataBase, tumor,
  tumorData = TRUE, env)
```

Arguments

table	A matrix or data frame with data used in this package (i.e Gene, isoform and protein expression, methylation and mutation data)
dataType	Type of data. It could be "methylation", "mutation", "clinical_supplement", "bios • Only present in "Legacy" database:"protein", "Exon quantification", "miRNA gen quantification", "miRNA isoform quantification", "isoform", "image", "cl • Only present in "GDC" database:"miRNA Expression Quantification", "Isoform
normalization	Logical value where TRUE specify the desire to work with normalized files only. When FALSE, in the second run, do not forget to set env argument. This argument is only applyable to gene and isoform expression data from GDC Legacy Archive. The default is TRUE.
dataBase	If the data was downloaded from GDC/Legacy data base, however not using DOAGDC, please specified which data base. If the data do not come from GDC/Legacy data base, and it is related with genome version GRCh 37, please insert "legacy" in this argument. Otherwise, if it is related with genome version GRCh 38, please insert "GDC" in this argument.
tumor	A character string contaning one of the 33 tumors available in the TCGA project. For instance, the "BRCA" stands for breast cancer.
tumorData	Logical value where TRUE specifies the desire to work with tumor tissue files only. When set to FALSE, it creates two matrices, one containing tumor data and other containing data from not-tumor tissue. The default is TRUE.
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Examples

```
## Not run:
patient <- paste0(paste0("patient_", LETTERS[1:4]))
genes <- paste0("gene_", seq(1, 5))
# generate a simulated gene expression matrix
example_gene_table <- matrix(runif(20, 0.0, 90.5), 5, 4, TRUE, list(genes, patient))
# without env created
table2DOAGDC(example_gene_table, dataType = "gene", dataBase = "legacy", tumor = "UCS")
# with env created
table2DOAGDC(example_gene_table, env = "env name without quotes")

## End(Not run)
```

---

venn_diagram *Venn diagram of differential expression genes list*

Description

Venn diagram of differential expression genes list

Usage

```
venn_diagram(FinalData, n_pack = 3, packageNames,
  pairName = "G2_over_G1", Width = 2000, Height = 2000, Res = 300,
  Unit = "px", Colors = c("green", "blue", "red"),
  VennDiagram_imagetype = "png", workDir, env)
```

Arguments

FinalData	A character string indicating the name of which diferential expression package should be used to get the statistical values in the final list. The default is "EBSeq".
n_pack	A numerical value indicating the number of expression analysis to be used in venn diagram. It is expected the number 2 or 3. The default is 3.
packageNames	A character vector indicating the names of at least two diferential expression packages used in previous steps: "DESeq2", "edgeR", "DESeq2, or "All".
pairName	A character string indicating which condition name should be used. When there are only two groups the default is "G2_over_G1".
Width	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Height	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Res	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Unit	Graphical parameters. See par for more details. As default Width = 2000, Height = 1500, 300 and Unit = "px".
Colors	A character vector indicating the colors to be used in the venn diagram. The default is c('green', 'blue', "red").
VennDiagram_imagetype	A character string indicating the image_format (e.g. "tiff", "png" or "svg"). The default is "png".
workDir	A character string specifying the path to work directory.
env	A character string containing the environment name that should be used. If none has been set yet, the function will create one in global environment following the standard criteria: • 'tumor_dataBase_dataType_tumor_data' or • 'tumor_dataBase_dataType_both_data' (for tumor and not tumor data in separated matrices).

Examples

```
## Not run:
venn_diagram("DESeq2", 2, packageNames = c("edgeR", "DESeq2"), env = "env name without quotes")

## End(Not run)
```

Index