Package 'rprism'

August 17, 2022

Title Support for the Meta PRISM project

Version 2.2.5

Maintainer 'Yoann Pradat' <yoann.pradat@centralesupelec.fr>

Description Meta PRISM project is an institutional project from the French University Hospital Gustave Roussy. This project aims at describing the molecular landscape of metastatic cancers from all solid tumor types.

License BSD_3_clause + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.1

URL https://github.com/gustaveroussy/MetaPRISM/tree/master/functions/rprism

Depends R (>= 3.0)

Imports Exact,
doParallel,
dplyr,
foreach,
magrittr,
rlang,
readr,
readxl,
stats,
tibble,
tidyr,
writexl,
yaml

Suggests testthat

R topics documented:

add_annotations_bio_cln 2
add_pvals_tables 3
compute_pvals_table 4
correct_multiple_testing 5
delineate_data 5

load_bio . . . 6
load_cln . . . 7
load_colors . . . 8
load_dsg . . . 8
load_from_data . . . 9
load_ids . . . 9
load_resource . . . 10
load_rna_fus . . . 10
load_rna_gex . . . 11
load_stat_for_wes_mut . . . 12
load_summary_rna_fus . . . 13
load_summary_rna_gex . . . 13
load_summary_wes_mut . . . 14
load_table . . . 15
load_wes_mut . . . 15
merge_on_rows . . . 16
preprocess_wes_mut . . . 17
rprism . . . 18
save_to_data . . . 18
select_tumor_types . . . 18
setwd_to_data . . . 19
setwd_to_logs . . . 19
setwd_to_results . . . 20
setwd_to_scripts . . . 20
split_targeted_therapy_targets . . . 21
subset_data . . . 21

Index **22**

`add_annotations_bio_cln` *Add columns*

Description

Add annotations from bio and cln files.

Usage

```
add_annotations_bio_cln(
  df,
  df_bio = NULL,
  df_cln = NULL,
  col_on_bio = NULL,
  col_on_cln = NULL,
  cols_bio = NULL,
  cols_cln = NULL
)
```

Arguments

df	the dataframe to be annotated
df_bio	the dataframe of biospecimen annotations
df_cln	the dataframe of subjects annotations
col_on_bio	column in the dataframe to be used to connect to df_bio
col_on_cln	column in the dataframe to be used to connect to df_cln
cols_bio	columns from the df_bio dataframe to be added to df
cols_cln	columns from the df_cln dataframe to be added to df

Author(s)

Yoann Pradat

---

add_pvals_tables *Perform multiple testing correction from a set of pvalues*

Description

From a list of multiple tables of pvalues, perform one multiple testing comparison correction considering all tests of all tables at the same time. Corrected pvalues (qvalues) are returned in the same format as the input list of pvalue tables.

Usage

```
add_pvals_tables(
  dfs_plot,
  cohorts_a,
  cohorts_b,
  test,
  col_evt,
  in_plot_evt,
  tt_keep,
  n_cores,
  suffix = ""
)
```

Arguments

dfs_pvals	a list of dataframes. Names must contain the values in cohorts.
cohorts	A character vector.
method	The multiple testing correction method.

Value

a list of dataframes. Names and format are identical to that of dfs_pvals.

Author(s)

Yoann Pradat

`compute_pvals_table` *Compute pvalues of multiple 2x2 tables.*

Description

From a list of 2 data.frames and their margins, build 2 tables with identical rows and columns and run Fisher tests on each of the 2x2 contigency table built from each pair of cells of both tables.

Usage

```
compute_pvals_table(
  dfs,
  cohort_a,
  cohort_b,
  row_col,
  row_names,
  col_names,
  test = "fisher",
  n_cores = 1
)
```

Arguments

`dfs`	a 2-level list of data.frames. `cohort_a` and `cohort_b` must be 1st level names while 2nd-level names must contain 'count' (the count matrix) and 'count_col' (the vector of columns margins).
`cohort_a`	Name of the cohort A.
`cohort_b`	Name of the cohort B.
`row_col`	Name of the column to be used as rownames.
`row_names`	Values of `row_col` to run tests on.
`col_names`	Subcohort names of each of cohort A and B to run tests on.
`test`	Name of the statistical test employed. Use "fisher" for Fisher-Booschloo, and "cmh" for CMH stratified fisher test.
`n_cores`	Number of cores to be used for parallel computations.

Author(s)

Yoann Pradat

correct_multiple_testing

Perform multiple testing correction from a set of pvalues

Description

From a list of multiple tables of pvalues, perform one multiple testing comparison correction considering all tests of all tables at the same time. Corrected pvalues (qvalues) are returned in the same format as the input list of pvalue tables.

Usage

```
correct_multiple_testing(dfs_pvals, cohorts, method = "fdr")
```

Arguments

dfs_pvals	a list of dataframes. Names must contain the values in cohorts.
cohorts	A character vector.
method	The multiple testing correction method.

Value

a list of dataframes. Names and format are identical to that of dfs_pvals.

Author(s)

Yoann Pradat

delineate_data

Delineate treatments from compact format to more user-friendly format

Description

Compressed data must consist of a name with additional data enclosed in brackets and separated by a common delimiter.

Usage

```
delineate_data(
  df,
  col_name,
  format,
  separate_rows = T,
  prefix = "",
  suffix = "",
  delineate_dates = F,
  convert_dates_to_datetime = T,
  sep = ";",
  name = "Name"
)
```

Arguments

df	a data.frame that must contain the column col_name.
col_name	Name of the column to be delineated.
format	Character vector of delineated column names corresponding the format of the compressed data.
separate_rows	(optional) Should data be first unnested on the "|" separator"?
prefix	(optional) Prefix to the new columns that contain the delineated data.
suffix	(optional) Suffix to the new columns that contain the delineated data.
delineate_dates	(optional) Used only if "Dates" is in format. Split "Dates" into "Date_Min" and "Date_Max".
convert_dates_to_datetime	(optional) Used only if delineate_dates=True. Should dates be converted to datetime type?
sep	(optional) The delimiter of the compressed data.
name	(optional) Name given to the column containing the value before the brackets. Prefix and suffixes specified will be added to this name.

Value

a data.frame with possibly more rows (if separate_rows is set to TRUE) and more columns (as many as names in the list format) built from df.

---

load_bio *Load biospecimen files.*

Description

This function loads biospecimen tables (one line is one sample id) for the specified study and optionally specified identifiers. The mode parameters allow to choose which table should be loaded between 'in_design' and 'all'.

Usage

```
load_bio(
  study,
  identifiers = NULL,
  identifiers_name = NULL,
  mode = "in_design",
  ...
)
```

Arguments

`study`	Name of the cohort used in the naming of the files. Choose "met500", "prism" or "tcga".
`identifiers`	(optional) If not NULL, return clinical tables only for the individuals specified
`identifiers_name`	(optional) The name of the column the `identifiers` values correspond to.
`mode`	(optional) Choose "in_design" to load only data for samples in the design or 'all' to load data for all samples of the cohort.
`...`	Extra parameters passed to `load_from_data`

Value

a data.frame

Author(s)

Yoann Pradat

`load_cln` *Load clinical files.*

Description

This function loads clinical tables (one line is one subject id) for the specified study and optionally specified identifiers. The `mode` parameters allow to choose which table should be loaded between 'in_design' and 'all'.

Usage

```
load_cln(
  study,
  identifiers = NULL,
  identifiers_name = NULL,
  mode = "in_design",
  ...
)
```

Arguments

`study`	Name of the cohort used in the naming of the files. Choose "met500", "prism" or "tcga".
`identifiers`	(optional) If not NULL, return clinical tables only for the individuals specified.
`identifiers_name`	(optional) The name of the column the `identifiers` values correspond to.
`mode`	(optional) Either 'in_design' or 'all' for 'pancancer' annotations or 'brca' for per-cancer annotations.
`...`	Extra parameters passed to `load_from_data`

Value

a data.frame

Author(s)

Yoann Pradat

`load_colors` *Load lists of colors from Excel table*

Description

Load one or multiple sheets from an excel workbook containing lists of colors for categories of different fields.

Usage

```
load_colors(sheet, as_tibble = F)
```

Arguments

`sheet`	Name of the excel spreadsheet.
`as_tibble`	(optional) Set to TRUE if you would like to have the list of colors returned in a `tibble` dataframe. If FALSE, colors are returned as a named list.

Value

A named list or a `tibble` dataframe. Names are categories of some fields or custom names and values are colors in hexadecimal character vectors.

`load_dsg` *Load the design table*

Description

Load the design table for the specified study. Only "prism" is supported.

Usage

```
load_dsg(study)
```

Arguments

`study`	Name of the cohort used in the naming of the files. Choose "met500", "prism" or "tcga".

Value

a data.frame

Author(s)

Yoann Pradat

load_from_data *Load a table located in the data folder*

Description

Load a table located in the data folder

Usage

```
load_from_data(path, ...)
```

Arguments

path	Path to the file.
...	Extra parameters passed to read_delim.

load_ids *Load identifiers files.*

Description

This function loads the table linking all subject, sample and idspsy identifiers..

Usage

```
load_ids(study, ...)
```

Arguments

study	Name of the cohort used in the naming of the files. Choose "met500", "prism" or "tcga".
...	Extra parameters passed to load_from_data

Value

a data.frame

Author(s)

Yoann Pradat

load_resource *Load resources from different databases or used for different analyses.*

Description

Load resources from different databases or used for different analyses.

Usage

```
load_resource(database, name, ...)
```

Arguments

database	Name of the database to be load from. Supported values are • civic • oncokb • cosmic • gencode • curated
name	Name of the resource
...	Extra parameters passed to load_from_data

Author(s)

Yoann Pradat

load_rna_fus *Load fusions called from rna-seq data.*

Description

This function loads RNA gene fusion tables for the specified cohort or subcohort.

Usage

```
load_rna_fus(
  study,
  mode = NULL,
  identifiers = NULL,
  identifiers_name = NULL,
  ...
)
```

Arguments

study	Name of the cohort used in the naming of the files. Choose 'met500', 'prism' or 'tcga'.
mode	The mode determines the caller and/or the filtering level. • "met500": arriba, ericscript, pizzly, starfusion, oncokb_civic. • "prism": arriba, ericscript, pizzly, starfusion, oncokb_civic. • "tcga": aggregated, oncokb_civic.
identifiers	(optional) If not NULL, return RNA data only for the identifiers values. Used only if identifiers_name is not NULL.
identifiers_name	(optional) Field name of the identifiers.
...	Extra parameters passed to load_from_data

Value

a tibble data.frame

Author(s)

Yoann Pradat

---

load_rna_gex *Load raw gene expression data files.*

Description

This function loads RNA gene expression tables for the specified cohort or subcohort.

Usage

```
load_rna_gex(
  study,
  level = c("genes", "transcripts"),
  metric = c("counts", "TPM", "length"),
  test_mode = F,
  identifiers = NULL,
  identifiers_name = NULL,
  ...
)
```

Arguments

study	Name of the cohort used in the naming of the files. Choose one of. • "met500": Trim Galore > Kallisto > Tximport. See https://github.com/gustaveroussy/MetaPRISM • "prism": Trim Galore > Kallisto > Tximport. See https://github.com/gustaveroussy/MetaPRISM • "tcga": From https://stanfordmedicine.app.box.com/s/lu703xuaulfz02vgd2lunxnvt4m

- "tcga_6_samples_prism": PRISM pipeline on 6 TCGA FASTQs.
- "tcga_6_samples_gao": Gao pipeline on 6 TCGA FASTQs.

`level` Level at which counts are aggregated. You may choose "genes" or "transcripts".

`metric` string or vector of strings. If not specified, every metric is loaded.

- "TPM" to load "abundance.kallisto"
- "counts" to load "counts.kallisto"
- "length" to load "length.kallisto"

`test_mode` (optional) Set to True to use submatrices generated with `generate_rna_test` from `_util_rna` module.

`identifiers` (optional) If not NULL, return RNA data only for the identifiers values. Used only if `identifiers_name` is not NULL.

`identifiers_name` (optional) Field name of the identifiers.

`...` Extra parameters passed to `load_from_data`

Value

a tibble data.frame

Author(s)

Yoann Pradat

---

load_stat_for_wes_mut *Build a dataframe with per patient statistics of MAF files loading.*

Description

Used to see which data was loaded for which patient in MAF files.

Usage

```
load_stat_for_wes_mut(
  df_maf,
  patient_field,
  tumor_field = NULL,
  normal_field = NULL
)
```

Arguments

`df_maf` The dataframe in MAF format

`patient_field` The name of the patient column

`tumor_field` The name of the tumor column

`normal_field` The name of the normal column

Value

a list of `data.frame`

Author(s)

Yoann Pradat

`load_summary_rna_fus` *Load summary of rna-seq gene fusion tables*

Description

The summary table links Sample_Id to Subject_Id. This allows easier connection to other data tables using either of Sample_Id or Subject_Id.

Usage

```
load_summary_rna_fus(study, mode = NULL)
```

Arguments

`study`	Name of the cohort used in the naming of the files. Choose 'met500', 'prism' or 'tcga'.
`mode`	The mode determines the caller and/or the filtering level. • "met500": arriba, ericscript, pizzly, starfusion, oncokb_civic. • "prism": arriba, ericscript, pizzly, starfusion, oncokb_civic. • "tcga": aggregated, oncokb_civic.

Value

a data.frame with summary

`load_summary_rna_gex` *Load summary of RNA tables*

Description

The summary table links original names in the RNA table to Sample_Id and Subject_Id. This allows easier connection to other data tables using either of Sample_Id or Subject_Id.

Usage

```
load_summary_rna_gex(study, level, metric, test_mode = F)
```

Arguments

study	Name of the cohort used in the naming of the files. Choose one of. • "met500": Trim Galore > Kallisto > Tximport. See https://github.com/gustaveroussy/MetaPRISM • "prism": Trim Galore > Kallisto > Tximport. See https://github.com/gustaveroussy/MetaPRISM • "tcga": From https://stanfordmedicine.app.box.com/s/lu703xuaulfz02vgd2lunxnvt4m • "tcga_6_samples_prism": PRISM pipeline on 6 TCGA FASTQs. • "tcga_6_samples_gao": Gao pipeline on 6 TCGA FASTQs.
level	Level at which counts are aggregated. You may choose "genes" or "transcripts".
metric	string or vector of strings. If not specified, every metric is loaded. • "TPM" to load "abundance.kallisto" • "counts" to load "counts.kallisto" • "length" to load "length.kallisto"
test_mode	(optional) Set to True to use submatrices generated with `generate_rna_test` from `_util_rna` module.

Value

a data.frame with summary

---

load_summary_wes_mut *Load summary of MAF tables*

Description

The summary table links original names in the WES table to Sample_Id and Subject_Id. This allows easier connection to other data tables using either of Sample_Id or Subject_Id.

Usage

```
load_summary_wes_mut(study, mode = "oncotator_filtered")
```

Arguments

study	Name of the cohort used in the naming of the files. Choose one of "met500", "prism" or "tcga".
mode	(optional) Choose which MAF file to load.

Value

a data.frame with summary

load_table *Load a table.*

Description

Load a table.

Usage

```
load_table(path, header_prefix = NULL, ...)
```

Arguments

path	Path to the file.
header_prefix	If not NULL, skip header rows that will be identified as first lines starting with this prefix.
...	Extra parameters passed to read_delim.

load_wes_mut *Load mutation tables.*

Description

Load mutation tables for the specified study and optionally specified identifiers. This function is an R reimplementation of the function load_wes_mut from the pyprism package.

Usage

```
load_wes_mut(
  study,
  identifiers = NULL,
  identifiers_name = NULL,
  mode = "somatic_maf",
  ...
)
```

Arguments

study	Name of the cohort used in the naming of the files. Choose one of "met500", "prism" or "tcga".
identifiers	(optional) If not NULL, return clinical tables only for the individuals specified
identifiers_name	(optional) The name of the column the identifiers values correspond to.
mode	(optional) Choose which MAF file to load.
...	Extra parameters passed to load_from_data

Details

- If "somatic_maf", load data/study/wes/somatic_maf/somatic_calls.maf.gz
- If "somatic_filters", load data/study/wes/somatic_maf/somatic_calls_filters.tsv.gz
- If "somatic_oncokb", load data/study/wes/somatic_maf/somatic_calls_oncokb.maf.gz
- If "somatic_civic", load data/study/wes/somatic_maf/somatic_calls_civic.maf.gz

Value

a `data.frame`

Author(s)

Yoann Pradat

`merge_on_rows` *Merge two dataframes on their row names.*

Description

This function merges two dataframes on their row names and provides the user with the choice on which columns should be kept.

Usage

```
merge_on_rows(df_x, df_y, how_cols = "inner")
```

Arguments

`df_x`	a data.frame
`df_y`	a data.frame
`how_cols`	a character vector specifying how the merge should proceed 1. 'inner' keep only the columns common to both `df_x` and `df_y` 2. 'x' keep only the columns of `df_x` 3. 'y' keep only the columns of `df_y` 4. 'outer' keep all the columns of `df_x` and `df_y` This function raises an error if a row entry (identified by its row name) is present in both `df_x` and `df_y` and has divergent values on a column that is present in both data frames.

Value

a data.frame

Author(s)

Yoann Pradat

`preprocess_wes_mut` *Preprocess mutations table*

Description

Add bio and/or cln attributes to the mutations table and perform a selection of one pair tumor/normal per subject if requested.

Usage

```
preprocess_wes_mut(
  df_mut,
  cohort,
  cols_bio = c(),
  cols_cln = c(),
  select_pairs = FALSE,
  selection_mut = "all",
  verbose = TRUE
)
```

Arguments

`df_mut`	data.frame of mutations
`cohort`	name of the cohort
`cols_bio`	(optional) list ist of bio attributes
`cols_cln`	(optional) list of cln attributes.
`select_pairs`	(optional) If set to TRUE, 1 pair tumor/normal is selected for each subject.
`selection_mut`	(optional) Here are the available modes 1. 'all'no selection of variants 2. 'annotated'no selection of variants 3. 'non_synonymous'select variants for which the value of 'Variant_Classification' is one of * Frame_Shift_Del * Frame_Shift_Ins, * Splice_Site * Translation_Start_Site * Nonsense_Mutation * Nonstop_Mutation * In_Frame_Del * In_Frame_Ins * Missense_Mutation * Start_Codon_Del * Start_Codon_SNP * Stop_Codon_Del * Stop_Codon_Ins 4. 'truncating'select variants for which the value of 'Variant_Classification' is one of * Frame_Shift_Del * Frame_Shift_Ins, * Splice_Site * Nonsense_Mutation
`verbose`	(optional) Should info messages be printed?

Value

A dataframe of mutations with possibly additional attributes and possibly a reduced number of lines.

rprism — *rprism: An R package accompanying the analysis of the META-PRISM cohort*

Description

The rprism package provides handy functions for curating data, loading row and curated data and for performing recurrent operatons on the data. It also provides util functions for running statistical tests or for setting the working directory to specific locations for instance.

save_to_data — *Write a table to a path in the data folder*

Description

Write a table to a path in the data folder

Usage

```
save_to_data(df, path, ...)
```

Arguments

df	a data.frame that contains the column col_name.
path	a path relative to the folder data/.
...	Extra parameters passed to write_xlsx or write_delim.
col_name	Name of the the column to be used to susbet the dataframe.

select_tumor_types — *Function for selecting tumor types.*

Description

Using the column col_tt, this function subsets each input dataframe to the values of col_tt with at least tt_min_size entries.

Usage

```
select_tumor_types(dfs, col_tt, tt_min_size, tt_drop)
```

Arguments

dfs	A named list of data.frame. They must contain columns col_tt.
tt_min_size	Numeric value.
tt_drop	Values that should be discarded regarless of the size.

Value

a list containing a named list of dataframes and a named list of character vectors, each with the sames names as `dfs`.

Author(s)

Yoann Pradat

---

setwd_to_data *Set the working directory to the data folder of the project.*

Description

Set the working directory to the data folder of the project.

Usage

```
setwd_to_data()
```

Value

current_wd name of the working directory before change

Author(s)

Yoann Pradat

---

setwd_to_logs *Set the working directory to the logs folder of the project.*

Description

Set the working directory to the logs folder of the project.

Usage

```
setwd_to_logs()
```

Value

current_wd name of the working directory before change

Author(s)

Yoann Pradat

`setwd_to_results` *Set the working directory to the results folder of the project.*

Description

Set the working directory to the results folder of the project.

Usage

```
setwd_to_results()
```

Value

current_wd name of the working directory before change

Author(s)

Yoann Pradat

`setwd_to_scripts` *Set the working directory to the scripts folder of the project.*

Description

Set the working directory to the scripts folder of the project.

Usage

```
setwd_to_scripts()
```

Value

current_wd name of the working directory before change

Author(s)

Yoann Pradat

split_targeted_therapy_targets *Split targeted therapy classes with multiple targets*

Description

Split multiple-targets therapies into multiple single-target therapies.

Usage

```
split_targeted_therapy_targets(x, sep = "|")
```

Arguments

x	a string containing entries (e.g drug names) separated by sep
sep	the separator, default is "|"

Value

a string

subset_data *Load a table located in the data folder*

Description

Load a table located in the data folder

Usage

```
subset_data(df, values = NULL, col_name = NULL)
```

Arguments

df	a data.frame that contains the column col_name.
values	a list of values to be selected.
col_name	Name of the the column to be used to susbet the dataframe.

Value

a data.frame containing a subset of the rows of df.

Index

∗ **export**
merge_on_rows, 16

add_annotations_bio_cln, 2
add_pvals_tables, 3

compute_pvals_table, 4
correct_multiple_testing, 5

delineate_data, 5

load_bio, 6
load_cln, 7
load_colors, 8
load_dsg, 8
load_from_data, *7*, *9*, 9, *10–12*, *15*
load_ids, 9
load_resource, 10
load_rna_fus, 10
load_rna_gex, 11
load_stat_for_wes_mut, 12
load_summary_rna_fus, 13
load_summary_rna_gex, 13
load_summary_wes_mut, 14
load_table, 15
load_wes_mut, 15

merge_on_rows, 16

preprocess_wes_mut, 17

rprism, 18

save_to_data, 18
select_tumor_types, 18
setwd_to_data, 19
setwd_to_logs, 19
setwd_to_results, 20
setwd_to_scripts, 20
split_targeted_therapy_targets, 21
study, *16*
subset_data, 21