Package ‘semseeker’

August 17, 2023

Type Package

Title Stochastic Epigenetic Mutations SEM Seeker

Version 0.9.7

Author Luigi Corsaro, Davide Sacco

Maintainer Luigi Corsaro <lcorsaro69@gmail.com>

Description Stochastic epimutation and enriched region upstream and downstream tool for EWAS.

License AGPL-3

Encoding UTF-8

URL https://github.com/drake69/semseeker

BugReports https://github.com/drake69/semseeker/issues

Imports coxed, dplyr, doFuture, doRNG, FactoMineR, factoextra, foreach, FSA, fst, future, future.apply, ggplot2, gtools, Hmisc, lqmm, openxlsx, plyr, progressr, quantreg, readxl, reshape, reshape2, R.utils, rlang, stats, stringr, utils, withr, zoo

RoxygenNote 7.2.3

Suggests pathfindR, GEOquery, stringi, testthat

Depends R (>= 4.3.0)

LazyData true

LazyDataCompression gzip

SysDataCompression bzip2

Config/testthat/parallel false

Config/testthat/edition 3

NeedsCompilation no

R topics documented:

`analyze_population` *Calculate stochastic epi mutations from a methylation dataset as outcome report of pivot*

Description

Calculate stochastic epi mutations from a methylation dataset as outcome report of pivot

Usage

```
analyze_population(
  methylation_data,
  sliding_window_size,
  sample_sheet,
  beta_thresholds,
  bonferroni_threshold = 0.05,
  probe_features
)
```

Arguments

`methylation_data`
whole matrix of data to analyze.

`sliding_window_size`
size of the sliding widows to compute epilesions default 11 probe_features.

`sample_sheet` name of samplesheet's column to use as control population selector followed by selection value,

`beta_thresholds` thresholds defined to calculate epimutations lesions definition

`bonferroni_threshold` threshold to define which pValue accept for

`probe_features` probe_features detail from 27 to EPIC illumina dataset

Value

files into the result folder with pivot table and bedgraph.

`analyze_single_sample` *analyze_single_sample*

Description

analyze_single_sample

Usage

```
analyze_single_sample(
  values,
  sliding_window_size,
  thresholds,
  figure,
  sample_detail,
  bonferroni_threshold = 0.05,
  probe_features
)
```

Arguments

`values` values of methylation

`sliding_window_size` size of window sliding to calculate hypergeometric

`thresholds` threshold to use for comparison

`figure` which figure's of sasmple will be analized HYPO or HYPER

`sample_detail` details of the sample to analyze

`bonferroni_threshold` bonferroni threshold to validate pValue

`probe_features` probe_features details to be used

Value

list of lesion count and probe_features count

`annotate_bed` *create an annotated file for each marker, figure, area and subarea, each file has all the sample_groups used to calculate epimutation*

Description

create an annotated file for each marker, figure, area and subarea, each file has all the sample_groups used to calculate epimutation

Usage

```
annotate_bed()
```

Value

nothing

`apply_stat_model` *Title*

Description

Title

Usage

```
apply_stat_model(
  tempDataFrame,
  g_start,
  family_test,
  covariates = NULL,
  key,
  transformation,
  dototal,
  session_folder,
  independent_variable,
  depth_analysis = 3,
  ...
)
```

Arguments

`tempDataFrame`	data frame to apply association
`g_start`	index of starting data
`family_test`	family of test to run
`covariates`	vector of covariates
`key`	key to identify file to elaborate
`transformation`	transformation to apply to covariates, burden and independent variable

`dototal` do a total per area

`session_folder` where to save log file

`independent_variable`
independent variable name

`depth_analysis` depth's analysis

`...` extra parameters

`association_analysis` *Association analysis of SEMseeker's results*

Description

Association analysis of SEMseeker's results

Usage

```
association_analysis(
  inference_details,
  result_folder,
  maxResources = 90,
  parallel_strategy = "multisession",
  ...
)
```

Arguments

`inference_details`
independent variable: deve essere nalla sample sheet passata a semseeker quando lo abbiamo eseguito la prima volta tipo di regressioni: gaussian, poisson, binomial,quantreg_tau_runs(both as number) eg quantreg_0.25_2000 tipi di test: wilcoxon, stats::t.test, tipi di correlazioni: pearson, kendall, spearman MUTATIONS_* ~ tcdd_mother + exam_age transformation to be applied to dependent variable (mutations and lesions): scale, log, log2, log10, exp, none, quantile_quantiles(as number) eg quantile_3 depth analysis: 1: sample level 2: type level (gene, DMR, cpgisland) (includes 1) 3: genomic area: gene, body, gene tss1550, gene whole, gene tss200, (includes 1 and 2) filter_p_value report after adjusting saves only significant nominal p-value

`result_folder` where semseeker's results are stored, the root folder

`maxResources` percentage of max system's resource to use

`parallel_strategy`
which strategy to use for parallel execution see future vignete: possible values, none, multisession,sequential, multicore, cluster

`...` other options to filter elaborations

`build_data_set_from_geo` *build_data_set_from_geo*

Description

build_data_set_from_geo

Usage

```
build_data_set_from_geo(GEOgse, workingFolder, downloadFiles = 0)
```

Arguments

`GEOgse`	geo accession dataset identification
`workingFolder`	where sample sheet and files will be saved
`downloadFiles`	0 means download all files from Gene Expression Ombibus (GEO), different than zero means how many download

Value

samplesheet, and sample's file saved and samplesheet csv

`compute_qr_beta_boot_p` *Title*

Description

Title

Usage

```
compute_qr_beta_boot_p(sig.formula, tau, localDataFrame)
```

Arguments

`sig.formula`	formula to use for regression application
`tau`	tau to apply the quantile regression
`localDataFrame`	dataframe to apply th regression model

compute_quantreg_beta_boot_np *Title*

Description

Title

Usage

```
compute_quantreg_beta_boot_np(sig.formula, df, tau, lqm_control)
```

Arguments

sig.formula	formula to apply
df	dataframe to use
tau	tau at which apply the wuantile regression
lqm_control	specification of the lqmm package

create_heatmap *create_heatmap load the multiple bed resulting from analysis organized into files and folders per marker and produce a pivot*

Description

create_heatmap load the multiple bed resulting from analysis organized into files and folders per marker and produce a pivot

Usage

```
create_heatmap()
```

Value

nothing

`data_preparation` *Title*

Description

Title

Usage

```
data_preparation(
  family_test,
  transformation,
  tempDataFrame,
  independent_variable,
  g_start,
  dototal,
  covariates,
  depth_analysis
)
```

Arguments

`family_test`	test or regression to apply
`transformation`	transformation to apply to data
`tempDataFrame`	data frame to use for test/regression
`independent_variable`	regressor
`g_start`	starting column of the dataframe
`dototal`	boolean to calculate the total burden test/regression
`covariates`	vector of covariates to be found in the sample sheet
`depth_analysis`	1 only sample, 2 chr, 3 alle genomic areas

`delta_single_sample` *delta_single_sample*

Description

delta_single_sample

Usage

```
delta_single_sample(
  values,
  high_thresholds,
  low_thresholds,
  sample_detail,
  beta_medians,
  probe_features
)
```

Arguments

`values`	values of methylation
`high_thresholds`	highest threshold to use for comparison
`low_thresholds`	lowest threshold to use for comparison
`sample_detail`	details of sample to analyze
`beta_medians`	median to use for calculation
`probe_features`	genomic position of probe_features

Value

summary detail about the analysis

`deltar_single_sample` *delta_single_sample*

Description

delta_single_sample

Usage

```
deltar_single_sample(
  values,
  high_thresholds,
  low_thresholds,
  sample_detail,
  beta_medians,
  probe_features
)
```

Arguments

`values`	values of methylation
`high_thresholds`	highest threshold to use for comparison
`low_thresholds`	lowest threshold to use for comparison
`sample_detail`	details of sample to analyze
`beta_medians`	median to use for calculation
`probe_features`	genomic position of probe_features

Value

summary detail about the analysis

dir_check_and_create *dir_check_and_create*

Description

dir_check_and_create

Usage

```
dir_check_and_create(baseFolder, subFolders)
```

Arguments

baseFolder	folder to look in
subFolders	sub folders to create, complete tree

Value

full path

dump_sample_as_bed_file *given data and colnames dump as bed file*

Description

given data and colnames dump as bed file

Usage

```
dump_sample_as_bed_file(data_to_dump, fileName)
```

Arguments

data_to_dump	data frame to dump into bed file with CHR, START, END
fileName	name of the file to save data in

Value

nothing

glm_model *Title*

Description

Title

Usage

```
glm_model(family_test, tempDataFrame, sig.formula)
```

Arguments

`family_test`	regression model to apply
`tempDataFrame`	data frame to use for the model
`sig.formula`	formula to apply the model

init_env *init ssEnvonment*

Description

init ssEnvonment

Usage

```
init_env(
  result_folder,
  maxResources = 90,
  parallel_strategy = "multicore",
  ...
)
```

Arguments

`result_folder`	where result of semseeker will bestored
`maxResources`	percentage of how many available cores will be used default 90 percent, rounded to the lowest integer
`parallel_strategy`	which strategy to use for parallel executio see future vignete: possibile values, none, multisession,sequential, multicore, cluster
`...`	other options to filter elaborations

Value

the working ssEnvonment

manhattan_plot_per_area *Title*

Description

Title

Usage

```
manhattan_plot_per_area(
  marker,
  figure,
  group,
  subgroup,
  family,
  adjust_method,
  phenotype,
  only_significant_areas = FALSE
)
```

Arguments

marker	investigated marker eg. MUTATIONS, DELTAR, DELTAQ
figure	HYPO, HYPER
group	genomic area (eg. GENE, ISLAND, DMR)
subgroup	sub genomic area (TSS1550), depending on the genomic area
family	fullname of the family used for the association analysis
adjust_method	colnames of the pvalue adjusted to use
phenotype	variable to select from the sample_sheet to use for coloring point
only_significant_areas	TRUE if filter for pvalue < 0.05

mutations_get *mutations_get*

Description

mutations_get

Usage

```
mutations_get(values, figure, thresholds, probe_features, sampleName)
```

Arguments

`values`	values of methylation
`figure`	figure to get Mutaions of HYPO or HYPER methylation
`thresholds`	threshold to use for comparison
`probe_features`	probe_features features probe, chr, start,end
`sampleName`	name of the sample

Value

mutations

`pivot_to_long_format` *Get the pivot in long format instead of wide format*

Description

Get the pivot in long format instead of wide format

Usage

```
pivot_to_long_format(
  marker,
  figure,
  group,
  subgroup,
  phenotype_column,
  sample_sheet,
  areas_selection = NULL
)
```

Arguments

`marker`	marker to filer HYPER, HYPO, BOTH
`figure`	DELTAS, DELTAQ,DELTAR, MUTATIONS
`group`	GENE, DMR ...
`subgroup`	TSS1500 ...
`phenotype_column`	column from the sample sheet to pair to each sample
`sample_sheet`	sample sheet of samples
`areas_selection`	genomic area to select, if NULL all areas will be selected

Value

the pivot in a long format of 3 columnns, the phontype column with name phenotype, the value of the marker and the area investigated

pp_tot *PROBES_CHR_CHR*

Description

Full data set probe_features as defined by Illumina

Usage

```
pp_tot
```

Format

A data frame with five variables: `year`, `sex`, `name`, `n` and `prop` (`n` divided by total number of applicants in that year, which means proportions are of people of that sex with that name born in that year).

PROBES *PROBES_CHR_CHR*

Description

Full data set probe_features as defined by Illumina

Usage

```
PROBES
```

Format

A data frame with five variables: `year`, `sex`, `name`, `n` and `prop` (`n` divided by total number of applicants in that year, which means proportions are of people of that sex with that name born in that year).

PROBES_CHR_CHR *PROBES_CHR_CHR*

Description

Full data set probe_features as defined by Illumina

Usage

```
PROBES_CHR_CHR
```

Format

A data frame with five variables: `year`, `sex`, `name`, `n` and `prop` (`n` divided by total number of applicants in that year, which means proportions are of people of that sex with that name born in that year).

`quantreg_model` *Title*

Description

Title

Usage

```
quantreg_model(
  family_test,
  sig.formula,
  tempDataFrame,
  independent_variable,
  boot_success,
  tests_count
)
```

Arguments

`family_test`	family lqmm, quantreg
`sig.formula`	formula of the model
`tempDataFrame`	data
`independent_variable`	name of regressor
`boot_success`	number of success tests to calculate corrected confidence interval
`tests_count`	count of total executed tests

`quantreg_summary` *Quantile regression result value, confidence interval and pvalue*

Description

Quantile regression result value, confidence interval and pvalue

Usage

```
quantreg_summary(
  boot_vector,
  estimate,
  conf.level,
  boot_success = 0,
  tests_count = 1
)
```

Arguments

`boot_vector`	vector of boot statistc beta regression
`estimate`	beta regression
`conf.level`	confidence intervals alpha level
`boot_success`	number of success respecting the null hypothesis
`tests_count`	how many tests were done

Value

ci and pvalue with BCA method

`range_beta_values` *calculate the range of beta values to define the outlier*

Description

calculate the range of beta values to define the outlier

Usage

```
range_beta_values(populationMatrix, iqrTimes = 3)
```

Arguments

`populationMatrix`	matrix of methylation for the population under calculation
`iqrTimes`	inter quartile ratio used to normalize

Value

methylation matrix as normalized distribution

`read_multiple_bed` *read multiple bed with annotated data as per input parameter*

Description

read multiple bed with annotated data as per input parameter

Usage

```
read_multiple_bed(sample_group, marker, figure)
```

Arguments

`sample_group`	name of the population used to build the data path
`marker`	marker definition used to label folder and files eg MUTATIONS, LESIONS
`figure`	figures like hypo/hyper to built the data path

Value

list of pivot by column identified with column Label and by Sample

semseeker *Calculate stochastic epi mutations from a methylation dataset as outcome report of pivot*

Description

Calculate stochastic epi mutations from a methylation dataset as outcome report of pivot

Usage

```
semseeker(
  sample_sheet,
  methylation_data,
  result_folder,
  bonferroni_threshold = 0.05,
  maxResources = 90,
  iqrTimes = 3,
  parallel_strategy = "multisession",
  ...
)
```

Arguments

`sample_sheet` dataframe with at least a column Sample_ID to identify samples

`methylation_data` matrix of methylation data

`result_folder` where the result will be saved

`bonferroni_threshold` = 0.05 #threshold to define which pValue adjusted to define an epilesion

`maxResources` percentage of how many available cores will be used default 90 percent, rounded to the lowest integer

`iqrTimes` how many times below the first quartile and over the third quartile the interqauartile is "added" to define the outlier

`parallel_strategy` which strategy to use for parallel executio see future vignete: possibile values, none, multisession,sequential, multicore, cluster

`...` other options to filter elaborations

Value

files into the result folder with pivot table and bedgraph.

sort_by_chr_and_start *sort the dataframe using CHR and START sorting column first for CHR and after for START*

Description

sort the dataframe using CHR and START sorting column first for CHR and after for START

Usage

```
sort_by_chr_and_start(dataframe)
```

Arguments

dataframe	dataframe to be sorted

Value

sorted dataframe

test_match_order *Title*

Description

Title

Usage

```
test_match_order(x, y)
```

Arguments

x	vector to compare
y	vector to compare

Value

true if the order matches otherwise is false

`test_model` *Title*

Description

Title

Usage

```
test_model(
  family_test,
  tempDataFrame,
  sig.formula,
  burdenValue,
  independent_variable
)
```

Arguments

`family_test`	which family test to apply
`tempDataFrame`	data frame to use with the test
`sig.formula`	formula to apply
`burdenValue`	burden colon name
`independent_variable`	independent variable for regressor

Index