Package 'semseeker'

March 30, 2023

Type Package

Title Stochastic Epigenetic Mutations SEM Seeker

Version 0.7.6

Author Luigi Corsaro, Davide Gentilini, Lucio Calzari, 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 http://github.com/drake69/semseeker

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

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

RoxygenNote 7.2.3

Suggests pathfindR, GEOquery, gtools, stringi, testthat

Depends R (>= 2.10)

R topics documented:

`analize_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

```
analize_population(
  envir,
  methylation_data,
  sliding_window_size,
  beta_superior_thresholds,
  beta_inferior_thresholds,
  sample_sheet,
  beta_medians,
  bonferroni_threshold = 0.05,
  probe_features
)
```

Arguments

`envir`	semseekere working infos
`methylation_data`	whole matrix of data to analyze.
`sliding_window_size`	size of the sliding widows to compute epilesions default 11 probes.
`beta_superior_thresholds`	data frame to select, from the sample sheet, samples to use as control as study population and as refereces two vectors within the first vector the names of the selection colum and tha second vector the study population selector,
`beta_inferior_thresholds`	name of samplesheet's column to use as control population selector followed by selection value,
`sample_sheet`	name of samplesheet's column to use as control population selector followed by selection value,

beta_medians	name of samplesheet's column to use as control population selector followed by selection value,
bonferroni_threshold	threshold to define which pValue accept for
probe_features	probes detail from 27 to EPIC illumina dataset lesions definition

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

Arguments

envir	environment to get globals
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	probes details to be used

Value

list of lesion count and probes count

`annotate_bed` *takes a bed and its location (build with the details of popuationa nd genomic area) and annoate with detail about genomic area*

Description

takes a bed and its location (build with the details of popuationa nd genomic area) and annoate with detail about genomic area

Usage

```
annotate_bed(
  envir,
  populations,
  figures,
  anomalies,
  groups,
  probes_prefix,
  columnLabel,
  groupingColumnLabel
)
```

Arguments

`envir`	semseekere working infos
`populations`	vector of population to cycle with to build the folder path
`figures`	vector of hyper /hypo to use to build the folder path
`anomalies`	vector of lesions/mutations to use to build the folder path
`groups`	vector of genomic area to cycle and group the annotated data
`probes_prefix`	prefix to use to get the annotated probes dataset
`columnLabel`	label of the column of the genomic area gene, island ,dmr etc..
`groupingColumnLabel`	label of the column of the genomic sub area body, tss1500

Value

original bed with genomic area infos

`apply_stat_model` *Title*

Description

Title

Usage

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

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
logFolder	where to save log file
independent_variable	independent variable name
depth_analysis	depth's analysis
envir	object environment
...	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

```
localDataFrame
```

compute_quantreg_beta_boot_np *Title*

Description

Title

Usage

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

Arguments

```
lqm_control
```

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

Description

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

Usage

```
create_heatmap(
  envir,
  inputBedDataFrame,
  anomalies,
  file_prefix,
  groupColumnLabels
)
```

Arguments

envir	semseekere working infos
inputBedDataFrame	data frame to chart
anomalies	vector of anomalies to manage
file_prefix	main genomic area to char eg: gene
groupColumnLabels	positions of the group coplumn id

Value

list of pivot by column identified with file_prefix and by Sample

---

data_preparation *Title*

Description

Title

Usage

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

Arguments

envir

delta_single_sample *delta_single_sample*

Description

delta_single_sample

Usage

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

Arguments

envir	environment to get globals
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 probes

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

`sig.formula`

init_env *init environment*

Description

init environment

Usage

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

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 environment

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	probes features probe, chr, start,end
sampleName	name of the sample

Value

mutations

`quantreg_model` *Title*

Description

Title

Usage

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

Arguments

`tests_count`

`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
working_data	data to regress
sig.formula	formula for model
tau	quantile to regress at
independent_variable	name of indenpendent variable
lqm_control	controls of lqmm packages

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(
  envir,
  anomalyLabel,
  figureLable,
  probe_features,
  columnLabel,
  populationName,
  groupingColumnLabel
)
```

Arguments

envir	semseekere working infos
anomalyLabel	anomaly definition used to label folder and files eg MUTATIONS, LESIONS
figureLable	figures like hypo/hyper to built the data path
probe_features	features of probe CHR and START and NAME
columnLabel	name of column in the annotation dataset to select genomic area (gene, island etc..)
populationName	name of the population used to build the data path
groupingColumnLabel	name of the genomic sub area

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

independent_variable

Index