Package 'PhosMap'

December 23, 2018

Type Package

Title A Comprehensive R Package For Analyzing Quantitative Phosphoproteomics Data

Version 0.99.12

Author Dongdong Zhan [aut, cre], Mengsha Tong [aut, cre]

Maintainer Dongdong Zhan <ecnuzdd@163.com>

Description PhosMap is a comprehensive R package for analyzing quantitative phosphoproteomics data. Modules in PhosMap were classified into two major categories: (1) data pre-processing and (2) data analysis and presentation. An intact data pre-processing procedure of phosphoproteomics data covered three main steps: merging input files after quality control, mapping phosphorylation sites (p-sites) to the corresponding protein sequence and data normalization. PhosMap incorporated four analysis modules, including clustering and differential expression analysis, time course analysis, kinase activity prediction to find activated/deactivated kinases and motif enrichment analysis.

License GPL (>= 2)

Encoding UTF-8

LazyData FALSE

RoxygenNote 6.0.1

Date 2018-12-23

Depends R (>= 2.10)

Imports graphics, grDevices, stats, utils, stringr, ggseqlogo, samr, limma, e1071, ClueR, Rtsne, glmnet

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

`analysis_deps_anova` *Differential expression analysis using ANOVA*

Description

Differential expression analysis using ANOVA

Usage

```
analysis_deps_anova(expr_data_frame, group, log2_label = FALSE,
  return_padjust = TRUE, adjust_method = "BH")
```

Arguments

`expr_data_frame`	A data frame containing ID and quantification value.
`group`	A factor representing groups.
`log2_label`	A boolean value for representing whether or not the value is logarithmic, the default is FALSE.
`return_padjust`	A boolean value for representing whether or not the pvalue is adjusted, the default is TRUE.
`adjust_method`	Correction method, such as "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr".

Value

A data frame containing ID, log2(FC) and pvalue.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
anova_result <- analysis_deps_anova(
  expr_data_frame, group, log2_label = FALSE,
  return_padjust = TRUE, adjust_method = 'BH'
)

## End(Not run)
```

`analysis_deps_limma` *Differential expression analysis using limma.*

Description

Differential expression analysis using limma.

Usage

```
analysis_deps_limma(expr_data_frame, group, comparison_factor,
  log2_label = FALSE, adjust_method = "BH")
```

Arguments

`expr_data_frame`	A data frame containing ID and quantification value.
`group`	A factor for representing groups.
`comparison_factor`	A vector for comparison factor.
`log2_label`	A boolean value for representing whether or not the value is logarithmic, the default is FALSE.
`adjust_method`	method used to adjust the p-values for multiple testing. See p.adjust for the complete list of options, the default is "BH"

Value

A list containing results from limma analysis.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression \ analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

Examples

```
## Not run:
limma_results_list = analysis_deps_limma(
  expr_data_frame, group, comparison_statement,
  log2_label = FALSE, adjust_method = 'BH'
)

## End(Not run)
```

`analysis_deps_sam` *Differential expression analysis using SAM*

Description

Differential expression analysis using SAM

Usage

```
analysis_deps_sam(expr_data_frame, group, log2_label = FALSE, nperms = 100,
  rand = NULL, minFDR = 0.05, samr_plot = TRUE)
```

Arguments

`expr_data_frame`	A data frame containing ID and quantification value.
`group`	A factor representing groups.
`log2_label`	A boolean value for representing whether or not the value is logarithmic, the default is FALSE.
`nperms`	Number of permutations used to estimate false discovery rates.
`rand`	if specified, the random number generator will be put in a reproducible state.
`minFDR`	A numeric value for filtering significant genes, the default is 0.05.
`samr_plot`	A boolean value for representing whether or not samr graph is plotted.

Value

A list containing results from sam analysis.

Author(s)

Dongdong Zhan and Mengsha Tong

References

R. Tibshirani, G. Chu, T. Hastie and Balasubramanian Narasimhan (2010). samr: SAM: Significance Analysis of Microarrays.\ Rpackage version 1.28. https://CRAN.R-project.org/package=samr

Examples

```
## Not run:
sam_results_list = analysis_deps_sam(
  expr_data_frame, group, log2_label = FALSE,
  nperms = 100, rand = NULL, minFDR = 0.05,samr_plot = T
)

## End(Not run)
```

BRAFi *BRAFi data on quantification proteomics*

Description

Data came from Ressa et al. MS experiments, they performed (phospho)proteomic profiling of the WiDr colorectal cancer cells harboring the BRAF(V600E) mutation after treatment using vemurafenib (BRAF inhibitor, abbr. BRAFi) in a time course at 0, 2, 6, 24, and 48 hour, respectively.

Usage

```
data(BRAFi)
```

Format

An list containning intermediate result from demo.

background_df A data frame for motif enrichment analysis as background.

data_frame_normalization_with_control_no_pair A data frame containning phosphoproteomics data normalized by proteomics data.

foreground_df A data frame for motif enrichment analysis as foreground.

fuzzy_input_df A data frame for time course analysis as input.

merge_df_with_phospho_peptides A merged phosphoproteomics data frame based on peptides files (unique ID).

motif_group_m_ratio_df_mat A matrix for motif profile.

phospho_data_normalization_and_filtering_STY A phosphoproteomics data frame after normalization and filtering.

profiling_data_normalized A proteomics data frame after normalization and filtering.

summary_df_of_unique_proteins_with_sites A data frame that phosphorylation sites had been mapping to protein sequence and eliminated redundancy.

group A factor for experiment group information.

Source

https://www.ebi.ac.uk/pride/archive/projects/PXD007740/

References

Ressa, A., et al. (2018) A System-wide Approach to Monitor Responses to Synergistic BRAF and EGFR Inhibition in Colorectal Cancer Cells, Molecular & cellular proteomics : MCP, 17, 1892-1908.

Examples

```
data(BRAFi)
phospho_data_normalization_and_filtering_STY <- BRAFi$phospho_data_normalization_and_filtering_STY
ID <- paste(
  phospho_data_normalization_and_filtering_STY$GeneSymbol,
  phospho_data_normalization_and_filtering_STY$AA_in_protein,
  sep = '_'
```

```
)
Value <- phospho_data_normalization_and_filtering_STY[,-seq(1,6)]
phospho_data <- data.frame(ID, Value)
phospho_data_top75 = keep_psites_with_max_in_topX(
                                  phospho_data,
                                  percent_of_kept_sites = 0.75
                                )
```

---

check_mea_input *Check input for motif enrichment analysis (mea)*

Description

Check input for motif enrichment analysis (mea)

Usage

```
check_mea_input(foreground, background, center)
```

Arguments

foreground	A vector for AA sequences with fixed length as foreground input.
background	A vector for AA sequences with fixed length as background input.
center	A character for center of k-mer.

Value

A list passing check steps

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
check_result_list <- check_mea_input(
  foreground,
  background,
  center
)

## End(Not run)
```

compute_kses *computing kinase-substrate enrichment score*

Description

computing kinase-substrate enrichment score

Usage

```
compute_kses(substate_vector, regulons_of_kinase,
  substrates_of_kinase_in_exp_count)
```

Arguments

substate_vector
: a vector for substrates with value indentified in current experiment.

regulons_of_kinase
: a vector for substrates of a specific kinase, which identified in current experiment.

substrates_of_kinase_in_exp_count
: a numeric for number in regulons_of_kinase vector.

Value

A numeric or NA for enrichment_score.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Hernandez-Armenta C et al. Benchmarking substrate-based kinase activity inference using phosphoproteomic data[J]. Bioinformatics, 2017, 33(12): 1845-1851.

Examples

```
## Not run:
enrichment_score <- compute_kses(
  substate_vector,
  regulons_of_kinase,
  substrates_of_kinase_in_exp_count
)

## End(Not run)
```

construct_pwm *Construct position weight matrix*

Description

Construct position weight matrix

Usage

```
construct_pwm(sequences, width, frequency_flag = TRUE)
```

Arguments

sequences	A vector for aligned sequences with fixed length.
width	A numeric for specific k-mer.
frequency_flag	A boolean for showing real frequency or frequency probability, the default is TRUE for showing real frequency.

Value

A position weight matrix.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Wagih O, Sugiyama N, Ishihama Y, et al. Uncovering phosphorylation-based specificities through functional interaction networks[J]. Molecular & Cellular Proteomics, 2016, 15(1): 236-245.

Examples

```
## Not run:
pwm <- construct_pwm(
  sequences,
  width,
  frequency_flag = TRUE
)

## End(Not run)
```

`extract_psites_score` *Create R code to call python for parsering mascot xml.*

Description

Extract the confidence probability of phosphorylation sites (psites) from mascot xml. One file containing experiment codes and one folder containing mascot xml as input, the another folder is required as output. Python is required and the corresponding xml package is also required.

Usage

```
extract_psites_score(phosphorylation_exp_design_info_file_path, mascot_xml_dir,
  mascot_txt_dir)
```

Arguments

`phosphorylation_exp_design_info_file_path`	A string representing the file path of experiment code, for examples: experiment_code.txt
`mascot_xml_dir`	A folder containing identification xml files searched by Mascot as input, for examples: Exp020901_F1_R1.xml
`mascot_txt_dir`	A folder used for saving files which contains the confidence of phosphorylation sites, for examples: Exp020901_F1_R1.txt

Value

A series of output file saved in the mascot_txt_dir

Examples

```
## Not run:
extract_psites_score(
  phosphorylation_exp_design_info_file_path,
  mascot_xml_dir, mascot_txt_dir
)

## End(Not run)
```

`fore_seq_to_motif` *Convert the list that consists of motifs and the corresponding sequences to data frame.*

Description

Convert the list that consists of motifs and the corresponding sequences to data frame.

Usage

```
fore_seq_to_motif(foreground_sequences_mapped_to_motifs)
```

Arguments

foreground_sequences_mapped_to_motifs
: A list that consists of motifs and the corresponding sequences.

Value

A data frame that consist of aligned sequences and the corresponding motifs.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
df <- fore_seq_to_motif(
  foreground_sequences_mapped_to_motifs
)

## End(Not run)
```

---

formatted_output_mef_results

Output formatted sequences in foreground that are mapped to specific motif.

Description

Output formatted sequences in foreground that are mapped to specific motif.

Usage

```
formatted_output_mef_results(foreground_sequences_mapped_to_motifs)
```

Arguments

foreground_sequences_mapped_to_motifs
: A list that consists of motifs and their corresponding aligned sequences from foreground.

Value

A data frame that motifs and their corresponding aligned sequences from foreground.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
formatted_output_df <- formatted_output_mef_results(
  foreground_sequences_mapped_to_motifs
)

## End(Not run)
```

`generate_psites_score_file` *Generate peptide identification files with psites scores.*

Description

Based on mascot txt files with psites and peptide identification files downloaded from Firmiana, the file including phosphorylation modifications is generated.

Usage

```
generate_psites_score_file(mascot_txt_dir, firmiana_peptide_dir,
  psites_score_dir)
```

Arguments

`mascot_txt_dir`	A folder containing identification xml files with psites scores as input.
`firmiana_peptide_dir`	A folder containing identification txt file downloaded from Firmiana as input.
`psites_score_dir`	A folder used for saving files of peptide identification file with psites scores

Value

A series of output files saved in the psites_score_dir

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
generate_psites_score_file(mascot_txt_dir, firmiana_peptide_dir, psites_score_dir)

## End(Not run)
```

`get_aligned_seq_for_mea` *Taking S/T/Y as the center, align sequence to fasta library by specific length.*

Description

Taking S/T/Y as the center, align sequence to fasta library by specific length.

Usage

```
get_aligned_seq_for_mea(GI, Sequence, AA_in_protein, fixed_length, species)
```

Arguments

`GI`	A vector for protein gi.
`Sequence`	A vector for sequence of peptide.
`AA_in_protein`	A vector for the location of S/T/Y in sequence of protein.
`fixed_length`	A numeric value for aligned sequence,the default is 15.
`species`	A string for that the alignment is based on which species, the options are 'human' and 'mouse'

Value

A data frame containing GI, Sequence, AA_in_protein, aligned_seq.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Hadley Wickham (2018). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.3.0.\ https://CRAN.R-project.org/package=stringr.

Examples

```
## Not run:
aligned_sequence_df_based_on_fasta_library =
 get_aligned_seq_for_mea(
   GI,
   Sequence,
   AA_in_protein,
   fixed_length,
   species
 )

## End(Not run)
```

`get_colors_for_discrete_value` *Generate custom colors from discrete values for heatmap.*

Description

Generate custom colors from discrete values for heatmap.

Usage

```
get_colors_for_discrete_value(color_intervals_list, value_intervals_list)
```

Arguments

`color_intervals_list`	a list for building color intervals.
`value_intervals_list`	a list for building value intervals.

Value

A vectors containing colors distribution.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
value_intervals_list = list(
seq(-4, -2, 0.2),
seq(-2, -1, 0.2),
seq(-1, 1, 0.2),
seq(1, 2, 0.2),
seq(2, 4, 0.2)
)
color_intervals_list = list(
  c('blue', '#33CCFF'),
  c('#33CCFF', 'green'),
  c('green', 'white', '#FF6600'),
  c('#FF6600', 'red'),
  c('red', 'firebrick')
)
colors = get_colors_for_discrete_value(
  color_intervals_list,
  value_intervals_list
)

## End(Not run)
```

get_combined_data_frame *Get a data frame mapped GI number to Gene Symbol.*

Description

This is an intermediate file and a dataframe with Gene Symbol is exported. Based on a library file consisting of mapping relationships about Gene Symbol, GeneID and GI, a new dataframe with Sequence, GI, Modification, Gene Symbol, Area and PSMs,is contructed.

Usage

```
get_combined_data_frame(merge_df_with_phospho_peptides, species = "human")
```

Arguments

merge_df_with_phospho_peptides	A dataframe consisting of IDs (Sequence_GI_Psite) and Area values.
species	A string, the options is human and mouse, the default is human.

Value

A dataframe with Sequence, GI, Modification, Gene Symbol, Area values and PSMs

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
combinated_df_with_mapped_gene_symbol = get_combined_data_frame(
  merge_df_with_phospho_peptides
)

## End(Not run)
```

get_df_with_AAs_i *Get data frame of amino acide sequences for a protein.*

Description

Get data frame of amino acide sequences for a protein.

Usage

```
get_df_with_AAs_i(unique_proteins, i, id_data_only_peptide2gi,
  proteins_in_id_data_only_peptide2gi, sequences_in_id_data_only_peptide2gi,
  modification_index_in_protein_seq_list)
```

Arguments

unique_proteins	a vector for unique proteins.
i	the ith unique proteins.
id_data_only_peptide2gi	a data frame for peptides with protein gi.
proteins_in_id_data_only_peptide2gi	a vector for proteins with only protein gi.
sequences_in_id_data_only_peptide2gi	a vector for peptides with only protein gi.
modification_index_in_protein_seq_list	a list for the index of modifications in protein sequence.

Value

A data frame with sequences for a protein.

Author(s)

Dongdong Zhan and Mengsha Tong

---

get_file_info_from_dir *Get data lists from and corresponding file ids.*

Description

Read batch files (.txt or .csv) from specific directory.

Usage

```
get_file_info_from_dir(specific_dir, experiment_ID)
```

Arguments

specific_dir	A folder containing files as input.
experiment_ID	A vector containing experiment codes as input

Value

A list containing data from files and corresponding file ids

Examples

```
## Not run:
result_list = get_file_info_from_dir(
  specific_dir,
  experiment_ID
)

## End(Not run)
```

get_file_suffix *Get sufffix of input file.*

Description

Get sufffix of input file.

Usage

```
get_file_suffix(file_name)
```

Arguments

file_name	A string for file name.

Value

A string for file format.

get_filtered_df *Get filtered data frame based on the Mascot and reference file.*

Description

Get filtered data frame based on the Mascot and reference file.

Usage

```
get_filtered_df(mascotfileName, refFileName)
```

Arguments

mascotfileName	a string for mascot name as input.
refFileName	a string for reference file name.

Value

A filtered data frame

Author(s)

Dongdong Zhan and Mengsha Tong

`get_foreground_df_to_motifs` *Get filtered foreground data frame that its aligned sequences with specific motif.*

Description

Get filtered foreground data frame that its aligned sequences with specific motif.

Usage

```
get_foreground_df_to_motifs(foreground_sequences_mapped_to_motifs, foreground,
  foreground_df)
```

Arguments

`foreground_sequences_mapped_to_motifs`	A list that consists of motifs and its corresponding aligned sequences.
`foreground`	A vector for aligned sequences.
`foreground_df`	A data frame from the initial foreground data frame.

Value

A data frame that its aligned sequences with specific motif.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
foreground_df_mapped_to_motifs <- get_foreground_df_to_motifs(
  foreground_sequences_mapped_to_motifs,
  foreground,
  foreground_df
)

## End(Not run)
```

`get_foreground_seq_to_motifs` *Get motifs and their corresponding aligned sequences form from foreground.*

Description

Get motifs and their corresponding aligned sequences form from foreground.

Usage

```
get_foreground_seq_to_motifs(motifs_list, foreground)
```

Arguments

motifs_list	A list from motif enrichment analysis.
foreground	A vector for aligned sequences.

Value

A list containing motifs and the corresponding sequences from foreground.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Hadley Wickham (2018). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.3.0.\ https://CRAN.R-project.org/package=stringr.

Examples

```
## Not run:
mots_match_list = get_foreground_seq_to_motifs(
  motifs_list,
  foreground
)

## End(Not run)
```

---

get_global_background_df

Get background data frame (fasta library from Refseq).

Description

Get background data frame (fasta library from Refseq).

Usage

```
get_global_background_df(species)
```

Arguments

species	A string for that the alignment is based on which species, the options are 'human' and 'mouse'.

Value

A data frame of background

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
background_df = get_global_background_df(species)

## End(Not run)
```

`get_ka_by_mean_or_mlr` *Computing kinase activity using mean value and multiple linear regression (ridge regression) except KSEA*

Description

Computing kinase activity using mean value and multiple linear regression (ridge regression) except KSEA

Usage

```
get_ka_by_mean_or_mlr(ptypes_data, species = "human", log2_label = FALSE,
  method = "mean")
```

Arguments

`ptypes_data`	A data frame of phosphorylation data after normalization.
`species`	A string representing the species of imported data, the options are human, mouse and rat.
`log2_label`	A boolean value representing whether data is logarithmetics, the default is FALSE.
`method`	A string for the method to compute kinase activity, the options are 'mean' and 'mlr' (multiple linear regression), the default is mean.

Value

A data frame that consists of kinase, psite, substrate, counting byond ratio_cutoff and corresponding original value.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
kinase_activity_df <- get_ka_by_mean_or_mlr(
  ptypes_data,
  species = 'human',
  log2_label = TRUE,
  method = 'mean'
)
```

```
## End(Not run)
```

`get_ksea_regulons_info` *Get informational data frame by combining result from all experiments*

Description

Get informational data frame by combining result from all experiments

Usage

```
get_ksea_regulons_info(ksea_regulons, ksea_trans_list, ksea_x_list,
  ptypes_data_ratio_colnames)
```

Arguments

`ksea_regulons`	A kinase vector from all experiments.
`ksea_trans_list`	A list that consits of regulation direction of kinase from each experiment by ksea.
`ksea_x_list`	A list that consits of sepecific information from each experiment by ksea, like regulation direction, pvalue and activity etc..
`ptypes_data_ratio_colnames`	A vector that consists of column names from experiments.

Value

A data frame containing sepecific information of all experiments from ksea results, like regulation direction, pvalue and activity etc..

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
information_dataframe = get_ksea_regulons_info(
  ksea_regulons,
  ksea_trans_list,
  ksea_x_list,
  ptypes_data_ratio_colnames
)

## End(Not run)
```

`get_ksea_result_list` *Kinase activity analysis based on known and predicted relationship of kinase-substrate*

Description

Kinase activity analysis based on known and predicted relationship of kinase-substrate

Usage

```
get_ksea_result_list(ptypes_data_ratio_in_sigle_exp, ID,
  kinase_substrate_regulation_relationship, ksea_activity_i_pvalue = 0.05)
```

Arguments

`ptypes_data_ratio_in_sigle_exp`
: A quantifcation vector from single experiment.

`ID`
: A phosporylation ID vector like VIM_S56 (GeneSymbol_psite).

`kinase_substrate_regulation_relationship`
: A data frame contanning relationship of kinase-substrate that consists of "kinase", "substrate", "site", "sequence" and "predicted" columns.

`ksea_activity_i_pvalue`
: A cutoff used for filtering significant activities computed from KSEA.

Value

A list containing results from ksea.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
ksea_result_list = get_ksea_result_list(
  ptypes_data_ratio_in_sigle_exp,
  ID,
  kinase_substrate,
  0.05
)

## End(Not run)
```

get_kses *computing kinase-substrate enrichment significance (pvalue)*

Description

computing kinase-substrate enrichment significance (pvalue)

Usage

```
get_kses(substate_vector, regulons_of_kinase, trial = 1000)
```

Arguments

substate_vector	a vector for substrates with value indentified in current experiment.
regulons_of_kinase	a vector for substrates of a specific kinase, which identified in current experiment.
trial	a numeric for the number of random samples, the default is 1000.

Value

A list for expected enrichment score and its significance

Author(s)

Dongdong Zhan and Mengsha Tong

References

Hernandez-Armenta C et al. Benchmarking substrate-based kinase activity inference using phosphoproteomic data[J]. Bioinformatics, 2017, 33(12): 1845-1851.

Examples

```
## Not run:
pvalue <- get_kses(
  substate_vector,
  regulons_of_kinase,
  1000
)

## End(Not run)
```

`get_list_with_filted_sites` *Filter phosphorylation sites.*

Description

Filter phosphorylation sites by extracting all peptides with ion score>=20 and FDR<0.01 from Firmiana and having psites scores.

Usage

```
get_list_with_filted_sites(peptide_id, files, files_site_score, qc, min_score,
  min_FDR)
```

Arguments

`peptide_id`	A vector containing experiment ids as input.
`files`	A data list containing peptides identificaton from Firmiana as input.
`files_site_score`	A data list containing psites scores extracted from mascot xml. The default is NULL, which represents no QC file.
`qc`	A boolean value representing whether it has QC files. The default is True.
`min_score`	A numeric for the minimum score of credible peptides, the default is 20 for Mascot ion score.
`min_FDR`	A numeric for the minimum FDR of credible peptides, the default is 0.01.

Value

A list containing peptides dataframe with area values and psm, IDs with mergered sequences, gi and sites, new peptides dataframe combined previous peptides dataframe and IDs.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
result_list_with_filtered_sites = get_list_with_filted_sites(
  peptide.id,
  files,
  files_site_score
)

## End(Not run)
```

get_merged_phospho_df *Get merged data frame with phospho-peptides.*

Description

Get merged data frame with phospho-peptides.

Usage

```
get_merged_phospho_df(peptide_id, peptide_df_with_area_psm_list,
  ID_of_seq_gi_site_list, ID_DF_list)
```

Arguments

peptide_id	a vector for peptide ID.
peptide_df_with_area_psm_list	a list for peptides with area and PSM.
ID_of_seq_gi_site_list	a list for peptides ID with sequence, gi and site.
ID_DF_list	a list for ID and value.

Value

a merged data frame with phospho-peptides.

get_modification_index *Get index of modifications in protein sequence.*

Description

Get index of modifications in protein sequence.

Usage

```
get_modification_index(id_data_only_peptide2gi, gi_fasta)
```

Arguments

id_data_only_peptide2gi	a data frame for peptides with protein gi.
gi_fasta	a fasta data for a specific species.

Value

A vector for index of modifications in protein sequence.

Author(s)

Dongdong Zhan and Mengsha Tong

get_motifs_list *Motif enrichment using rmotifx.*

Description

Motif enrichment using rmotifx.

Usage

```
get_motifs_list(foreground, background, center_vector, motifx_pvalue)
```

Arguments

foreground	A vector for aligned sequences as the foreground input.
background	A vector for aligned sequences as the background input.
center_vector	A vector for aligned centers.
motifx_pvalue	A numeric value for selecting motifs that meets the minimum cutoff.

Value

A list from results of motif enrichment.

Author(s)

Dongdong Zhan and Mengsha Tong

References

Omar Wagih (2014). rmotifx: An iterative statistical approach to the discovery of biological sequence motifs. R package version 1.0.

Examples

```
## Not run:
motifs_list = get_motifs_list(foreground, background, center_vector, motifx_pvalue)

## End(Not run)
```

get_motif_analysis_summary *Get summary result of motif analysis for specific input*

Description

Get summary result of motif analysis for specific input

Usage

```
get_motif_analysis_summary(foreground, background, center = "S",
  min_sequence_count = 1, min_pvalue = 0.01)
```

Arguments

foreground	A vector for AA sequences with fixed length as foreground input.
background	A vector for AA sequences with fixed length as background input.
center	A character for center of k-mer.
min_sequence_count	A numeric for the minimum sequence number assigned to a motif.
min_pvalue	A numeric for the minimum pvalue for found motif.

Value

A list for summary result of motif analysis

Author(s)

Dongdong Zhan and Mengsha Tong

References

Omar Wagih (2014). rmotifx: An iterative statistical approach to the discovery of biological sequence motifs. R package version 1.0.

Examples

```
## Not run:
summry_list <- get_motif_analysis_summary(
  foreground,
  background,
  center,
  min_sequence_count,
  min_pvalue
)

## End(Not run)
```

---

get_normalized_data_FOT5

Normailization on basis of sum

Description

Normailization on basis of sum

Usage

```
get_normalized_data_FOT5(data_frame, experiment_code_file_path)
```

Arguments

data_frame	A data frame containing IDs and values merged from multi-experiments as input.
experiment_code_file_path	A file path of storing experiment codes as input. The experiment codes are required to keep pace with column names of Values.

Value

A data frame after normalization

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
data_frame_normalization = get_normalized_data_FOT5(
  data_frame,
  experiment_code_file_path
)

## End(Not run)
```

---

get_normalized_data_of_psites

To normalize data and filter data only including phosphorylation sites.

Description

To normalize data and filter data only including phosphorylation sites.

Usage

```
get_normalized_data_of_psites(data_frame, experiment_code_file_path)
```

Arguments

data_frame	A data frame containing IDs and quantification values merged from multi-experiments as input.
experiment_code_file_path	A file path of storing experiment codes as input. The experiment codes are required to keep pace with column names of Value.

Value

A data frame after normalization and filtering (x 1e5)

Examples

```
## Not run:
ptypes_df = get_normalized_data_of_psites(data_frame, experiment_code_file_path)

## End(Not run)
```

`get_substrate_expr_df` *Get a data frame only containing kinase inferred by KSEA*

Description

Get a data frame only containing kinase inferred by KSEA

Usage

```
get_substrate_expr_df(ID, kinase_substrate_regulation_relationship,
  ksea_regulons, ptypes_data_ratio, ratio_cutoff = 3)
```

Arguments

`ID`	A phosporylation ID vector like VIM_S56 (GeneSymbol_psite).
`kinase_substrate_regulation_relationship`	A data frame contanning relationship of kinase-substrate that consists of "kinase", "substrate", "site", "sequence" and "predicted" columns.
`ksea_regulons`	A kinase vector from ksea
`ptypes_data_ratio`	A data frame that the ratio of phosphorylation and profiling data
`ratio_cutoff`	A cutoff that depicts quantification changes at phosphorylation level relative to profiling level, the default is 3.

Value

A data frame that consists of kinase, psite, substrate, counting byond ratio_cutoff and corresponding original value.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
kinase_site_substrate_original_ratio_df <- get_substrate_expr_df(
  ID,
  kinase_substrate_regulation_relationship,
  ksea_regulons,
  ptypes_data_ratio,
  ratio_cutoff = 3
)

## End(Not run)
```

`get_summary_from_ksea` *Get a data frame only containing inforamtion of kinase inferred by KSEA*

Description

Get a data frame only containing inforamtion of kinase inferred by KSEA

Usage

```
get_summary_from_ksea(ptypes_data, species = "human", log2_label = TRUE,
  ratio_cutoff = 3)
```

Arguments

`ptypes_data`	A data frame of phosphorylation data after normalization.
`species`	A string representing the species of imported data, the options are human, mouse and rat.
`log2_label`	A boolean value representing whether data is logarithmetics, the default is FALSE.
`ratio_cutoff`	A cutoff that depicts quantification changes at phosphorylation level relative to profiling level, the default is 3.

Value

A data frame that consists of kinase, psite, substrate, counting byond ratio_cutoff and corresponding original value.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
summary_df_list_from_ksea = get_summary_from_ksea(
  ptypes_data,
  species = 'human',
  log2_label = TRUE,
  ratio_cutoff = 3
)

## End(Not run)
```

get_summary_with_unique_sites *Assign psites to protein sequence.*

Description

Construct the data frame with unique phosphorylation site for each protein sequence and eliminate redundancy.

Usage

```
get_summary_with_unique_sites(combinated_df_with_mapped_gene_symbol,
  species = "human")
```

Arguments

combinated_df_with_mapped_gene_symbol	A dataframe with Sequence, GI, Modification, Gene Symbol, Area and PSMs as input.
species	A string, the options is human and mouse, the default is human.

Value

A dataframe that all redundant psites are assigned to protein sequence.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
summary_df_of_unique_proteins_with_sites = get_summary_with_unique_sites(
  combinated_df_with_mapped_gene_symbol
)

## End(Not run)
```

get_unique_AAs_i_df *Get data frame without redundancy.*

Description

Get data frame without redundancy.

Usage

```
get_unique_AAs_i_df(df_with_AAs_i)
```

Arguments

`df_with_AAs_i` a data frame for peptides of the ith protein.

Value

A data frame with sites in unique protein.

Author(s)

Dongdong Zhan and Mengsha Tong

`hello` *Hello, world!*

Description

This is an example function named 'hello' which prints 'Hello, world!' and 'Welcome to PhosMap!'.

Usage

```
hello()
```

Details

Some useful keyboard shortcuts for package authoring: Build and Reload Package: 'Ctrl + Shift + B' Check Package: 'Ctrl + Shift + E' Test Package: 'Ctrl + Shift + T'

Author(s)

Dongdong Zhan and Mengsha Tong

`keep_psites_with_max_in_topX` *Keep psites whose row maximum is top N (percentage).*

Description

Compute row maximum each psites, sort row maximum in decreasing order and keep top N (percentage).

Usage

```
keep_psites_with_max_in_topX(phospho_data, percent_of_kept_sites = 3/4)
```

Arguments

`phospho_data` A data frame of phospho-data.

`percent_of_kept_sites` A numeric value representing a cutoff used for filter psites. The default is 3/4.

Value

A data frame meeting specific cutoff.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
phospho_data_meet_percent = keep_psites_with_max_in_topX(
  phospho_data,
 percent_of_kept_sites = 3/4
)

## End(Not run)
```

`mea_based_on_background` *Motif enrichment based on global background (fasta library from Refseq).*

Description

Motif enrichment based on global background (fasta library from Refseq).

Usage

```
mea_based_on_background(foreground, AA_in_protein, background, motifx_pvalue)
```

Arguments

`foreground`	A vector for aligned sequence of foreground.
`AA_in_protein`	A vector for the location of S/T/Y in sequence of protein.
`background`	A vector for aligned sequence of background.
`motifx_pvalue`	A numeric value for selecting motifs that meets the minimum cutoff.

Value

A list containing motifs and the corresponding sequences

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
motifs_list = mea_based_on_background(
  foreground,
  AA_in_protein,
  background,
  motifx_pvalue
)

## End(Not run)
```

`merge_profiling_file_from_Firmiana` *Merge profiling files downloaded from Firmiana.*

Description

Filter data based on US (Unique and Ionscore > 20) peptide counts then merge profiling files.

Usage

```
merge_profiling_file_from_Firmiana(firmiana_gene_dir, US_cutoff = 1,
  experiment_gene_file_path)
```

Arguments

`firmiana_gene_dir`	a folder containing gene identification results as input.
`US_cutoff`	a numerical value as a cutoff to filter data, the default is 1.
`experiment_gene_file_path`	a file path for storing experiemnt design of proteomics data.

Value

A merged data frame after filtering (US_cutoff) and replacing NAs to zeros.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
merged_df = merge_profiling_file_from_Firmiana(firmiana_gene_dir, US_cutoff = 1)

## End(Not run)
```

motif_data_frame_to_sequence

Convert data frame of motif to the sequence pattern

Description

Convert data frame of motif to the sequence pattern

Usage

```
motif_data_frame_to_sequence(motif_data_frame, center, width)
```

Arguments

motif_data_frame	A data frame with two columns including amino acid and index on sequence with fixed length.
center	A character for center of k-mer.
width	A numeric for specific k-mer.

Value

A string for motif pattern

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
motif_pattern <- motif_data_frame_to_sequence(
  motif_data_frame,
  center,
  width
)

## End(Not run)
```

normalize_nopair_ctrl_by_col

For data without pairs but with control, normalize them to the control.

Description

For data without pairs but with control, normalize them to the control.

Usage

```
normalize_nopair_ctrl_by_col(data_frame, experiment_design_file, control_label)
```

Arguments

data_frame	a data frame as input.
experiment_design_file	a data frame for design of experiment.
control_label	a string for a control.

Value

A data frame after normalization.

---

normalize_nopair_noctrl_by_colmed

For data without pairs and control, normalize them to the median.

Description

For data without pairs and control, normalize them to the median.

Usage

```
normalize_nopair_noctrl_by_colmed(data_frame)
```

Arguments

data_frame	a data frame as input.

Value

A data frame after normalization.

---

normalize_phos_data_to_profiling

Normalize phospho-data to profiling

Description

Normalize phospho-data to profiling

Usage

```
normalize_phos_data_to_profiling(phospho_data_normalized,
  profiling_data_normalized, phosphorylation_exp_design_info_file_path,
  profiling_exp_design_info_file_path, control_label = NA,
  pair_flag = FALSE)
```

Arguments

phospho_data_normalized	A data frame of phospho-data after normalization
profiling_data_normalized	A data frame of profiling after normalization
phosphorylation_exp_design_info_file_path	A file path about phosphorylation experiment design, it has 2 kinds of file configuration as follows: 1. experiment_design_noPair.txt must contain columns of Experiment_Code, Group. 2. experiment_design_Pair.txt must contain columns of Experiment_Code, Group, and Pair. (Pair: 1 -> case, -1 -> control)
profiling_exp_design_info_file_path	A file path about profiling experiment design, it has 2 kinds of file configuration as same as phosphorylation_exp_design_info_file_path.
control_label	A string represents label of control group. The default is NA which shows no control group.
pair_flag	A boolean value represents whether experiments have pairs. The default is FALSE which shows no pairs.

Value

A data frame which comes from results that phospho-data is normalizated base on the abundance of proteins in the profiling experiments.

Examples

```
## Not run:
df_phospho_Value_vs_profiling = normalize_phos_data_to_profiling(
  phospho_data_normalized,
  profiling_data_normalized,
  experiment_design_file_path,
  control_label = NA,
  pair_flag = FALSE
)

## End(Not run)
```

---

normalize_to_Pair *For data with pairs, normalize them to the sample with flag eaqul to -1.*

Description

For data with pairs, normalize them to the sample with flag eaqul to -1.

Usage

```
normalize_to_Pair(data_frame, experiment_design_file)
```

Arguments

data_frame	a data frame as input.
experiment_design_file	a data frame for design of experiment.

Value

A data frame after normalization.

---

plot_seqlogo *Plot sequence logo based on list that consist of motifs and sequences.*

Description

Plot sequence logo based on list that consist of motifs and sequences.

Usage

```
plot_seqlogo(base_dir, foreground_sequences_mapped_to_motifs,
  plot_min_seqs = 5)
```

Arguments

base_dir	A path used for saving plots.
foreground_sequences_mapped_to_motifs	A list that consist of motifs and sequences.
plot_min_seqs	A numeric value for cutoff, sequences of motifs greater than the cutoff are plotted, the default is 5.

Author(s)

Dongdong Zhan and Mengsha Tong

References

(1) Omar Wagih (2017). ggseqlogo: A 'ggplot2' Extension for Drawing Publication-Ready Sequence Logos. R package version 0.1.\ https://github.com/omarwagih/ggseqlogo; (2) Hadley Wickham (2018). stringr: Simple, Consistent Wrappers for Common String Operations. \ R package version 1.3.0. https://CRAN.R-project.org/package=stringr

Examples

```
## Not run:
plot_seqlogo(base_dir, foreground_sequences_mapped_to_motifs, plot_min_seqs = 50)

## End(Not run)
```

pre_process_filter_psites *Get peptides data frame passed phosphorylation sites quality control.*

Description

Filter phosphorylation sites by extracting all peptides with ion score>=20 and FDR<0.01 from Firmiana and having psites scores. Generate new IDs consisting of sequence, gi, psite. Quantification values containing area and psm.

Usage

```
pre_process_filter_psites(firmiana_peptide_dir, psites_score_dir,
  phospho_experiment_design_file_path, qc, min_score = 20, min_FDR = 0.01)
```

Arguments

firmiana_peptide_dir	A folder containing peptide identification files from Firmiana as input.
psites_score_dir	A folder containing psites scores files extracted from mascot xml as input.
phospho_experiment_design_file_path	A string representing the path of phospho-experiment design file as input.
qc	A boolean value representing whether it has QC files. The default is True.
min_score	A numeric for the minimum score of credible peptides, the default is 20 for Mascot ion score.
min_FDR	A numeric for the minimum FDR of credible peptides, the default is 0.01.

Value

A merged data frame containing sequence, gi, psite, area and psm.

Examples

```
## Not run:
merge_df_with_phospho_peptides = pre_process_filter_psites(
  firmiana_peptide_dir,
  psites_score_dir
)

## End(Not run)
```

seach_motif_pattern *Convert data frame of motif to the sequence pattern*

Description

Convert data frame of motif to the sequence pattern

Usage

```
seach_motif_pattern(foreground_sequence, background_sequence,
  min_sequence_count = 1, min_pvalue = 0.01, center = "S", width)
```

Arguments

foreground_sequence	A vector for AA sequences with fixed length as foreground input.
background_sequence	A vector for AA sequences with fixed length as background input.
min_sequence_count	A numeric for the minimum sequence number assigned to a motif.
min_pvalue	A numeric for the minimum pvalue for found motif.
center	A character for center of k-mer.
width	A numeric for specific k-mer.

Value

A list for information summary of searching mortif

Author(s)

Dongdong Zhan and Mengsha Tong

References

Omar Wagih (2014). rmotifx: An iterative statistical approach to the discovery of biological sequence motifs. R package version 1.0.

Examples

```
## Not run:
seach_motif_pattern(
    foreground_sequence,,
    background_sequence,
    min_sequence_count = 1,
    min_pvalue = 0.01,
    center = 'S',
    width = 15
)

## End(Not run)
```

`visualization_deps_with_scatter` *Visualize differentially expressed results with scatter*

Description

Visualize differentially expressed results with scatter

Usage

```
visualization_deps_with_scatter(deps_data, minFC = 2, minPvalue = 0.05,
  main = "Differentially expressed proteins", show_text = FALSE,
  min_up_text = 15, min_down_text = 15)
```

Arguments

`deps_data`	a data frame containing ID, logFC and pvalue.
`minFC`	a numeric for the minimum fold change.
`minPvalue`	a numeric for the significance cutoff.
`main`	an overall title for the plot.
`show_text`	a boolean value representing whether or not the text is showed, the default is FALSE.
`min_up_text`	cutoff value for showing up-IDs. Only IDs with lower than min_up_text are showed.
`min_down_text`	cutoff value for showing down-IDs. Only IDs with lower than min_down_text are showed.

Value

A scatter plot for showing differentially expressed results.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
visualization_deps_with_scatter(
  deps_data,
  minFC = 2,
  minPvalue = 0.05,
  main = 'Differentially expressed proteins',
  show_text = FALSE,
  min_up_text = 15,
  min_down_text = 15
)

## End(Not run)
```

`visualization_fuzzycluster` *Visualize results from fuzzy clusters with line chart*

Description

Visualize results from fuzzy clusters with line chart

Usage

```
visualization_fuzzycluster(input_data, group, group_levels, k_cluster,
  iteration = 100, mfrow = c(3, 3), min_mem = 0.1, plot = TRUE)
```

Arguments

`input_data`	a data frame containing ID and expression profile.
`group`	a factor for representing groups.
`group_levels`	a factor levels for group.
`k_cluster`	number of clusters fuzzy cluster.
`iteration`	a numeric value for interation, the defualt is 100.
`mfrow`	a vector containing 2 elements for controling the subplots in graphic window, the default is mfrow = c(3,3)
`min_mem`	cutoff value for membership. Only results with greater than min_mem are showed.
`plot`	a boolean value for deciding whether ploting, the default is TRUE.

Value

A lines chart with fuzzy degree.

Author(s)

Dongdong Zhan and Mengsha Tong

References

(1) David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the \ Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8.https://CRAN.R-project.org/package=e1071 \ (2) Pengyi Yang (2018). ClueR: Cluster Evaluation. R package version 1.4. https://CRAN.R-project.org/package=ClueR

Examples

```
## Not run:
visualization_fuzzycluster(
  input_data,
  group,
  group_levels,
  k_cluster,
```

```
  iteration = 100,
  mfrow = c(3,3),
  min_mem = 0.1,
  plot = TRUE
)

## End(Not run)
```

`visualization_with_simple_pca` *A simple PCA plot.*

Description

A simple PCA plot.

Usage

```
visualization_with_simple_pca(expr_data_frame, main = "Simple PCA",
  point_cex = 2, point_col = "firebrick", point_type = 20, text_cex = 1)
```

Arguments

`expr_data_frame`	A data frame containing ID and quantification value.
`main`	The main title of plot.
`point_cex`	a numerical value for point size.
`point_col`	a color code or name for point color.
`point_type`	point type, see points.
`text_cex`	a numerical value for text size.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
visualization_with_simple_pca(expr_data_frame, main = 'Simple PCA',
point_cex = 2, point_col = 'firebrick', point_type = 20, text_cex = 1)

## End(Not run)
```

`visualization_with_simple_tsne` *A simple t-SNE plot.*

Description

A simple t-SNE plot.

Usage

```
visualization_with_simple_tsne(expr_data_frame, group, main = "Simple t-SNE",
  perplexity = 10)
```

Arguments

`expr_data_frame`	A data frame containing ID and quantification value.
`group`	A factor for group information.
`main`	The main title of plot.
`perplexity`	A numerical value for perplexity, the default is 10.

Author(s)

Dongdong Zhan and Mengsha Tong

Examples

```
## Not run:
visualization_with_simple_tsne(
  expr_data_frame,
  group,
  main = 'Simple t-SNE',
  perplexity = 10
)

## End(Not run)
```

`write_csv_pep_seq_conf` *Write data to specific direction with CSV format.*

Description

Write data to specific direction with CSV format.

Usage

```
write_csv_pep_seq_conf(expName, outputName, mascotfileNames, refFileName)
```

Arguments

`expName`	a string for experiment name as input.
`outputName`	a string for experiment name as output.
`mascotfileNames`	a vector for storing mascot file names.
`refFileName`	a string for reference file name.

Author(s)

Dongdong Zhan and Mengsha Tong

Index