Package: CheckSumStats (via r-universe)

October 1, 2024

Title CheckSumStats

Version 0.0.0.9000

Description CheckSumStats is an R package for checking the accuracy of meta- and summary-data from genome-wide association studies (GWAS) prior to their use in post-GWAS applications. For example, the package provides tools for checking that the reported effect allele and effect allele frequency columns are correct. It also checks for possible issues in the reported effect sizes that might introduce bias into downstream analyses.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.2.3

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Imports gwasrapidd, utils, stats, ggplot2, grid, gridExtra, cowplot, grDevices, ieugwasr, knitr, biomaRt, purrr, dplyr, tibble, magrittr, curl, plyr, ggpubr

Remotes bioc::release/biomaRt, ramiromagno/gwasrapidd

Depends R (>= 3.5.0)

Repository https://mrcieu.r-universe.dev

RemoteUrl https://github.com/MRCIEU/CheckSumStats

RemoteRef HEAD

RemoteSha ba670767d42a9f6041b65e4a155b64a3ac096434

Contents

`ara_test_dat` *A example dataset of genetic summary data for arachidonic acid*

Description

The dataset contains summary association statistics for 436 SNPs, generated in linear regression models, from a genome-wide association study of arachidonic acid conducted by the CHARGE consortium. No post-GWAS filtering on allele frequency, imputation info score or number of studies has been performed. The selected SNPs correspond to three groups: 1) A MAF 1KG reference set, 2) GWAS catalog top hits for arachidonic acid and 3) GWAS top hits for arachidonic acid in the CHARGE study

Usage

```
ara_test_dat
```

Format

A data frame with 436 rows and 9 variables:

snp SNP rsid

effect_allele effect allele

other_allele non-effect allele

effect_allele_freq effect allele frequency

beta change in arachidonic acid per copy of the effect allele

se standard error for beta

p p value statistic describing the association between the SNP and arachidonic acid

n number of study participants

path_to_target_file name of file used to generate example dataset

Source

http://www.chargeconsortium.com/main/results

---

charge_top_hits *GWAS top hits for arachidonic acid in the CHARGE consortium*

Description

The dataset contains rsids for single nucleotide polymorphisms extracted from a genome-wide association study of arachidonic acid in the CHARGE consortium. The list was generated by 1) extracting all SNPs with P values <5e-8 (1063 SNPs in total); and then 2) performing LD clumping on the 1063 extracted SNPs (clump_r2 = 0.01, clump_kb=10000) using European participants from UK Biobank as a reference dataset. Clumping was performed using ieugwasr::ld_clump. No post-GWAS filtering on allele frequency, imputation info score or number of studies was performed on the GWAS summary statistics prior to the extraction of the SNPs.

Usage

```
charge_top_hits
```

Format

A charactor vector of length 210:

Source

http://www.chargeconsortium.com/main/results

`charge_top_hits_cleaned` *GWAS top hits for arachidonic acid in the CHARGE consortium after post-GWAS cleaning*

Description

The dataset contains rsids for single nucleotide polymorphisms extracted from a genome-wide association study of arachidonic acid in the CHARGE consortium. Prior to extraction of the rsids, SNPs were excluded if they had a minor allele frequency <=5%, an imputation r2 score <=0.5 or were present in only 1 study (out of a total of 5 studies in the meta-analysis). This filtering steps are based on the post-GWAS filtering steps described in Guan et al (PMID=24823311). The list of rsids was then generated by: 1) extracting all SNPs with P values <5e-8 (219 SNPs in total); and then 2) performing LD clumping on the 219 extracted SNPs (clump_r2 = 0.01, clump_kb=10000) using European participants from UK Biobank as a reference dataset (6 SNPs remained after LD clumping). Clumping was performed using ieugwasr::ld_clump.

Usage

```
charge_top_hits_cleaned
```

Format

A charactor vector of length 6:

Source

http://www.chargeconsortium.com/main/results

`combine_plots` *Make cow plot*

Description

Combine all plots into a single plot using the cowplot package

Usage

```
combine_plots(
  Plot_list = NULL,
  out_file = NULL,
  return_plot = FALSE,
  width = 800,
  height = 1000,
  Title = "",
```

```
  Xlab = "",
  Ylab = "",
  Title_size = 0,
  Title_axis_size = 0,
  by2cols = TRUE,
  Ncol = 2,
  Tiff = FALSE
)
```

Arguments

Plot_list	plots to combine. Can either be vector of character strings giving the names of plot objects or a list of plot objects.
out_file	filepath to save the plot
return_plot	logical argument. If TRUE, plot is returned and is not save to out_file
width	width of plot
height	height of plot
Title	plot title
Xlab	label for X axis
Ylab	label for Y axis
Title_size	size of title
Title_axis_size	size of x axis title
by2cols	logical argument. If true, forces plot to have 2 columns
Ncol	number of columns
Tiff	save plot in tiff format. Default is set to FALSE. If set to FALSE, the plot is saved in png format. Not applicable if return_plot is set to TRUE.

Value

plot

---

compare_effect_to_gwascatalog

Compare the genetic effect sizes in the test dataset to the GWAS catalog

---

Description

Compare the direction of effects and effect allele frequency between the test dataset and the GWAS catalog, in order to identify effect allele meta data errors

Usage

```
compare_effect_to_gwascatalog(
  dat = NULL,
  efo = NULL,
  efo_id = NULL,
  trait = NULL,
  beta = NULL,
  se = NULL,
  gwas_catalog_ancestral_group = c("European", "East Asian"),
  exclude_palindromic_snps = TRUE,
  force_all_trait_study_hits = FALSE,
  distance_threshold = distance_threshold
)
```

Arguments

dat	the test dataset of interest
efo	trait of interest in the experimental factor ontology
efo_id	ID for trait of interest in the experimental factor ontology
trait	the trait of interest
beta	name of the column containing the SNP effect size
se	name of the column containing the standard error for the SNP effect size.
gwas_catalog_ancestral_group	restrict the comparison to these ancestral groups in the GWAS catalog. Default is set to (c("European","East Asian")
exclude_palindromic_snps	should the function exclude palindromic SNPs? default set to TRUE. If set to FALSE, then conflicts with the GWAS catalog could reflect comparison of different reference strands.
force_all_trait_study_hits	force the comparison to include GWAS hits from the test dataset if they are not in the GWAS catalog? This should be set to TRUE only if dat is restricted to GWAS hits for the trait of interest. This is useful for visualising whether the test trait study has an unusually larger number of GWAS hits, which could, in turn, indicate analytical issues with the summary statistics
distance_threshold	distance threshold for deciding if the GWAS hit in the test dataset is present in the GWAS catalog. For example, a distance_threshold of 25000 means that the GWAS hit in the test dataset must be within 25000 base pairs of a GWAS catalog association, otherwise it is reported as missing from the GWAS catalog.

Value

dataframe

`compare_effect_to_gwascatalog2` *Compare the genetic effect sizes in the test dataset to the GWAS catalog*

Description

Compare the direction of effects and effect allele frequency between the test dataset and the GWAS catalog, in order to identify effect allele meta data errors

Usage

```
compare_effect_to_gwascatalog2(
  dat = NULL,
  efo = NULL,
  efo_id = NULL,
  trait = NULL,
  gwas_catalog_ancestral_group = c("European", "East Asian"),
  exclude_palindromic_snps = TRUE,
  map_association_to_study = FALSE,
  beta = "beta",
  se = "se",
  gwas_catalog = NULL,
  force_all_trait_study_hits = FALSE,
  distance_threshold = distance_threshold
)
```

Arguments

`dat`	the test dataset of interest
`efo`	trait of interest in the experimental factor ontology
`efo_id`	ID for trait of interest in the experimental factor ontology
`trait`	the trait of interest
`gwas_catalog_ancestral_group`	restrict the comparison to these ancestral groups in the GWAS catalog. Default is set to (c("European","East Asian")
`exclude_palindromic_snps`	should the function exclude palindromic SNPs? default set to TRUE. If set to FALSE, then conflicts with the GWAS catalog could reflect comparison of different reference strands.
`map_association_to_study`	map associations to study in GWAS catalog. This supports matching of results on PMID and study ancestry, which increases accuracy of comparisons, but is slow when there are large numbers of associations. Default = FALSE.
`beta`	name of the column containing the SNP effect size

`se` name of the column containing the standard error for the SNP effect size.

`gwas_catalog` user supplied data frame containing results from the GWAS catalog for the trait of interest. If set to NULL then the function will retrieve results from the GWAS catalog.

`force_all_trait_study_hits`
force the comparison to include GWAS hits from the test dataset if they are not in the GWAS catalog? This should be set to TRUE only if dat is restricted to GWAS hits for the trait of interest. This is useful for visualising whether the test trait study has an unusually larger number of GWAS hits, which could, in turn, indicate analytical issues with the summary statistics

`distance_threshold`
distance threshold for deciding if the GWAS hit in the test dataset is present in the GWAS catalog. For example, a distance_threshold of 25000 means that the GWAS hit in the test dataset must be within 25000 base pairs of a GWAS catalog association, otherwise it is reported as missing from the GWAS catalog.

Value

dataframe

`extract_sig_snps` *Extract SNPs with P value below a specified threshold (e.g. significant SNPs)*

Description

Exract the rows of the summary dataset of interest with P values below the specified threshold. This only works on linux/mac operating systems.

Usage

```
extract_sig_snps(
  path_to_target_file = NULL,
  p_val_col_number = NULL,
  p_threshold = 5e-08
)
```

Arguments

`path_to_target_file`
path to the target file. This contains the summary data for the trait of interest

`p_val_col_number`
the column number corresponding to the P values for the SNP-trait associations

`p_threshold` Extract SNP-trait associtions with P values less than this value. Default set to 5e-8

Value

data frame

`extract_snps` *Extract SNPs*

Description

Exract the summary data for the rsids of interest from a target study. This only works on linux/ mac operating systems. Will not work on Windows.

Usage

```
extract_snps(
  snplist = NULL,
  path_to_target_file = NULL,
  exact_match = TRUE,
  path_to_target_file_sep = "\t",
  Test.gz = FALSE,
  fill = FALSE,
  Comment = "#",
  Head = TRUE,
  get_sig_snps = FALSE,
  p_val_col_number = NULL,
  p_threshold = 5e-08
)
```

Arguments

`snplist`	a list of rsids of interest, either a character vector or path_to_target_file with the list of rsids
`path_to_target_file`	path to the target file This contains the summary data for the trait of #' interest
`exact_match`	search for exact matches. Default TRUE
`path_to_target_file_sep`	column/field separator. Default assumes that data is tab separated
`Test.gz`	is the target data a gz file? Default set to FALSE
`fill`	argument from read.table. logical. If 'TRUE' then in case the rows have unequal length, blank fields are implicitly added. Default is FALSE
`Comment`	comment to pass to comment.char in read.table. default = "#"
`Head`	Does the file have a header ? Default set to TRUE
`get_sig_snps`	also extract the top hits from the target file, not just the SNPs specified in snplist. logic TRUE or FALSE. Default set to FALSE
`p_val_col_number`	the column number corresponding to the P values for the SNP-trait associations
`p_threshold`	Extract SNP-trait associtions with P values less than this value. Default set to 5e-8

Value

data frame

`find_hits_in_gwas_catalog` *Are hits in the GWAS catalog?*

Description

Identify GWAS hits in the test dataset and see if they overlap with GWAS hits in the GWAS catalog.

Usage

```
find_hits_in_gwas_catalog(
  gwas_hits = NULL,
  trait = NULL,
  efo = NULL,
  efo_id = NULL,
  distance_threshold = 25000
)
```

Arguments

`gwas_hits`	the "GWAS hits" in the test dataset (e.g. SNP-trait associations with P<5e-8)
`trait`	the trait of interest
`efo`	trait of interest in the experimental factor ontology
`efo_id`	ID for trait of interest in the experimental factor ontology
`distance_threshold`	distance threshold for deciding if the GWAS hit in the test dataset is present in the GWAS catalog. For example, a distance_threshold of 25000 means that the GWAS hit in the test dataset must be within 25000 base pairs of a GWAS catalog association, otherwise it is reported as missing from the GWAS catalog.

Value

list

flag_af_conflicts *Flag allele frequency conflicts*

Description

Flag allele frequency conflicts through comparison of reported allele frequency to minor allele frequency in the 1000 genomes super populations.

Usage

```
flag_af_conflicts(target_dat = NULL)
```

Arguments

target_dat	the dataset of interest. Data frame.

Value

list

flag_gc_conflicts *Flag conflicts with the GWAS catalog*

Description

Flag conflicts with the GWAS catalog through comparison of reported effect alleles and reported effect allele frequency.

Usage

```
flag_gc_conflicts(
  dat = NULL,
  beta = "lnor",
  se = "lnor_se",
  efo = NULL,
  trait = NULL,
  efo_id = NULL,
  gwas_catalog_ancestral_group = c("European", "East Asian"),
  exclude_palindromic_snps = TRUE
)
```

Arguments

dat	the test dataset of interest
beta	name of the column containing the SNP effect size
se	name of the column containing the standard error for the SNP effect size.
efo	trait of interest in the experimental factor ontology
trait	the trait of interest
efo_id	ID for trait of interest in the experimental factor ontology
gwas_catalog_ancestral_group	restrict the comparison to these ancestral groups in the GWAS catalog. Default is set to (c("European","East Asian")
exclude_palindromic_snps	should the function exclude palindromic SNPs? default set to TRUE. If set to FALSE, then conflicts with the GWAS catalog could reflect comparison of different reference strands.

Value

list

---

flag_gc_conflicts2 *Flag conflicts with the GWAS catalog*

Description

Flag conflicts with the GWAS catalog through comparison of reported effect alleles and reported effect allele frequency.

Usage

```
flag_gc_conflicts2(gc_dat = NULL)
```

Arguments

gc_dat	dataset generated by compare_effect_to_gwascatalog2()

Value

list

format_data *format data*

Description

Get the trait summary data ready for the QC checks.

Usage

```
format_data(
  dat = NULL,
  trait = NA,
  population = NA,
  ncase = NA,
  ncontrol = NA,
  rsid = NA,
  effect_allele = NA,
  other_allele = NA,
  beta = NA,
  se = NA,
  lnor = NA,
  lnor_se = NA,
  eaf = NA,
  p = NA,
  or = NA,
  or_lci = NA,
  or_uci = NA,
  chr = NA,
  pos = NA,
  z_score = NA,
  drop_duplicate_rsids = TRUE
)
```

Arguments

dat	the dataset to be formatted
trait	the name of the trait.
population	describe the population ancestry of the dataset
ncase	number of cases or name of the column specifying the number of cases
ncontrol	number of controls or name of the column specifying the number of controls. If your summary data was generated in a linear model of a continuous trait, use ncontrol to indicate the total sample size.
rsid	name of the column containing the rs number or identifiers for the genetic variants
effect_allele	name of the effect allele column

other_allele	name of the non-effect allele column
beta	name of the column containing the SNP effect sizes. Use this argument if your summary data was generated in a linear model of a continuous trait.
se	standard error for the beta. Use this argument if your summary data was generated in a linear model of a continuous trait.
lnor	name of the column containing the log odds ratio. If missing, tries to infer it from the odds ratio
lnor_se	name of the column containing the standard error for the log odds ratio. If missing, tries to infer it from 95% confidence intervals or pvalues
eaf	name of the effect allele frequency column
p	name of the pvalue columne
or	name of column containing the odds ratio
or_lci	name of column containing the lower 95% confidence interval for the odds ratio
or_uci	name of column containing the upper 95% confidence interval for the odds ratio
chr	name of the column containing the chromosome number for each genetic variant
pos	genomic position for the genetic variant in base pairs
z_score	effect size estimate divided by its standard error
drop_duplicate_rsids	drop duplicate rsids? logical. default TRUE. duplicate rsids may for example correspond to triallelic SNPs.

Value

data frame

`get_efo` *get efo*

Description

Retrieve the experimental factor ontology (EFO) for some trait of interest. EFOs are retrieved from ZOOMA https://www.ebi.ac.uk/spot/zooma/

Usage

```
get_efo(trait = NULL)
```

Arguments

trait	the trait of interest

Value

list

`glioma_test_dat` *A example dataset of genetic summary data*

Description

The dataset contains summary association statistics for 98 SNPs, generated in logistic regression models, from a genome-wide association study of glioma conducted by the GliomaScan consortium.

Usage

```
glioma_test_dat
```

Format

A data frame with 98 rows and 20 variables:

Locus SNP rsid

Allele1 non-effect allele

Allele2 effect allele

MAF SNP minor allele frequency in controls|cases

Geno_Counts genotype counts in controls/cases

Subjects Number of participants in study

p p value statistic describing the association between the SNP and glioma

OR odds ratio for glioma

OR_95._CI_l lower 95% confidence interval

OR_95._CI_u upper 95% confidence interval

CHROMOSOME chromosome number

LOCATION genomic coordinates in base pairs

controls number of controls

cases number of cases

eaf.controls effect allele frequency in controls

Source

https://pubmed.ncbi.nlm.nih.gov/22886559/

gwas_catalog_hits *GWAS top hits*

Description

Extract results for top hits for the trait of interest from the NHGRI-EBI GWAS catalog

Usage

```
gwas_catalog_hits(
  trait = NULL,
  efo = NULL,
  efo_id = NULL,
  map_association_to_study = FALSE
)
```

Arguments

trait	the trait of interest as reported in the GWAS catalog
efo	trait of intersest in the experimental factor ontology
efo_id	ID for trait of interest in the experimental factor ontology
map_association_to_study	map associations to study in GWAS catalog. This supports matching of results on PMID and study ancestry, which increases accuracy of comparisons, but is slow when there are large numbers of associations. It is recommended that you run this function with map_association_to_study set to FALSE. Then, if large numbers of conflicting effect sizes are identified, re-run with this argument set to TRUE. Default = FALSE.

Value

data frame

infer_ancestry *Infer ancestry*

Description

Infer possible ancestry through comparison of allele frequency amongst test dataset and 1000 genomes super populations. Returns list of Pearson correlation coefficients.

Usage

```
infer_ancestry(target_dat = NULL)
```

Arguments

target_dat	the dataset of interest. Data frame.

Value

list

---

make_plot_gwas_catalog *Plot comparing the test study to the GWAS catalog*

Description

Make a plot comparing signed Z scores, or effect allele frequency, between the test dataset and the GWAS catalog, in order to identify effect allele meta data errors

Usage

```
make_plot_gwas_catalog(
  dat = NULL,
  plot_type = "plot_zscores",
  efo_id = NULL,
  efo = NULL,
  trait = NULL,
  gwas_catalog_ancestral_group = c("European", "East Asian"),
  legend = TRUE,
  Title = "Comparison of Z scores between test dataset & GWAS catalog",
  Ylab = "Z score in test dataset",
  Xlab = "Z score in GWAS catalog",
  force_all_trait_study_hits = FALSE,
  exclude_palindromic_snps = TRUE,
  beta = "beta",
  se = "se",
  distance_threshold = 25000,
  return_dat = FALSE,
  map_association_to_study = FALSE,
  gwas_catalog = NULL,
  nocolour = FALSE,
  publication_quality = FALSE,
  gc_dat = NULL
)
```

Arguments

dat	the test dataset of interest

`plot_type` compare Z scores or effect allele frequency? For comparison of Z scores set plot_type to "plot_zscores". For comparison of effect allele frequency set to "plot_eaf". Default is set to "plot_zscores"

`efo_id` ID for trait of interest in the experimental factor ontology

`efo` trait of interest in the experimental factor ontology

`trait` the trait of interest

`gwas_catalog_ancestral_group`
restrict the comparison to these ancestral groups in the GWAS catalog. Default is set to (c("European","East Asian")

`legend` include legend in plot. Default TRUE

`Title` plot title

`Ylab` label for Y axis

`Xlab` label for X axis

`force_all_trait_study_hits`
force the plot to include GWAS hits from the outcome study if they are not in the GWAS catalog? This should be set to TRUE only if dat is restricted to GWAS hits for the trait of interest. This is useful for visualising whether the outcome/trait study has an unusually larger number of GWAS hits, which could, in turn, indicate that the summary statistics have not been adequately cleaned.

`exclude_palindromic_snps`
should the function exclude palindromic SNPs? default set to TRUE. If set to FALSE, then conflicts with the GWAS catalog could reflect comparison of different reference strands.

`beta` name of the column containing the SNP effect size

`se` name of the column containing the standard error for the SNP effect size.

`distance_threshold`
distance threshold for deciding if the GWAS hit in the test dataset is present in the GWAS catalog. For example, a distance_threshold of 25000 means that the GWAS hit in the test dataset must be within 25000 base pairs of a GWAS catalog association, otherwise it is reported as missing from the GWAS catalog.

`return_dat` if TRUE, the dataset used to generate the plot is returned to the user and no plot is made.

`map_association_to_study`
map associations to study in GWAS catalog. This supports matching of results on PMID and study ancestry, which increases accuracy of comparisons, but is slow when there are large numbers of associations. Default = FALSE

`gwas_catalog` user supplied data frame containing results from the GWAS catalog for the trait of interest. If set to NULL then the function will retrieve results from the GWAS catalog.

`nocolour` if TRUE, effect size conflicts are illustrated using shapes rather than colours. Default FALSE

`publication_quality`
produce a high resolution image e.g. for publication purposes. Default FALSE

`gc_dat` output of compare_effect_to_gwascatalog2. This will typically be ignored by most users. Default NULL

Value

plot

`make_plot_maf` *MAF plot*

Description

Make a plot comparing minor allele frequency between test dataset and reference studies.

Usage

```
make_plot_maf(
  ref_dat = NULL,
  ref_1000G = c("AFR", "AMR", "EAS", "EUR", "SAS", "ALL"),
  target_dat = NULL,
  eaf = "eaf",
  snp_target = "rsid",
  snp_reference = "SNP",
  ref_dat_maf = "MAF",
  target_dat_effect_allele = "effect_allele",
  target_dat_other_allele = "other_allele",
  ref_dat_minor_allele = "minor_allele",
  ref_dat_major_allele = "major_allele",
  trait = "trait",
  target_dat_population = "population",
  ref_dat_population = "population",
  target_study = "study",
  ref_study = "study",
  Title = "Comparison of allele frequency between test dataset & reference study",
  Ylab = "Allele frequency in test dataset",
  Xlab = "MAF in reference study",
  cowplot_title = "Allele frequency in test dataset vs 1000 genomes super populations",
  return_dat = FALSE,
  nocolour = FALSE,
  legend = TRUE,
  allele_frequency_conflict = 1,
  publication_quality = FALSE
)
```

Arguments

`ref_dat`	user supplied reference dataset. data frame. optional
`ref_1000G`	if ref_dat is NULL, the user should indicate the 1000 genomes reference study of interest. options are: AFR, AMR, EAS, EUR, SAS or ALL. Default is to make plots for all super populations

target_dat	the test dataset of interest. Data frame.
eaf	name of the effect allele frequency column in target_dat
snp_target	rsid column in target_dat
snp_reference	rsid column in ref_dat
ref_dat_maf	name of the minor allele frequency column in the reference dataset. Only necessary if ref_dat is specified
target_dat_effect_allele	name of the effect allele column in target_dat
target_dat_other_allele	name of the non-effect allele column in target_dat
ref_dat_minor_allele	name of the minor allele column in the reference dataset. Only necessary if ref_dat is specified
ref_dat_major_allele	name of the major allele column in the reference dataset. Only necessary if ref_dat is specified
trait	name of the trait corresponding to target_dat
target_dat_population	population ancestry of target_dat
ref_dat_population	name of column describing population ancestry of reference dataset. Only necessary if ref_dat is specified
target_study	column in target_dat indicating name of target study
ref_study	column in reference study indicating name of reference study. Only necessary if ref_dat is specified
Title	plot title
Ylab	Y label
Xlab	X label
cowplot_title	title of overall plot
return_dat	if TRUE, the dataset used to generate the plot is returned to the user and no plot is made.
nocolour	if TRUE, allele frequency conflicts are illustrated using shapes rather than colours.
legend	include legend in plot. Default TRUE
allele_frequency_conflict	how to define allele frequency conflicts. 1= flag SNPs in the test dataset whose reported minor allele has frequency >0.5. 2= additionally flag SNPs with allele frequency differening by more than 10 points from allele frequency in the reference dataset. Default = 1
publication_quality	produce a very high resolution image e.g. for publication purposes. Default FALSE

Value

plot

make_plot_pred_effect *Predicted versus reported effect sizes*

Description

Make a plot comparing the predicted effect sizes to the reported effect sizes.

Usage

```
make_plot_pred_effect(
  dat = NULL,
  Xlab = "Reported effect size",
  Ylab = "Expected effect size",
  subtitle = "",
  maf_filter = FALSE,
  bias = FALSE,
  Title = "Expected versus reported effect size",
  legend = TRUE,
  standard_errors = FALSE,
  pred_beta = "lnor_pred",
  pred_beta_se = "lnor_se_pred",
  beta = "lnor",
  se = "lnor_se",
  sd_est = "sd_est",
  exclude_1000G_MAF_refdat = TRUE,
  nocolour = FALSE,
  publication_quality = FALSE
)
```

Arguments

dat	the target dataset of interest
Xlab	label for X axis
Ylab	label for Y axis
subtitle	subtitle
maf_filter	minor allele frequency threshold. If not NULL, genetic variants with a minor allele frequency below this threshold are excluded
bias	logical argument. If TRUE, plots the % deviation of the expected from the reported effect size on the Y axis against the reported effect size on the X axis.
Title	plot title
legend	logical argument. If true, includes figure legend in plot
standard_errors	logical argument. If TRUE, plots the expected versus the reported standard errors for the effect sizes

pred_beta — name of column containing the predicted effect size

pred_beta_se — name of column containing the standard error for the predicted effect size

beta — name of column containing the reported effect size

se — name of column containing the standard error for the reported effect size

sd_est — the standard deviation of the phenotypic mean. Can either be a numeric vector of length 1 or name of the column in dat containing the standard deviation value (in which case should be constant across SNPs). Only applicable for continuous traits. If not supplied by the user, the standard deviation is approximated using sd_est, estimated by the predict_beta_sd() function. The sd_est is then used to standardise the reported effect size. If the reported effect size is already standardised (ie is in SD units) then sd_est should be set to NULL

exclude_1000G_MAF_refdat — exclude rsids from the 1000 genome MAF reference dataset.

nocolour — if TRUE, effect size conflicts are illustrated using shapes rather than colours. Default FALSE

publication_quality — produce a very high resolution image e.g. for publication purposes. Default FALSE

Value

plot

---

make_snplist *make a SNP list*

Description

Create a list of rsids corresponding to "top hits" in the GWAS catalog, the 1000 genomes super popualtions and SNPs of specific interest to the user (e.g. genetic instruments/proxies for the exposure of interest).

Usage

```
make_snplist(
  trait = NULL,
  efo_id = NULL,
  efo = NULL,
  ref1000G_superpops = TRUE,
  snplist_user = NULL
)
```

Arguments

trait	the name of the trait in the NHGRI-EBI GWAS catalog
efo_id	experimental factor ontology ID for trait of interest
efo	experimental factor ontology for the trait of interest
ref1000G_superpops	include reference SNPs from 1000 genomes super populations. Default=TRUE
snplist_user	character vector of user specified rsids.

Value

character vector

Examples

```
snplist<-make_snplist(efo_id="EFO_0006859",ref1000G_superpops=FALSE)
```

---

predict_beta_sd *Predicted standardised beta*

Description

Predict the standardised beta using sample sise, Z score and minor allele frequency. Returns the predicted standardised beta, proportion of phenotypic variance explained by the SNP (r2) and F statistic for each SNP

Usage

```
predict_beta_sd(
  dat = NULL,
  beta = "beta",
  se = "se",
  eaf = "eaf",
  sample_size = "ncontrol",
  pval = "p"
)
```

Arguments

dat	the outcome dataset of interest
beta	the effect size column
se	the standard error column
eaf	the effect allele frequency column
sample_size	the sample size column
pval	name of the p value column

Value

data frame with predicted standardised beta, r2 and F stat statistics and estimated standard deviation

---

`predict_lnor_sh` *Predicted log odds ratio*

Description

Predict the log odds ratio, using the Harrison approach. https://seanharrisonblog.com/2020/. The log odds ratio is inferred from the reported number of cases and controls, Z scores and minor allele frequency

Usage

```
predict_lnor_sh(dat = NULL)
```

Arguments

`dat`	the outcome dataset of interest

Value

data frame

`refdat_1000G_superpops` *A dataset of reference allele frequencies from 1000 genomes super-populations*

Description

The dataset contains minor allele frequency for 2297 SNPs that have minor allele frequency 0.1-0.3 across each superpopulation in the 1000 genomes project.

Usage

```
refdat_1000G_superpops
```

Format

A data frame with 13782 rows and 8 variables:

CHR chromosome number

SNP SNP rsid

minor_allele SNP minor allele

major_allele SNP major allele

MAF SNP minor allele frequency

NCHROBS number of observed chromosomes

population 1000 genomes superpopulation: AFR=African; ALL=all individuals; AMR = Ad Mixed American; EAS=East Asian; EUR=European; SAS=South Asian

Source

https://www.internationalgenome.org/home

`transform_betas` *Transform betas*

Description

Transform betas from a linear model to a log odds ratio scale. Assumes betas have been derived from a linear model of case-control status regressed on SNP genotype (additively coded).

Usage

```
transform_betas(dat = NULL, effect = "lnor", effect.se = "se")
```

Arguments

`dat`	the target dataset rsids
`effect`	the column containing the beta. We wish to transform this to a log odds ratio scale
`effect.se`	standard error for the beta

Value

data frame

zz_plot *ZZ plot*

Description

Calculate Z scores from the reported P values (Zp) and the reported log odds ratios (Zlnor). Construct a scatter plot of Zp and Zlnor

Usage

```
zz_plot(
  dat = NULL,
  Title = "ZZ plot",
  Ylab = "Z score inferred from p value",
  Xlab = "Z score inferred from effect size and standard error",
  beta = "lnor",
  se = "lnor_se",
  exclude_1000G_MAF_refdat = TRUE,
  publication_quality = FALSE
)
```

Arguments

dat	the target dataset of interest
Title	plot title
Ylab	label for Y axis
Xlab	label for X axis
beta	the name of the column containing the SNP effect size
se	the name of the column containing the standard error for the SNP effect size
exclude_1000G_MAF_refdat	exclude rsids from the 1000 genome MAF reference dataset.
publication_quality	produce a very high resolution image e.g. for publication purposes. Default FALSE

Value

plot

Index