Package 'MungeSumstats'

April 10, 2022

Type Package

Title Standardise summary statistics from GWAS

Version 1.2.4

Description The *MungeSumstats* package is designed to facilitate the standardisation of GWAS summary statistics. It reformats inputted summary statisitics to include SNP, CHR, BP and can look up these values if any are missing. It also removes duplicates across SNPs.

URL https://github.com/neurogenomics/MungeSumstats

BugReports https://github.com/neurogenomics/MungeSumstats/issues

License Artistic-2.0

Depends R(>= 4.1)

Imports magrittr, data.table, utils, R.utils, dplyr, stats, GenomicRanges, GenomeInfoDb, BSgenome, Biostrings, VariantAnnotation, stringr, googleAuthR, httr, jsonlite, methods, parallel, rtracklayer, RCurl

biocViews SNP, WholeGenome, Genetics, ComparativeGenomics, GenomeWideAssociation, GenomicVariation, Preprocessing

RoxygenNote 7.1.2

Encoding UTF-8

Roxygen list(markdown = TRUE)

Suggests SNPlocs.Hsapiens.dbSNP144.GRCh37, SNPlocs.Hsapiens.dbSNP144.GRCh38, BSgenome.Hsapiens.1000genomes.hs37d5, BSgenome.Hsapiens.NCBI.GRCh38, BiocGenerics, IRanges, S4Vectors, rmarkdown, markdown, knitr, testthat (>= 3.0.0), UpSetR, BiocStyle, covr, seqminer, Rsamtools, MatrixGenerics

Config/testthat/edition 3

VignetteBuilder knitr

git_url https://git.bioconductor.org/packages/MungeSumstats

git_branch RELEASE_3_14

git_last_commit a8af342

git_last_commit_date 2022-03-23

Date/Publication 2022-04-10

Author Alan Murphy [aut, cre] (<https://orcid.org/0000-0002-2487-8753>),
Brian Schilder [aut, ctb] (<https://orcid.org/0000-0001-5949-2191>),
Nathan Skene [aut] (<https://orcid.org/0000-0002-6807-3180>)

Maintainer Alan Murphy <alanmurph94@hotmail.com>

R topics documented:

`check_ldsc_format` *Ensures that parameters are compatible with LDSC format*

Description

Format summary statistics for direct input to Linkage Disequilibrium SCore (LDSC) regression without the need to use their `munge_sumstats.py` script first.

Usage

```
check_ldsc_format(
  sumstats_dt,
  ldsc_format,
  convert_n_int,
  allele_flip_check,
  compute_z,
```

```
  compute_n
)
```

Arguments

sumstats_dt	data table obj of the summary statistics file for the GWAS.
ldsc_format	Binary Ensure that output format meets all requirements to be fed directly into LDSC without the need for additional munging. Default is FALSE
convert_n_int	Binary, if N (the number of samples) is not an integer, should this be rounded? Default is TRUE.
allele_flip_check	Binary Should the allele columns be checked against reference genome to infer if flipping is necessary. Default is TRUE.
compute_z	Whether to compute Z-score column from P. Default is FALSE. **Note** that imputing the Z-score for every SNP will not correct be perfectly correct and may result in a loss of power. This should only be done as a last resort.
compute_n	Whether to impute N. Default of 0 won't impute, any other integer will be imputed as the N (sample size) for every SNP in the dataset. **Note** that imputing the sample size for every SNP is not correct and should only be done as a last resort. N can also be inputted with "ldsc", "sum", "giant" or "metal" by passing one of these for this field or a vector of multiple. Sum and an integer value creates an N column in the output whereas giant, metal or ldsc create an Neff or effective sample size. If multiples are passed, the formula used to derive it will be indicated.

Details

LDSC documentation.

Value

Formatted summary statistics

Source

LDSC GitHub

`download_vcf` *Download VCF file and its index file from Open GWAS*

Description

Ideally, we would use gwasvcf instead but it hasn't been made available on CRAN or Bioconductor yet, so we can't include it as a dep.

Usage

```
download_vcf(
  vcf_url,
  vcf_dir = tempdir(),
  vcf_download = TRUE,
  download_method = "download.file",
  force_new = FALSE,
  quiet = TRUE,
  nThread = 1
)
```

Arguments

vcf_url	Remote URL to VCF file.
vcf_dir	Where to download the original VCF from Open GWAS. *WARNING:* This is set to `tempdir()` by default. This means the raw (pre-formatted) VCFs be deleted upon ending the R session. Change this to keep the raw VCF file on disk (e.g. `vcf_dir="./raw_vcf"`).
vcf_download	Download the original VCF from Open GWAS.
download_method	`"axel"` (multi-threaded) or `"download.file"` (single-threaded) .
force_new	Overwrite a previously downloaded VCF with the same path name.
quiet	Run quietly.
nThread	Number of threads to parallelize over.

Value

List containing the paths to the downloaded VCF and its index file.

Examples

```
#only run the examples if user has internet access:
if(try(is.character(getURL("www.google.com")))==TRUE){
vcf_url <- "https://gwas.mrcieu.ac.uk/files/ieu-a-298/ieu-a-298.vcf.gz"
out_paths <- download_vcf(vcf_url = vcf_url)
}
```

---

find_sumstats *Search Open GWAS for datasets matching criteria*

Description

For each argument, searches for any datasets matching a case-insensitive substring search in the respective metadata column. Users can supply a single character string or a list/vector of character strings.

Usage

```
find_sumstats(
  ids = NULL,
  traits = NULL,
  years = NULL,
  consortia = NULL,
  authors = NULL,
  populations = NULL,
  categories = NULL,
  subcategories = NULL,
  builds = NULL,
  pmids = NULL,
  min_sample_size = NULL,
  min_ncase = NULL,
  min_ncontrol = NULL,
  min_nsnp = NULL,
  include_NAs = FALSE,
  access_token = check_access_token()
)
```

Arguments

ids	List of Open GWAS study IDs (e.g. `c("prot-a-664","ieu-b-4760")`).
traits	List of traits (e.g. `c("parkinson","Alzheimer")`).
years	List of years (e.g. `seq(2015,2021)` or `c(2010,2012,2021)`).
consortia	List of consortia (e.g. `c("MRC-IEU","Neale Lab")`.
authors	List of authors (e.g. `c("Elsworth","Kunkle","Neale")`).
populations	List of populations (e.g. `c("European","Asian")`).
categories	List of categories (e.g. `c("Binary","Continuous","Disease","Risk factor")`)).
subcategories	List of categories (e.g. `c("neurological","Immune","cardio")`)).
builds	List of genome builds (e.g. `c("hg19","grch37")`).
pmids	List of PubMed ID (exact matches only) (e.g. `c(29875488,30305740,28240269)`).
min_sample_size	Minimum total number of study participants (e.g. `5000`).
min_ncase	Minimum number of case participants (e.g. `1000`).
min_ncontrol	Minimum number of control participants (e.g. `1000`).
min_nsnp	Minimum number of SNPs (e.g. `200000`).
include_NAs	Include datasets with missing metadata for size criteria (i.e. `min_sample_size`, `min_ncase`, or `min_ncontrol`).
access_token	Google OAuth2 access token. Used to authenticate level of access to data

Details

By default, returns metadata for all studies currently in Open GWAS database.

Value

(Filtered) GWAS metadata table.

Examples

```
#only run the examples if user has internet access:
if(try(is.character(getURL("www.google.com")))==TRUE){
### By ID
metagwas <- find_sumstats(ids = c(
    "ieu-b-4760",
    "prot-a-1725",
    "prot-a-664"
))
### By ID amd sample size
metagwas <- find_sumstats(
    ids = c("ieu-b-4760", "prot-a-1725", "prot-a-664"),
    min_sample_size = 5000
)
### By criteria
metagwas <- find_sumstats(
    traits = c("alzheimer", "parkinson"),
    years = seq(2015, 2021)
)
}
```

`format_sumstats` *Check that summary statistics from GWAS are in a homogeneous format*

Description

Check that summary statistics from GWAS are in a homogeneous format

Usage

```
format_sumstats(
  path,
  ref_genome = NULL,
  convert_ref_genome = NULL,
  convert_small_p = TRUE,
  compute_z = FALSE,
  force_new_z = FALSE,
  compute_n = 0L,
  convert_n_int = TRUE,
  analysis_trait = NULL,
  INFO_filter = 0.9,
  FRQ_filter = 0,
  pos_se = TRUE,
```

```
  effect_columns_nonzero = FALSE,
  N_std = 5,
  N_dropNA = TRUE,
  rmv_chr = c("X", "Y", "MT"),
  rmv_chrPrefix = TRUE,
  on_ref_genome = TRUE,
  strand_ambig_filter = FALSE,
  allele_flip_check = TRUE,
  allele_flip_drop = TRUE,
  allele_flip_z = TRUE,
  allele_flip_frq = TRUE,
  bi_allelic_filter = TRUE,
  snp_ids_are_rs_ids = TRUE,
  remove_multi_rs_snp = FALSE,
  frq_is_maf = TRUE,
  sort_coordinates = TRUE,
  nThread = 1,
  save_path = tempfile(fileext = ".tsv.gz"),
  write_vcf = FALSE,
  tabix_index = FALSE,
  return_data = FALSE,
  return_format = "data.table",
  ldsc_format = FALSE,
  log_folder_ind = FALSE,
  log_mungesumstats_msgs = FALSE,
  log_folder = tempdir(),
  imputation_ind = FALSE,
  force_new = FALSE,
  mapping_file = sumstatsColHeaders
)
```

Arguments

path	Filepath for the summary statistics file to be formatted. A dataframe or datatable of the summary statistics file can also be passed directly to MungeSumstats using the path parameter.
ref_genome	name of the reference genome used for the GWAS ("GRCh37" or "GRCh38"). Argument is case-insensitive. Default is NULL which infers the reference genome from the data.
convert_ref_genome	name of the reference genome to convert to ("GRCh37" or "GRCh38"). This will only occur if the current genome build does not match. Default is not to convert the genome build (NULL).
convert_small_p	Binary, should p-values < 5e-324 be converted to 0? Small p-values pass the R limit and can cause errors with LDSC/MAGMA and should be converted. Default is TRUE.

`compute_z` Whether to compute Z-score column from P. Default is FALSE. **Note** that imputing the Z-score for every SNP will not correct be perfectly correct and may result in a loss of power. This should only be done as a last resort.

`force_new_z` When a "Z" column already exists, it will be used by default. To override and compute a new Z-score column from P set `force_new_z=TRUE`.

`compute_n` Whether to impute N. Default of 0 won't impute, any other integer will be imputed as the N (sample size) for every SNP in the dataset. **Note** that imputing the sample size for every SNP is not correct and should only be done as a last resort. N can also be inputted with "ldsc", "sum", "giant" or "metal" by passing one of these for this field or a vector of multiple. Sum and an integer value creates an N column in the output whereas giant, metal or ldsc create an Neff or effective sample size. If multiples are passed, the formula used to derive it will be indicated.

`convert_n_int` Binary, if N (the number of samples) is not an integer, should this be rounded? Default is TRUE.

`analysis_trait` If multiple traits were studied, name of the trait for analysis from the GWAS. Default is NULL.

`INFO_filter` numeric The minimum value permissible of the imputation information score (if present in sumstats file). Default 0.9.

`FRQ_filter` numeric The minimum value permissible of the frequency(FRQ) of the SNP (i.e. Allele Frequency (AF)) (if present in sumstats file). By default no filtering is done, i.e. value of 0.

`pos_se` Binary Should the standard Error (SE) column be checked to ensure it is greater than 0? Those that are, are removed (if present in sumstats file). Default TRUE.

`effect_columns_nonzero` Binary should the effect columns in the data BETA,OR (odds ratio),LOG_ODDS,SIGNED_SUMSTAT be checked to ensure no SNP=0. Those that do are removed(if present in sumstats file). Default FALSE.

`N_std` numeric The number of standard deviations above the mean a SNP's N is needed to be removed. Default is 5.

`N_dropNA` Drop rows where N is missing.Default is TRUE.

`rmv_chr` vector or character The chromosomes on which the SNPs should be removed. Use NULL if no filtering necessary. Default is X, Y and mitochondrial.

`rmv_chrPrefix` Remove "chr" or "CHR" from chromosome names. Default is TRUE.

`on_ref_genome` Binary Should a check take place that all SNPs are on the reference genome by SNP ID. Default is TRUE.

`strand_ambig_filter` Binary Should SNPs with strand-ambiguous alleles be removed. Default is FALSE.

`allele_flip_check` Binary Should the allele columns be checked against reference genome to infer if flipping is necessary. Default is TRUE.

`allele_flip_drop` Binary Should the SNPs for which neither their A1 or A2 base pair values match a reference genome be dropped. Default is TRUE.

`allele_flip_z` Binary should the Z-score be flipped along with effect and FRQ columns like Beta? It is assumed to be calculated off the effect size not the P-value and so will be flipped i.e. default TRUE.

`allele_flip_frq` Binary should the frequency (FRQ) column be flipped along with effect and z-score columns like Beta? Default TRUE.

`bi_allelic_filter` Binary Should non-biallelic SNPs be removed. Default is TRUE.

`snp_ids_are_rs_ids` Binary Should the supplied SNP ID's be assumed to be RSIDs. If not, imputation using the SNP ID for other columns like base-pair position or chromosome will not be possible. If set to FALSE, the SNP RS ID will be imputed from the reference genome if possible. Default is TRUE.

`remove_multi_rs_snp` Binary Sometimes summary statistics can have multiple RSIDs on one row (i.e. related to one SNP), for example "rs5772025_rs397784053". This can cause an error so by default, the first RS ID will be kept and the rest removed e.g."rs5772025". If you want to just remove these SNPs entirely, set it to TRUE. Default is FALSE.

`frq_is_maf` Conventionally the FRQ column is intended to show the minor/effect allele frequency (MAF) but sometimes the major allele frequency can be inferred as the FRQ column. This logical variable indicates that the FRQ column should be renamed to MAJOR_ALLELE_FRQ if the frequency values appear to relate to the major allele i.e. >0.5. By default this mapping won't occur i.e. is TRUE.

`sort_coordinates` Whether to sort by coordinates of resulting sumstats

`nThread` Number of threads to use for parallel processes.

`save_path` File path to save formatted data. Defaults to `tempfile(fileext=".tsv.gz")`.

`write_vcf` Whether to write as VCF (TRUE) or tabular file (FALSE).

`tabix_index` Index the formatted summary statistics with tabix for fast querying.

`return_data` Return `data.table`, `GRanges` or `VRanges` directly to user. Otherwise, return the path to the save data. Default is FALSE.

`return_format` If return_data is TRUE. Object type to be returned ("data.table","vranges","granges").

`ldsc_format` Binary Ensure that output format meets all requirements to be fed directly into LDSC without the need for additional munging. Default is FALSE

`log_folder_ind` Binary Should log files be stored containing all filtered out SNPs (separate file per filter). The data is outputted in the same format specified for the resulting sumstats file. The only exception to this rule is if output is vcf, then log file saved as .tsv.gz. Default is FALSE.

`log_mungesumstats_msgs` Binary Should a log be stored containing all messages and errors printed by MungeSumstats in a run. Default is FALSE

`log_folder` Filepath to the directory for the log files and the log of MungeSumstats messages to be stored. Default is a temporary directory.

`imputation_ind` Binary Should a column be added for each imputation step to show what SNPs have imputed values for differing fields. This includes a field denoting SNP allele flipping (flipped). On the flipped value, this denoted whether the alelles where switched based on MungeSumstats initial choice of A1, A2 from the input column headers and thus may not align with what the creator intended.**Note** these columns will be in the formatted summary statistics returned. Default is FALSE.

`force_new` If a formatted file of the same names as `save_path` exists, formatting will be skipped and this file will be imported instead (default). Set `force_new=TRUE` to override this.

`mapping_file` MungeSumstats has a pre-defined column-name mapping file which should cover the most common column headers and their interpretations. However, if a column header that is in youf file is missing of the mapping we give is incorrect you can supply your own mapping file. Must be a 2 column dataframe with column names "Uncorrected" and "Corrected". See data(sumstatsColHeaders) for default mapping and necessary format.

Value

The address for the modified sumstats file or the actual data dependent on user choice. Also, if log files wanted by the user, the return in both above instances are a list.

Examples

```
# Pass path to Educational Attainment Okbay sumstat file to a temp directory

eduAttainOkbayPth <- system.file("extdata", "eduAttainOkbay.txt",
    package = "MungeSumstats"
)

## Call uses reference genome as default with more than 2GB of memory,
## which is more than what 32-bit Windows can handle so remove certain checks

is_32bit_windows <-
    .Platform$OS.type == "windows" && .Platform$r_arch == "i386"
if (!is_32bit_windows) {
    reformatted <- format_sumstats(
        path = eduAttainOkbayPth,
        ref_genome = "GRCh37"
    )
} else {
    reformatted <- format_sumstats(
        path = eduAttainOkbayPth,
        ref_genome = "GRCh37",
        on_ref_genome = FALSE,
        strand_ambig_filter = FALSE,
        bi_allelic_filter = FALSE,
        allele_flip_check = FALSE
    )
}
# returned location has the updated summary statistics file
```

get_genome_builds *Infer genome builds*

Description

Infers the genome build of summary statistics files (GRCh37 or GRCh38) from the data. Uses SNP (RSID) & CHR & BP to get genome build.

Usage

```
get_genome_builds(
  sumstats_list,
  header_only = TRUE,
  sampled_snps = 10000,
  names_from_paths = FALSE,
  nThread = 1
)
```

Arguments

sumstats_list	A named list of paths to summary statistics, or a named list of data.table objects.
header_only	Instead of reading in the entire sumstats file, only read in the first N rows where N=sampled_snps. This should help speed up cases where you have to read in sumstats from disk each time.
sampled_snps	Downsample the number of SNPs used when inferring genome build to save time.
names_from_paths	Infer the name of each item in sumstats_list from its respective file path. Only works if sumstats_list is a list of paths.
nThread	Number of threads to use for parallel processes.

Details

Iterative version of get_genome_build.

Value

ref_genome the genome build of the data

Examples

```
# Pass path to Educational Attainment Okbay sumstat file to a temp directory

eduAttainOkbayPth <- system.file("extdata", "eduAttainOkbay.txt",
    package = "MungeSumstats"
)
```

```
sumstats_list <- list(ss1 = eduAttainOkbayPth, ss2 = eduAttainOkbayPth)

## Call uses reference genome as default with more than 2GB of memory,
## which is more than what 32-bit Windows can handle so remove certain checks
is_32bit_windows <-
    .Platform$OS.type == "windows" && .Platform$r_arch == "i386"
if (!is_32bit_windows) {

    #multiple sumstats can be passed at once to get all their genome builds:
    #ref_genomes <- get_genome_builds(sumstats_list = sumstats_list)
    #just passing first here for speed
    sumstats_list_quick <- list(ss1 = eduAttainOkbayPth)
    ref_genomes <- get_genome_builds(sumstats_list = sumstats_list_quick)
}
```

`hg19ToHg38` *UCSC Chain file hg19 to hg38*

Description

UCSC Chain file hg19 to hg38, .chain.gz file, downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/ on 09/10/21

Format

gunzipped chain file

Details

UCSC Chain file hg19 to hg38, .chain.gz file, downloaded on 09/10/21 To be used as a back up if the download from UCSC fails.

hg19ToHg38.over.chain.gz

NA

Source

The chain file was downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/
`utils::download.file('ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.`

hg38ToHg19 *UCSC Chain file hg38 to hg19*

Description

UCSC Chain file hg38 to hg19, .chain.gz file, downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/ on 09/10/21

Format

gunzipped chain file

Details

UCSC Chain file hg38 to hg19, .chain.gz file, downloaded on 09/10/21 To be used as a back up if the download from UCSC fails.

hg38ToHg19.over.chain.gz

NA

Source

The chain file was downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/
`utils::download.file('ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.`

ieu-a-298 *Local ieu-a-298 file from IEU Open GWAS*

Description

Local ieu-a-298 file from IEU Open GWAS, downloaded on 09/10/21.

Format

gunzipped tsv file

Details

Local ieu-a-298 file from IEU Open GWAS, downlaoded on 09/10/21. This is done in case the download in the package vignette fails.

ieu-a-298.tsv.gz

NA

Source

The file was downloaded with: `MungeSumstats::import_sumstats(ids = "ieu-a-298",ref_genome = "GRCH37")`

`import_sumstats` *Import full genome-wide GWAS summary statistics from Open GWAS*

Description

Requires internet access to run.

Usage

```
import_sumstats(
  ids,
  vcf_dir = tempdir(),
  vcf_download = TRUE,
  save_dir = tempdir(),
  write_vcf = FALSE,
  download_method = "download.file",
  quiet = TRUE,
  force_new_vcf = FALSE,
  nThread = 1,
  parallel_across_ids = FALSE,
  ...
)
```

Arguments

`ids`	List of Open GWAS study IDs (e.g. `c("prot-a-664","ieu-b-4760")`).
`vcf_dir`	Where to download the original VCF from Open GWAS. *WARNING:* This is set to `tempdir()` by default. This means the raw (pre-formatted) VCFs be deleted upon ending the R session. Change this to keep the raw VCF file on disk (e.g. `vcf_dir="./raw_vcf"`).
`vcf_download`	Download the original VCF from Open GWAS.
`save_dir`	Directory to save formatted summary statistics in.
`write_vcf`	Whether to write as VCF (TRUE) or tabular file (FALSE).
`download_method`	`"axel"` (multi-threaded) or `"download.file"` (single-threaded) .
`quiet`	Run quietly.
`force_new_vcf`	Overwrite a previously downloaded VCF with the same path name.
`nThread`	Number of threads to use for parallel processes.
`parallel_across_ids`	If `parallel_across_ids=TRUE` and `nThread>1`, then each ID in `ids` will be processed in parallel.
`...`	Additional arguments passed to format_sumstats.

Value

Either a named list of data objects or paths, depending on the arguments passed to `format_sumstats`.

Examples

```
#only run the examples if user has internet access:
if(try(is.character(getURL("www.google.com")))==TRUE){
### Search by criteria
metagwas <- find_sumstats(
    traits = c("parkinson", "alzheimer"),
    min_sample_size = 5000
)
### Only use a subset for testing purposes
ids <- (dplyr::arrange(metagwas, nsnp))$id

### Default usage
## You can supply \code{import_sumstats()}
## with a list of as many OpenGWAS IDs as you want,
## but we'll just give one to save time.

## Call uses reference genome as default with more than 2GB of memory,
## which is more than what 32-bit Windows can handle so remove certain checks
## commented out down to runtime
# datasets <- import_sumstats(ids = ids[1])
}
```

---

index_tabular *Convert summary stats file to tabix format*

Description

Convert summary stats file to tabix format

Usage

```
index_tabular(
  path,
  chrom_col = "CHR",
  start_col = "BP",
  end_col = start_col,
  verbose = TRUE
)
```

Arguments

`path`	Path to GWAS summary statistics file.
`chrom_col`	column for chromosome
`start_col`	column for start position

end_col	column for end position (is the same as start for snps)
verbose	Print messages.

Value

Path to tabix-indexed tabular file

Source

Borrowed function from echotabix.

Examples

```
eduAttainOkbayPth <- system.file("extdata", "eduAttainOkbay.txt",
                                 package = "MungeSumstats")
sumstats_dt <- data.table::fread(eduAttainOkbayPth, nThread = 1)
sumstats_dt <-
MungeSumstats:::standardise_sumstats_column_headers_crossplatform(
    sumstats_dt = sumstats_dt)$sumstats_dt
sumstats_dt <- MungeSumstats:::sort_coords(sumstats_dt = sumstats_dt)
path <- tempfile(fileext = ".tsv")
MungeSumstats::write_sumstats(sumstats_dt = sumstats_dt, save_path = path)

indexed_file <- MungeSumstats::index_tabular(path = path)
```

---

load_ref_genome_data *Load the reference genome data for SNPs of interest*

Description

Load the reference genome data for SNPs of interest

Usage

```
load_ref_genome_data(snps, ref_genome, msg = NULL)
```

Arguments

snps	Character vector SNPs by rs_id from sumstats file of interest.
ref_genome	Name of the reference genome used for the GWAS (GRCh37 or GRCh38)
msg	Optional name of the column missing from the dataset in question. Default is NULL

Value

data table of snpsById, filtered to SNPs of interest.

`load_snp_loc_data` *Loads the SNP locations and alleles for Homo sapiens extracted from NCBI dbSNP Build 144. Reference genome version is dependent on user input.*

Description

Loads the SNP locations and alleles for Homo sapiens extracted from NCBI dbSNP Build 144. Reference genome version is dependent on user input.

Usage

```
load_snp_loc_data(ref_genome, msg = NULL)
```

Arguments

`ref_genome`	name of the reference genome used for the GWAS (GRCh37 or GRCh38)
`msg`	Optional name of the column missing from the dataset in question

Value

SNP_LOC_DATA SNP positions and alleles for Homo sapiens extracted from NCBI dbSNP Build 144

Examples

```
SNP_LOC_DATA <- load_snp_loc_data("GRCH37")
```

`raw_ALSvcf` *GWAS Amyotrophic lateral sclerosis ieu open GWAS project - Subset*

Description

VCF (VCFv4.2) of the GWAS Amyotrophic lateral sclerosis ieu open GWAS project Dataset: ebi-a-GCST005647. A subset of 99 SNPs

Format

vcf document with 528 items relating to 99 SNPs

Details

A VCF file (VCFv4.2) of the GWAS Amyotrophic lateral sclerosis ieu open GWAS project has been subsetted here to act as an example summary statistic file in VCF format which has some issues in the formatting. MungeSumstats can correct these issues and produced a standardised summary statistics format.

ALSvcf.vcf

NA

Source

The summary statistics VCF (VCFv4.2) file was downloaded from https://gwas.mrcieu.ac.uk/datasets/ebi-a-GCST005647/ and formatted to a .rda with the following: `#Get example VCF dataset,use GWAS Amyotrophic lateral sclerosis ALS_GWAS_VCF <-readLines("ebi-a-GCST005647.vcf.gz") #Subset to just the first 99 SNPs ALSvcf <-ALS_GWAS_VCF[1:528] writeLines(ALSvcf,"inst/extdata/ALSvcf.vcf")`

---

raw_eduAttainOkbay *GWAS Educational Attainment Okbay 2016 - Subset*

Description

GWAS Summary Statistics on Educational Attainment by Okbay et al 2016: PMID: 27898078 PMCID: PMC5509058 DOI: 10.1038/ng1216-1587b. A subset of 93 SNPs

Format

txt document with 94 items

Details

GWAS Summary Statistics on Educational Attainment by Okbay et al 2016 has been subsetted here to act as an example summary statistic file which has some issues in the formatting. MungeSumstats can correct these issues.

eduAttainOkbay.txt

NA

Source

The summary statistics file was downloaded from https://www.nature.com/articles/ng.3552 and formatted to a .rda with the following: `#Get example dataset,use Educational-Attainment_Okbay_2016 link<-"Educational-Attainment_Okbay_2016/EduYears_Discovery_5000.txt" eduAttainOkbay<-readLines(link #There is an issue where values end with .0,this 0 is removed in func #There are also SNPs not on ref genome or arebi/tri allelic #So need to remove these in this dataset as its used for testing tmp <-tempfile() writeLines(eduAttainOkbay,con=tmp) eduAttainOkbay <-data.table::fread(tmp #DT read removes the .0's #remove those not on ref genome and withbi/tri allelic rmv <-c("rs192818565","rs7 eduAttainOkbay <-eduAttainOkbay[!MarkerName data.table::fwrite(eduAttainOkbay,file=tmp,sep="\t") eduAttainOkbay <-readLines(tmp) writeLines(eduAttainOkbay,"inst/extdata/eduAttainOkbay.txt")`

read_sumstats *Determine summary statistics file type and read them into memory*

Description

Determine summary statistics file type and read them into memory

Usage

```
read_sumstats(
  path,
  nThread = 1,
  nrows = Inf,
  standardise_headers = FALSE,
  mapping_file = sumstatsColHeaders
)
```

Arguments

path	Filepath for the summary statistics file to be formatted. A dataframe or datatable of the summary statistics file can also be passed directly to MungeSumstats using the path parameter.
nThread	Number of threads to use for parallel processes.
nrows	integer. The (maximal) number of lines to read. If Inf, will read in all rows.
standardise_headers	Standardise headers first.
mapping_file	MungeSumstats has a pre-defined column-name mapping file which should cover the most common column headers and their interpretations. However, if a column header that is in youf file is missing of the mapping we give is incorrect you can supply your own mapping file. Must be a 2 column dataframe with column names "Uncorrected" and "Corrected". See data(sumstatsColHeaders) for default mapping and necessary format.

Value

data.table of formatted summary statistics

Examples

```
path <- system.file("extdata", "eduAttainOkbay.txt",
    package = "MungeSumstats"
)
eduAttainOkbay <- read_sumstats(path = path)
```

sumstatsColHeaders *Summary Statistics Column Headers*

Description

List of uncorrected column headers often found in GWAS Summary Statistics column headers. Note the effect allele will always be the A2 allele, this is the approach done for VCF(https://www.ncbi.nlm.nih.gov/pmc/articles/PM
This is enforced with the column header corrections here and also the check allele flipping test.

Usage

```
data("sumstatsColHeaders")
```

Format

dataframe with 2 columns

Source

The code to prepare the .Rda file file from the marker file is: # Most the data in the below table comes from the LDSC github wiki data("sumstatsColHeaders") # Make additions to sumstatsColHeaders using github version of MungeSumstats-# shown is an example of adding columns for Standard Error (SE) #se_cols <-data.frame("Uncorrected"=c("SE","se","STANDARD.ERROR",# "STANDARD_ERROR","STAND
"Corrected"=rep("SE",5)) #sumstatsColHeaders <-rbind(sumstatsColHeaders,se_cols) #Once additions are made,order & save the new mapping dataset #now sort ordering -important for logic that # uncorrected=corrected comes first sumstatsColHeaders$ordering <-sumstatsColHeaders$Uncorre
sumstatsColHeaders <-sumstatsColHeaders[order(sumstatsColHeaders$Corrected,sumstatsColHeaders$orderi
= TRUE),] rownames(sumstatsColHeaders)<-1:nrow(sumstatsColHeaders) sumstatsColHeaders$ordering <-NULL usethis::use_data(sumstatsColHeaders,overwrite = TRUE,internal=TRUE) save(sumstatsColHeaders,f
You will need to restart your r session for effects to take account

write_sumstats *Write sum stats file to disk*

Description

Write sum stats file to disk

Usage

```
write_sumstats(
  sumstats_dt,
  save_path,
  sep = "\t",
  write_vcf = FALSE,
  tabix_index = FALSE,
```

```
  nThread = 1,
  return_path = FALSE
)
```

Arguments

sumstats_dt	data table obj of the summary statistics file for the GWAS.
save_path	File path to save formatted data. Defaults to `tempfile(fileext=".tsv.gz")`.
sep	The separator between columns. Defaults to the character in the set `[,\t |;:]` that separates the sample of rows into the most number of lines with the same number of fields. Use `NULL` or `""` to specify no separator; i.e. each line a single character column like `base::readLines` does.
write_vcf	Whether to write as VCF (TRUE) or tabular file (FALSE).
tabix_index	Index the formatted summary statistics with tabix for fast querying.
nThread	The number of threads to use. Experiment to see what works best for your data on your hardware.
return_path	Return `save_path`. This will have been modified in some cases (e.g. after compressing and tabix-indexing a previously un-compressed file).

Value

If `return_path=TRUE`, returns `save_path`. Else returns `NULL`.

Examples

```
path <- system.file("extdata", "eduAttainOkbay.txt",
    package = "MungeSumstats"
)
eduAttainOkbay <- read_sumstats(path = path)
write_sumstats(
    sumstats_dt = eduAttainOkbay,
    save_path = tempfile(fileext = ".tsv.gz")
)
```

Index