

# Package ‘MungeSumstats’

October 14, 2021

**Type** Package

**Title** Standardise summary statistics from GWAS

**Version** 1.0.1

**Description** The \*MungeSumstats\* package is designed to facilitate the standardisation of GWAS summary statistics. It reformats inputted summary statistics 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(>= 4.0)

**Imports** data.table, utils, stats, GenomicRanges, BSgenome, Biostrings

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

**RoxygenNote** 7.1.1

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

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

**Config/testthat/edition** 3

**VignetteBuilder** knitr

**git\_url** <https://git.bioconductor.org/packages/MungeSumstats>

**git\_branch** RELEASE\_3\_13

**git\_last\_commit** c1fe775

**git\_last\_commit\_date** 2021-06-22

**Date/Publication** 2021-10-14

**Author** Alan Murphy [cre] (<<https://orcid.org/0000-0002-2487-8753>>),  
Nathan Skene [aut] (<<https://orcid.org/0000-0002-6807-3180>>)

**Maintainer** Alan Murphy <alanmurph94@hotmail.com>

## R topics documented:

format_sumstats . . . . .	2
load_ref_genome_data . . . . .	4
load_snp_loc_data . . . . .	4
raw_ALSvcf . . . . .	5
raw_eduAttainOkbay . . . . .	5
sumstatsColHeaders . . . . .	6

<b>Index</b>	<b>7</b>
--------------	----------

---

format_sumstats	<i>Check that summary statistics from GWAS are in a homogeneous format</i>
-----------------	--

---

### Description

Check that summary statistics from GWAS are in a homogeneous format

### Usage

```
format_sumstats(
  path,
  ref_genome = "GRCh37",
  convert_small_p = TRUE,
  convert_n_int = TRUE,
  analysis_trait = NULL,
  INFO_filter = 0.9,
  N_std = 5,
  rmv_chr = c("X", "Y", "MT"),
  on_ref_genome = TRUE,
  strand_ambig_filter = FALSE,
  allele_flip_check = TRUE,
  bi_allelic_filter = TRUE
)
```

### Arguments

path	Filepath for the summary statistics file to be formatted
ref_genome	name of the reference genome used for the GWAS (GRCh37 or GRCh38). Default is GRCh37.

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.
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 sumstatsfile). Default 0.9
N_std	numeric The number of standard deviations above the mean a SNP's N is needed to be removed. Default is 5.
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.
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
bi_allelic_filter	Binary Should non-biallelic SNPs be removed. Default is TRUE

**Value**

The address for the modified sumstats file

**Examples**

```
#Pass path to Educational Attainment Okbay sumstat file to a temp directory
eduAttainOkbayPth <- system.file("extdata","eduAttainOkbay.txt",
package="MungeSumstats")
#pass path to format_sumstats
## 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 <- MungeSumstats::format_sumstats(eduAttainOkbayPth,
ref_genome="GRCh37")
} else{
reformatted <- MungeSumstats::format_sumstats(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
```

---

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

datatable 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\_eduAttain0kbay

*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 subsetting 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 are bi/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 with bi/tri allelic rmv <-c("rs192818565", "rs77112413") eduAttainOkbay <-eduAttainOkbay[!MarkerName %in% rmv, ] data.table::fwrite(eduAttainOkbay, file=tmp, sep="\t") eduAttainOkbay <-readLines(tmp) writeLines(eduAttainOkbay, "inst/extdata/eduAttainOkbay.txt")

---

sumstatsColHeaders	<i>Summary Statistics Column Headers</i>
--------------------	--

---

**Description**

List of uncorrected column headers often found in GWAS Summary Statistics column headers

**Usage**

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

**Format**

dataframe with 82 rows and 2 columns

**Source**

The code to prepare the .Rda file from the marker file is: # Most the data in the below table comes from the LDSC github wiki sumstatsColHeaders <-read.csv("inst/extdata/Magma\_Column\_headers.csv", sep=";", as.is=TRUE) usethis::use\_data(sumstatsColHeaders, overwrite = TRUE, internal=TRUE)

# Index

## \* datasets

sumstatsColHeaders, [6](#)

format\_sumstats, [2](#)

load\_ref\_genome\_data, [4](#)

load\_snp\_loc\_data, [4](#)

raw\_ALSvcf, [5](#)

raw\_eduAttainOkbay, [5](#)

sumstatsColHeaders, [6](#)