Package 'RapidoPGS'

October 12, 2022

Title A Fast and Light Package to Compute Polygenic Risk Scores

Version 2.2.0

Description Quickly computes polygenic scores from GWAS summary statistics of either case-control or quantitative traits, without LD matrix computation or parameter tuning. Reales,G., Vigorito, E., Kelemen,M., Wallace,C. (2021) <doi:10.1101/2020.07.24.220392> ``RápidoPGS: A rapid polygenic score calculator for summary GWAS data without a test dataset".

License GPL-3

Depends R (>= 4.0), data.table, RCurl, curl

Imports dplyr (>= 1.0.6), GenomicRanges (>= 1.42.0), IRanges (>= 2.24.1), bigsnpr (>= 1.8.1), coloc (>= 5.1.0), bigreadr (>= 0.2.4)

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Guillermo Reales [aut, cre] (<https://orcid.org/0000-0001-9993-3916>), Chris Wallace [aut] (<https://orcid.org/0000-0001-9755-1703>), Olly Burren [ctb] (<https://orcid.org/0000-0002-3388-5760>)

Maintainer Guillermo Reales <gr440@cam.ac.uk>

Repository CRAN

Date/Publication 2022-06-15 13:40:05 UTC

R topics documented:

`create_1000G` *Download 1000 Genomes Phase III panel*

Description

`create_1000G` downloads and gets 1000 Genomes Phase III panel in PLINK format, and apply quality control for being used to compute PGS using `rapidopgs_multi`. Given the size of the files, running this function can take long, depending on broadband speed and server status. We also recommend to ensure that there is at least 60GB free space available in disk.

Usage

```
create_1000G(
  directory = "ref-data",
  remove.related = TRUE,
  qc.maf = 0.01,
  qc.hwe = 1e-10,
  qc.geno = 0,
  autosomes.only = TRUE
)
```

Arguments

`directory`	a string indicating the directory to download the panel
`remove.related`	a logical stating if related individuals should be removed. Default TRUE.
`qc.maf`	a numeric to set the MAF threshold for variants to be removed. DEFAULT 0.01
`qc.hwe`	a numeric indicating the threshold for Hardy-Weinberg exact test p-value, below which variants will be removed. DEFAULT 1e-10.
`qc.geno`	a numeric to set maximum missing call rates for variants. DEFAULT = 0.
`autosomes.only`	If FALSE, it will include X and Y chromosomes, too.

Value

bed, fam and bim files for each chromosome in the chosen directory.

Author(s)

Guillermo Reales

Examples

```
## Not run:
create_1000G()

## End(Not run)
```

`EUR_ld.blocks` *LD block architecture for European populations (hg19).*

Description

A GRanges object containing the LD block for European ancestry, in hg19 build. This dataset was obtained from Berisa and Pickrell (2016), in bed format, then converted to GRanges. See manuscript for more details.

Usage

```
EUR_ld.blocks
```

Format

A GRanges object containing 1703 ranges

seqnames chromosome

ranges start and stop positions for the block

strand genomic strand, irrelevant here

Source

https://bitbucket.org/nygcresearch/ldetect-data/src

`EUR_ld.blocks38` *LD block architecture for European populations (hg38).*

Description

A GRanges object containing the LD block for European ancestry, in hg38 build. This dataset was obtained from MacDonald et al. (2022), in bed format, then transformed to GRanges. See manuscript for more details.

Usage

```
EUR_ld.blocks38
```

Format

A GRanges object containing 1361 ranges

seqnames chromosome

ranges start and stop positions for the block

strand genomic strand, irrelevant here

Source

https://raw.githubusercontent.com/jmacdon/LDblocks_GRCh38/master/data/EUR_LD_blocks.bed

---

`gwascat.download` *Retrieve GWAS summary datasets from GWAS catalog* '`gwascat.download` *takes a PMID from the user and downloads the associated summary statistics datasets published in GWAS catalog*

Description

This function, takes PUBMED ids as an input, searches at the GWAS catalog for harmonised datasets associated to that, interactively asking the user to choose if there are more than one, and fetches the dataset.

Usage

```
gwascat.download(ID, filenum = NULL, hm_only = TRUE)
```

Arguments

ID	a numeric. A PubMed ID (PMID) reference number from a GWAS paper.
filenum	a numeric. If multiple files are available, which one to choose? If NULL (DEFAULT), R will prompt an interactive prompt, asking for the number.
hm_only	a logical. Should GWAS catalog harmonised columns be retained?

Details

If multiple files are available for the same study, R will prompt an interactive dialogue to select a specific file, by number. If you know the number and prefer to select it automatically, you can provide it using file argument.

Value

a data.table containing the dataset.

Author(s)

Guillermo Reales

Examples

```
## Not run:
ds <- gwascat.download(29059683, hm_only = FALSE) # This should work: Michailidou dataset
wrongds <- gwascat.download(01223247236) # This shouldn't work: The Empress pub phone number

## End(Not run)
```

---

logsum *Helper function to sum logs without loss of precision*

Description

Sums logs without loss of precision This function is verbatim of its namesake in cupcake package (github.com/ollyburren/cupcake/)

Usage

```
logsum(x)
```

Arguments

x	a vector of logs to sum

Value

a scalar

Author(s)

Chris Wallace

`michailidou` *Subset of Michailidou BRCA GWAS sumstat dataset.*

Description

A data.table containing a subset of Michailidou et al., 2017 breast cancer summary statistic dataset, in hg38 build. This dataset is freely available in GWAS catalog (see link below). I removed unnecessary and all-missing columns, and rows with missing data at hm_beta and hm_effect_allele_frequency, and took a random sample of 100,000 SNPs without replacement.

Usage

```
michailidou
```

Format

A data.table object containing 100,000 SNPs

hm_rsid rsids, or SNP ids

hm_chrom chromosome

hm_pos base position, in hg38

hm_other_allele reference, or non-effect allele

hm_effect_allele alternative, or effect allele

hm_beta beta, log(OR), or effect size

hm_effect_allele_frequency effect allele frequency

standard_error standard error of beta

p_value p-value

Source

ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MichailidouK_29059683_GCST004988/harmonised/29059683-GCST004988-EFO_0000305.h.tsv.gz

`michailidou19` *Subset of Michailidou BRCA GWAS sumstat dataset.*

Description

A data.table containing a subset of Michailidou et al., 2017 breast cancer summary statistic dataset, in hg19 build. This dataset is freely available in GWAS catalog (see link below). I used "chromosome", "base_pair_location columns", removed unnecessary and all-missing columns, and took a random sample of 100,000 SNPs without replacement.

Usage

```
michailidou19
```

Format

A data.table object containing 100,000 SNPs

SNPID, CHR, BP, REF, ALT, ALT_FREQ, BETA, SE, P

SNPID rsids, or SNP ids

CHR chromosome

BP base position, in hg38

REF reference, or non-effect allele

ALT alternative, or effect allele

ALT_FREQ effect allele frequency

BETA beta, log(OR), or effect size

SE standard error of beta

P p-value

Source

ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MichailidouK_29059683_GCST004988/harmonised/29059683-GCST004988-EFO_0000305.h.tsv.gz

`rapidopgs_multi` *Compute PGS from GWAS summary statistics using Bayesian sum of single-effect (SuSiE) linear regression using z scores*

Description

'`rapidopgs_multi` computes PGS from a from GWAS summary statistics using Bayesian sum of single-effect (SuSiE) linear regression using z scores

Usage

```
rapidopgs_multi(
  data,
  trait = c("cc", "quant"),
  reference = NULL,
  LDmatrices = NULL,
  N = NULL,
  ancestry = "EUR",
  pi_i = 1e-04,
  ncores = 1,
  alpha.block = 1e-04,
  alpha.snp = 0.01,
  sd.prior = NULL
)
```

Arguments

`data`	a data.table containing GWAS summary statistic dataset with all required information.
`trait`	a string indicating if trait is a case-control ("cc") or quantitative ("quant").
`reference`	a string representing the path to the directory containing the reference panel (eg. "../ref-data/").
`LDmatrices`	a string representing the path to the directory containing the pre-computed LD matrices.
`N`	a numeric indicating the number of individuals used to generate input GWAS dataset, or a string indicating the column name containing per-SNP sample size. Required for quantitative traits only.
`ancestry`	a string indicating the ancestral population (DEFAULT: "EUR")
`pi_i`	a scalar representing the prior probability (DEFAULT: 1×10^{-4}).If you wish SuSiE to estimate this internally, set p=NULL.
`ncores`	a numeric specifying the number of cores (CPUs) to be used. If using pre-computed LD matrices, one core is enough for best performance.
`alpha.block`	a numeric threshold for minimum P-value in LD blocks. Blocks with minimum P above `alpha.block` will be skipped. Default: 1e-4.
`alpha.snp`	a numeric threshold for P-value pruning within LD block. SNPs with P above `alpha.snp` will be removed. Default: 0.01.
`sd.prior`	the prior specifies that BETA at causal SNPs follows a centred normal distribution with standard deviation sd.prior. If NULL (default) it will be automatically estimated (recommended).

Details

This function will take a GWAS summary statistic dataset as an input, will assign LD blocks to it, then use user-provided LD matrices or a preset reference panel in Plink format to compute LD matrices for each block. Then SuSiE method will be used to compute posterior probabilities of

variants to be causal and generate PGS weights by multiplying those posteriors by effect sizes (β). Unlike `rapidopgs_single`, this approach will assume one or more causal variants.

The GWAS summary statistics file to compute PGS using our method must contain the following minimum columns, with these exact column names:

CHR Chromosome

BP Base position (in GRCh37/hg19).

REF Reference, or non-effect allele

ALT Alternative, or effect allele, the one β refers to

BETA β (or log(OR)), or effect sizes

SE standard error of β

P P-value for the association test

In addition, quantitative traits must have the following extra column:

ALT_FREQ Minor allele frequency.

Also, for quantitative traits, sample size must be supplied, either as a number, or indicating the column name, for per-SNP sample size datasets (see below). Other columns are allowed, and will be ignored.

Reference panel should be divided by chromosome, in Plink format. Both reference panel and summary statistic dataset should be in GRCh37/hg19. For 1000 Genomes panel, you can use `create_1000G` function to set it up automatically.

If prefer to use LD matrices, you must indicate the path to the directory where they are stored. They must be in RDS format, named LD_chrZ.rds (where Z is the 1-22 chromosome number). If you don't have LD matrices already, we recommend downloading those gently provided by Prive et al., at https://figshare.com/articles/dataset/European_LD_reference/13034123. These matrices were computed using for 1,054,330 HapMap3 variants based on 362,320 European individuals of the UK biobank.

Value

a data.table containing the sumstats dataset with computed PGS weights.

Author(s)

Guillermo Reales, Chris Wallace

Examples

```
## Not run:
sumstats <- data.table(
CHR=c("4","20","14","2","4","6","6","21","13"),
BP=c(1479959, 13000913, 29107209, 203573414, 57331393, 11003529, 149256398,
25630085, 79166661),
REF=c("C","C","C","T","G","C","C","G","T"),
ALT=c("A","T","T","A","A","A","T","A","C"),
ALT_FREQ=c(0.2611,0.4482,0.0321,0.0538,0.574,0.0174,0.0084,0.0304,0.7528),
```

```
BETA=c(0.012,0.0079,0.0224,0.0033,0.0153,0.058,0.0742,0.001,-0.0131),
SE=c(0.0099,0.0066,0.0203,0.0171,0.0063,0.0255,0.043,0.0188,0.0074),
P=c(0.2237,0.2316,0.2682,0.8477,0.01473,0.02298,0.08472,0.9573,0.07535))
PGS  <- rapidopgs_multi(sumstats, trait="cc", reference = "ref-data/", ncores=2)

## End(Not run)
```

`rapidopgs_single` *Compute PGS from GWAS summary statistics using posteriors from Wakefield's approximate Bayes Factors*

Description

'`rapidopgs_single` computes PGS from a from GWAS summary statistics using posteriors from Wakefield's approximate Bayes Factors

Usage

```
rapidopgs_single(
  data,
  N = NULL,
  trait = c("cc", "quant"),
  build = "hg19",
  pi_i = 1e-04,
  sd.prior = if (trait == "quant") {     0.15 } else {     0.2 },
  filt_threshold = NULL,
  recalc = TRUE,
  reference = NULL
)
```

Arguments

`data`	a data.table containing GWAS summary statistic dataset with all required information.
`N`	a scalar representing the sample in the study, or a string indicating the column name containing it. Required for quantitative traits only.
`trait`	a string specifying if the dataset corresponds to a case-control ("cc") or a quantitative trait ("quant") GWAS. If trait = "quant", an ALT_FREQ column is required.
`build`	a string containing the genome build of the dataset, either "hg19" (for hg19/GRCh37) or "hg38" (hg38/GRCh38). DEFAULT "hg19".
`pi_i`	a scalar representing the prior probability (DEFAULT: 1×10^{-4}).
`sd.prior`	the prior specifies that BETA at causal SNPs follows a centred normal distribution with standard deviation sd.prior. Sensible and widely used DEFAULTs are 0.2 for case control traits, and 0.15 * var(trait) for quantitative (selected if trait == "quant").

`filt_threshold` a scalar indicating the ppi threshold (if `filt_threshold` < 1) or the number of top SNPs by absolute weights (if `filt_threshold` >= 1) to filter the dataset after PGS computation. If NULL (DEFAULT), no thresholding will be applied.

`recalc` a logical indicating if weights should be recalculated after thresholding. Only relevant if `filt_threshold` is defined.

`reference` a string indicating the path of the reference file SNPs should be filtered and aligned to, see Details.

Details

This function will take a GWAS summary statistic dataset as an input, will assign align it to a reference panel file (if provided), then it will assign SNPs to LD blocks and compute Wakefield's ppi by LD block, then will use it to generate PGS weights by multiplying those posteriors by effect sizes (β). Optionally, it will filter SNPs by a custom filter on ppi and then recalculate weights, to improve accuracy.

Alternatively, if filt_threshold is larger than one, RapidoPGS will select the top `filt_threshold` SNPs by absolute weights (note, not ppi but weights).

The GWAS summary statistics file to compute PGS using our method must contain the following minimum columns, with these exact column names:

CHR Chromosome

BP Base position (in GRCh37/hg19 or GRCh38/hg38). If using hg38, use build = "hg38" in parameters

REF Reference, or non-effect allele

ALT Alternative, or effect allele, the one β refers to

ALT_FREQ Minor/ALT allele frequency in the tested population, or in a close population from a reference panel. Required for Quantitative traits only

BETA β (or log(OR)), or effect sizes

SE standard error of β

If a reference is provided, it should have 5 columns: CHR, BP, SNPID, REF, and ALT. Also, it should be in the same build as the summary statistics. In both files, column order does not matter.

Value

a data.table containing the formatted sumstats dataset with computed PGS weights.

Author(s)

Guillermo Reales, Chris Wallace

Examples

```
sumstats <- data.table(SNPID=c("rs139096444","rs3843766","rs61977545", "rs544733737",
"rs2177641", "rs183491817", "rs72995775","rs78598863", "rs1411315"),
CHR=c("4","20","14","2","4","6","6","21","13"),
BP=c(1479959, 13000913, 29107209, 203573414, 57331393, 11003529, 149256398,
```

```
25630085, 79166661),
REF=c("C","C","C","T","G","C","C","G","T"),
ALT=c("A","T","T","A","A","A","T","A","C"),
BETA=c(0.012,0.0079,0.0224,0.0033,0.0153,0.058,0.0742,0.001,-0.0131),
SE=c(0.0099,0.0066,0.0203,0.0171,0.0063,0.0255,0.043,0.0188,0.0074))

PGS  <- rapidopgs_single(sumstats,  trait = "cc")
```

`sd.prior.est` *Compute Standard deviation prior (SD prior) for quantitative traits using pre-computed heritability.*

Description

`sd.prior.est` function will take the dataset as an input, a h^2 value obtained from a public repository such as LDhub, (http://ldsc.broadinstitute.org/ldhub/), sample size and number of variants, and will provide a sd.prior estimate that can be used to improve prediction performance of RapidoPGS functions on quantitative traits.

Usage

```
sd.prior.est(data, h2, N, pi_i = 1e-04)
```

Arguments

`data`	a data.table containing the GWAS summary statistic input dataset. Must contain SNPID and SE columns.
`h2`	a numeric. Heritability estimate or h^2 (See details).
`N`	a numeric. Sample size of the GWAS input dataset.
`pi_i`	a numeric. Prior that a given variant is causal. DEFAULT = 1e-4.

Author(s)

Guillermo Reales, Elena Vigorito, Chris Wallace

Examples

```
sumstats <- data.table(SNPID=c("4:1479959","20:13000913","14:29107209","2:203573414",
"4:57331393","6:11003529","6:149256398","21:25630085","13:79166661"),
REF=c("C","C","C","T","G","C","C","G","T"),
ALT=c("A","T","T","A","A","A","T","A","C"),
ALT_FREQ=c(0.2611,0.4482,0.0321,0.0538,0.574,0.0174,0.0084,0.0304,0.7528),
BETA=c(0.012,0.0079,0.0224,0.0033,0.0153,0.058,0.0742,0.001,-0.0131),
SE=c(0.0099,0.0066,0.0203,0.0171,0.0063,0.0255,0.043,0.0188,0.0074),
P=c(0.2237,0.2316,0.2682,0.8477,0.01473,0.02298,0.08472,0.9573,0.07535))
sd.prior <- sd.prior.est(sumstats, h2 = 0.2456, N = 45658, pi_i=1e-4)
```

`sdY.est` *Estimate trait variance, internal function*

Description

Estimate trait standard deviation given vectors of variance of coefficients, MAF and sample size

Usage

```
sdY.est(vbeta, maf, n)
```

Arguments

`vbeta`	vector of variance of coefficients
`maf`	vector of MAF (same length as vbeta)
`n`	sample size

Details

Estimate is based on var(beta-hat) = var(Y) / (n * var(X)) var(X) = 2* maf*(1-maf) so we can estimate var(Y) by regressing n*var(X) against 1/var(beta) This function is verbatim from its namesake in coloc package (github.com/chr1swallace/coloc/), by Chris Wallace

Value

estimated standard deviation of Y

Author(s)

Chris Wallace

`wakefield_pp` *compute posterior probabilities using Wakefield's approximate Bayes Factors* `wakefield_pp` *computes posterior probabilities for a given SNP to be causal for a given SNP under the assumption of a single causal variant.*

Description

This function was adapted from its namesake in cupcake package (github.com/ollyburren/cupcake/) to no longer require allele frequencies.

Usage

```
wakefield_pp(beta, se, pi_i = 1e-04, sd.prior = 0.2)
```

Arguments

beta	a vector of effect sizes (β) from a quantitative trait GWAS
se	vector of standard errors of effect sizes (β)
pi_i	a scalar representing the prior probability (DEFAULT 1×10^{-4})
sd.prior	a scalar representing our prior expectation of β (DEFAULT 0.2). The method assumes a normal prior on the population log relative risk centred at 0 and the DEFAULT value sets the variance of this distribution to 0.04, equivalent to a 95\ is in the range of 0.66-1.5 at any causal variant.

Value

a vector of posterior probabilities.

Author(s)

Olly Burren, Chris Wallace, Guillermo Reales

---

wakefield_pp_quant *Compute posterior probabilities using Wakefield's approximate Bayes Factors for quantitative traits*

Description

wakefield_pp_quant computes posterior probabilities for a given SNP to be causal for a given SNP under the assumption of a single causal variant.

Usage

```
wakefield_pp_quant(beta, se, sdY, sd.prior = 0.15, pi_i = 1e-04)
```

Arguments

beta	a vector of effect sizes (β) from a quantitative trait GWAS
se	vector of standard errors of effect sizes (β)
sdY	a scalar of the standard deviation given vectors of variance of coefficients, MAF and sample size. Can be calculated using sdY.est
sd.prior	a scalar representing our prior expectation of β (DEFAULT 0.15).
pi_i	a scalar representing the prior probability (DEFAULT 1×10^{-4}) The method assumes a normal prior on the population log relative risk centred at 0 and the DEFAULT value sets the variance of this distribution to 0.04, equivalent to a 95\ is in the range of 0.66-1.5 at any causal variant.

Details

This function was adapted from wakefield_pp in cupcake package (github.com/ollyburren/cupcake/)

Value

a vector of posterior probabilities.

Author(s)

Guillermo Reales, Chris Wallace

Index