Package 'RapidoPGS'

August 7, 2020

Title A Fast and Light Package to Compute Polygenic Risk Scores

Version 1.0.2

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., Kelemen,M., Wallace,C. (2020) <doi:10.1101/2020.07.24.220392> ``RápidoPGS: A rapid polygenic score calculator for summary GWAS data without validation dataset".

License GPL-3

Depends R (>= 3.6.0), data.table, RCurl, curl

Imports GenomicRanges, IRanges, bigsnpr

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

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 2020-08-07 14:20:02 UTC

R topics documented:

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

Description

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

Usage

```
computePGS(
  data,
  N0,
  N1 = NULL,
  build = "hg19",
  pi_i = 1e-04,
  sd.prior = if (is.null(N1)) {     0.15 } else {     0.2 },
  log.p = FALSE,
  filt_threshold = NULL,
  recalc = TRUE,
  reference = NULL,
  forsAUC = FALSE,
  altformat = FALSE
)
```

Arguments

`data`	a data.table containing GWAS summary statistic dataset with all required information.
`N0`	a scalar representing the number of controls in the study (or the number of subjects in quantitative trait GWAS), or a string indicating the column name containing it.
`N1`	a scalar representing the number of cases in the case-control study, or a string indicating the column name containing it. If NULL (DEFAULT), quantitative trait will be assumed.
`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 N1 is NULL).

`log.p`	if FALSE (DEFAULT), p is a p value. If TRUE, p is a log(p) value. Use this if your dataset holds p values too small to be accurately stored without using logs.
`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.
`forsAUC`	a logical indicating if output should be in sAUC evaluation format as we used it for the paper.
`altformat`	a logical indicating if output should be in a format containing pid (chr:pos), ALT, and weights only. DEFAULT FALSE

Details

Main RápidoPGS function. 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, RápidoPGS 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

SNPID rsids, or SNP identifiers. If not available, they can be anything (eg. CHR_BP)

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

BETA β (or log(OR)), or effect sizes

SE standard error of β

P P-value for the association test

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"),
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  <- computePGS(sumstats,  N0= 119078 ,N1=137045, build = "hg38")
```

`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 Belisa 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 Belisa and Pickrell (2016), in bed format, then liftovered to hg38 using UCSC liftOver tool, then converted to GRanges. See manuscript for more details.

Usage

```
EUR_ld.blocks38
```

Format

A GRanges object containing 1625 ranges

seqnames chromosome

ranges start and stop positions for the block

strand genomic strand, irrelevant here

Source

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

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

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 is verbatim of its namesake in cupcake package (github.com/ollyburren/cupcake/)

Usage

```
wakefield_pp(p, f, N, s, pi_i = 1e-04, sd.prior = 0.2, log.p = FALSE)
```

Arguments

p	a vector of univariate pvalues from a GWAS
f	a vector of minor allele frequencies taken from some reference population.
N	a scalar or vector for total sample size of GWAS
s	a scalar representing the proportion of cases (n.cases/N)
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.
log.p	if FALSE (DEFAULT), p is a p value. If TRUE, p is a log(p) value. Use this if your dataset holds p values too small to be accurately stored without using logs

Value

a vector of posterior probabilities.

Author(s)

Olly Burren, Chris Wallace

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

Arguments

`beta`	a vector of effect sizes (β) from a quantitative trait GWAS
`se.beta`	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