Package 'SCAT'

February 1, 2019

Type Package

Title Summary Based Conditional Association Test

Version 0.5.0

Date 2019-02-01

Author Han Zhang, Kai Yu

Maintainer Han Zhang <han.zhang2@nih.gov>

Depends stats, utils

Description Conditional association test based on summary data from genome-wide association study (GWAS). SCAT adjusts for heterogeneity in SNP coverage that exists in summary data if SNPs are not present in all of the participating studies of a GWAS meta-analysis. This commonly happens when different reference panels are used in participating studies for genotype imputation. This could happen when ones simply do not have data for some SNPs (e.g. different array, or imputated data is not available). Without properly adjusting for this kind of heterogeneity leads to inflated false positive rate. SCAT can also be used to conduct conventional conditional analysis when coverage heterogeneity is absent. For more details, refer to Zhang et al. (2018) Brief Bioinform. 19(6):1337-1343. <doi: 10.1093/bib/bbx072>.

License GPL-2 | GPL-3

LazyData TRUE

SystemRequirements C++11

NeedsCompilation yes

Repository CRAN

Date/Publication 2019-02-01 16:43:27 UTC

R topics documented:

read.bed *Reading data from binary PLINK files*

Description

Loads genotype data from PLINK format files .bed, .bim, and .fam.

Usage

```
read.bed(bed, bim, fam, sel.snps = NULL, sel.subs = NULL, encode012 = TRUE)
```

Arguments

bed	the name of the bed file.
bim	the name of the bim file.
fam	the name of the fam file.
sel.snps	a character vector of SNPs to be extracted from the plink files. The default is NULL, i.e., all SNPs are extracted.
sel.subs	an optional character vector specifying a subset of subject IDs to be extracted from the plink files. These IDs should be matched with the second column of fam files. The default is NULL, i.e., all subjects are extracted.
encode012	logical. Encoding the genotypes using 0/1/2 if TRUE, or using symbols of the reference and effect alleles if FALSE. The default is TRUE.

Value

A data frame of genotypes of specified subjects in the plink files.

Examples

```
library(SCAT)

# Load the sample data

bed <- system.file("extdata", package = 'SCAT', 'chr5.bed')
bim <- system.file("extdata", package = 'SCAT', 'chr5.bim')
fam <- system.file("extdata", package = 'SCAT', 'chr5.fam')

## first five SNPs
b <- read.table(bim, header = FALSE, as.is = TRUE, nrows = 5)
## first 50 subjects
f <- read.table(fam, header = FALSE, as.is = TRUE, nrows = 50)
geno <- read.bed(bed, bim, fam, sel.snps = b[, 2], sel.subs = f[, 2])

dim(geno) # 50 x 5
```

scat *Summary based conditional association test accounting for heterogeneity in SNP coverage*

Description

scat can be used to perform the conventional conditional test based on summary data generated from genome-wide association studies. These summary data are usually created from a meta-analysis, in which multiple studies are merged together to increase testing power in detecting novel associations. However, heterogeneity in SNP coverage widely exists in such data, even if genotype imputation is done. This is because imputation is usually conducted in each of the participating studies. As a result, SNPs may be missing in some of participating studies. Without properly dealing with such heterogeneity in SNP coverage can overestimate the correlation between association evidence between SNPs, and thus lead to flated false positive. The scat test accounts for the heterogeneity and has been demonstrated its ability in maintaining false positive rate at nominal level.

scat is used to test the association of a specified SNP conditioned on a set of index SNPs.

Usage

```
scat(summary.files, model, reference, lambda, nsamples,
     min.maf = 0.05, max.R2 = 0.9)
```

Arguments

summary.files	a character vector of file names containing the summary results of SNPs included in one or multiple studies. Each file must be able to be read by read.table. Each file must have columns called SNP, Chr, Pos, RefAllele, EffectAllele, Beta, and at least one of SE, P. An optional column called Direction describing studies information can also be included if the summary results were calculated from multiple studies by inverse weighting method. See Details.
model	a data.frame to specify index SNPs to be conditioned on and targeted SNPs to be tested. It allows for some sort of flexibility in format. It must contain two columns cond and test, while redundant columns would be ignored. In each line, the index SNPs specified in cond would be conditioned on simultaneously, while SNPs specified in test would be tested one by one. See Examples.
reference	a data.frame containing the paths of binary PLINK files of reference dataset. It must have columns called bed, bim and fam. The current version allows users to specify multiple sets of bed/bim/fam PLINK files as separate rows of the data frame. See Examples.
lambda	a numeric vector of inflation factors. Each file in summary.files should have one inflation factor specified in lambda.
nsamples	a list of numeric vectors specifying sample size of each participating study in each summary file. Each file in summary.files should corresponds to an element in this list. See Examples.

`min.maf` SNPs with minor allele frequencies (MAF) smaller than `min.maf` would be excluded from analysis. Low MAF usually leads to unstable estimation of correlation between SNPs. The default value is 0.05.

`max.R2` If the r-square between targeted SNP to be tested and any conditioned SNP is larger than `max.R2`, `scat` will not test them. Conditioning on highly correlated index SNPs would lead to unstable or misleading conditional p-values. The default value is 0.9.

Details

This function performs conditional association test if only summary data is available. The PLINK files provide information of LD between SNPs. Only SNPs that are simultaneously available in `model`, PLINK files, and at least one of the files in `summary.files` are tested, otherwise are simply dropped. SNPs that are conflict in alleles or genetic location, or that are not compatible with `min.maf` or `max.R2` are also discarded.

Each file in `summary.files` must contain

- `SNP` SNP name
- `Chr` chromosome.
- `Pos` base-pair position (bp units).
- `RefAllele` reference allele. Can be different in studies
- `EffectAllele` effect allele. Can be different in studies
- `Beta` estimated effect in linear regression model or log odds ratio in logistic regression model

and must contain one of the optional columns

- `SE` estimated standard error of Beta
- `P` p-value of Wald's, LRT or score test for testing `H_0: Beta = 0`. Can be generated by `lm`, `glm`, `anova` in `R` or other standard statistical softwares.

An optional column `Direction` is encouraged to be provided by the user

- `Direction` a character vector indicating which studies include a SNP. Any symbol except for '?' means a SNP is included in that study. Please note that the real direction of a SNP in studies ('+' or '-') does not matter, e.g., '++-?+' and '**+?-' provide exact the same information. See `Examples`.

The order of columns in each summary file and in `reference` are arbitrary, and all unnecessary columns (if any) are discarded in the analysis. The allele information in `RefAllele` and `EffectAllele` should be compatible with those in PLINK files, but case is not sensitive.

A file in `summary.files` can be considered as the result of a meta-analysis, in which one or multiple sub-studies are analyzed together. `scat` allows for multiple files specified in `summary.files` so that a meta-analysis is conducted on results from multiple meta-analyses.

The availability of the column `Direction` in a summary file are critical in adjusting for heterogeneity. If all SNPs in a summary file are tested on exactly the same set of subjects (e.g. all SNPs are completely imputed, or uniform coverage), then this column could be ignored in that file. Accordingly, the corresponding element in the list `nsamples` should be a single integer, the total sample size of all sub-studies in the file. If this column is missing in a file, a warning will be given to remind

the users to verify this strong assumption. This warning could be safely ignored if the coverage is uniform for all sub-studies in that file. Otherwise, users should consider to collect accurate coverage information before running the analysis.

If the SNP coverage in a summary file is not uniform, characters like '++-?*' are needed for every SNP in each line. As an example, '++-?*' means that there are in total five sub-studies used to generate that file, but for this particular SNP, the fourth sub-study is missing, and the `Beta` are positive in the first two sub-studies, is negative in the third sub-study, and the sign of `Beta` in the fifth sub-study is unknown for some reason, but we do know that the fifth sub-study has tested that SNP. The characters in `Direction` in the same summary file should have the same length. In this example, the corresponding element in the list `nsamples` should be a vector of five integers, corresponding to sample sizes of each of the five sub-studies. Please see `Examples` for more details.

Value

This function return a data frame of the following columns:

`Idx.SNP`	RS number of index SNPs being conditioned on. Separated by comma.
`Test.SNP`	RS number of SNP being tested.
`Idx.Pos`	Position information of `Idx.SNP`.
`Test.Pos`	Position information of `Test.SNP`.
`Idx.Dir`	Direction information of `Idx.SNP`. Separated by slash.
`Test.Dir`	Direction information of `Test.SNP`.
`Max.R2`	Maximum r-square between `Idx.SNP` and `Test.SNP`. Conditional test would be valuable only if `Max.R2` is relatively large, otherwise it is equivalent to the marginal test. However, `Max.R2` could not be too large, otherwise numerical concern may exist.
`Cor.Dir`	Direction of the greatest correlation between `Idx.SNP` and `Test.SNP` (squared correlation reaches `Max.R2`)
`Cond.P`	P-value of conditional association test.

References

Zhang H, Wheeler W, Song L, Yu K. (2017) Proper joint analysis of summary association statistics requires the adjustment of heterogeneity in SNP coverage pattern. Brief Bioinform. 19(6):1337-1343.

Examples

```
library(SCAT)

## Path of files containing summary statistics
## Only required columns will be loaded, so your files could contain redundant columns.
study1 <- system.file("extdata", package = "SCAT", "study1.txt.gz")
study2 <- system.file("extdata", package = "SCAT", "study2.txt.gz")
summary.files <- c(study1, study2)

## Prepare the PLINK files
```

```
## PLINK files for examples are built-in
fam <- vector("character", 2)
bim <- vector("character", 2)
bed <- vector("character", 2)

## suppose SNPs at chromosomes 5 and 8 are going to be tested
chr <- c(5, 8)
for(i in 1:2){
  fam[i] <- system.file("extdata", package = "SCAT", paste("chr", chr[i], ".fam", sep = ""))
  bim[i] <- system.file("extdata", package = "SCAT", paste("chr", chr[i], ".bim", sep = ""))
  bed[i] <- system.file("extdata", package = "SCAT", paste("chr", chr[i], ".bed", sep = ""))
}

reference <- data.frame(fam, bim, bed, stringsAsFactors = FALSE)

## different inflation factors are adjusted in two studies
## length of lambda and summary.files should be equal
lambda <- c(1.10, 1.08)

## we have two summary files, so there are two elements in the list nsamples
## the first summary file includes data calculated from meta-analysis of two sub-studies,
## each with sample size 63390 and 5643
## see a few rows in study1
# s <- read.table(study1, header = TRUE, as.is = TRUE, nrows = 10)
# s$direction
## [1] "+?" "++" "+?" "++" "++" "+?" "++" "+?" "+?" "+?"
## '?' means a SNP is not included in that sub-study
## any other symbols means a SNP is included in that sub-study
## the second summary file includes data calculated from a single sub-study with sample size 61957
nsamples <- list(c(63390, 5643),
                 c(61957))

## Space in model is okay, would be ignored
cond <- c('5:14957027, 5:32521333- 32522000',
          '5 : 179741534',
          '8:144662353 ,8:144663075,8:144663661')
test <- c('5:32525000 - 32526000, 5:98440820',
          '5:33930441 ,5:179738100-179740000',
          '8:144657269, 8:144664594')

model <- data.frame(cond, test, stringsAsFactors = FALSE)

## for each line in model, every single SNP specified in the
## column 'test' would be tested by conditioned on all SNPs
## in the column 'cond'
model

##                                  cond                               test
## 1     5:14957027, 5:32521333- 32522000 5:32525000 - 32526000, 5:98440820
## 2                         5 : 179741534 5:33930441 ,5:179738100-179740000
## 3 8:144662353 ,8:144663075,8:144663661          8:144657269, 8:144664594

## run it
```

```
scat(summary.files, model, reference, lambda, nsamples, min.maf = 0.01, max.R2 = 0.9)
```

Index