GSMR

Generalised Summary-data-based Mendelian Randomisation

Overview

The **gsmr** R-package implements the GSMR (Generalised Summary-data-based Mendelian Randomisation) method to test for causal association between a risk factor and disease[1]. The R package is developed by Zhihong Zhu, Zhili Zheng, Futao Zhang and Jian Yang at Institute for Molecular Bioscience, the University of Queensland. Bug reports or questions: z.zhu1@uq.edu.au or jian.yang@uq.edu.au.

Citation

Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. BioRxiv, 168674.

Installation

The **gsmr** requires R >= 2.15, you can install it in R by:

```
# gsmr requires the R-package survey
install.packages("survey")
# install gsmr
install.packages("http://cnsgenomics.com/software/gsmr/static/gsmr_1.0.3.tar.gz",repos=NULL,type="source")
```

The gsmr source codes are available in gsmr_1.0.3.tar.gz.

This online document has been integrated in the gsmr R-package, we can check that by the standard “?function_name” command in R.

Tutorial

The GSMR analysis only requires summary-level data from genome-wide association studies (GWAS). Here is an example, where the risk factor (*x*) is LDL cholesterol (LDL-c) and the disease (*y*) is coronary artery disease (CAD). GWAS summary data for both LDL-c and CAD are available in the public domain (Global Lipids Genetics Consortium et al. 2013, Nature Genetics; Nikpay, M. et al. 2015, Nature Genetics).

1. Prepare data for GSMR analysis

1.1 Load the example data

```
library("gsmr")
data("gsmr")
head(gsmr_data)
```

```
##          SNP a1 a2      freq     bzx bzx_se   bzx_pval    bzx_n       bzy
## 1  rs2419604  A  G 0.2830715  0.0302 0.0040  7.490e-14 172807.0  0.010183
## 2   rs676385  A  G 0.3116318 -0.0354 0.0043  1.169e-15 171609.0 -0.022094
## 3   rs648673  C  G 0.1315721 -0.0503 0.0057  1.155e-18 163522.0 -0.026150
## 4 rs17035630  A  G 0.1352071  0.0505 0.0061  1.438e-16 167679.5  0.031693
## 5   rs646776  C  T 0.2241335 -0.1602 0.0044 1.630e-272 173021.0 -0.101049
## 6    rs10410  A  G 0.1075555  0.0410 0.0061  6.197e-11 168300.0  0.037495
##      bzy_se     bzy_pval  bzy_n
## 1 0.0103044 3.230472e-01 184305
## 2 0.0101795 2.997370e-02 184305
## 3 0.0141759 6.508460e-02 184305
## 4 0.0131027 1.557110e-02 184305
## 5 0.0114222 9.010000e-19 184305
## 6 0.0173902 3.107610e-02 184305
```

```
dim(gsmr_data)
```

```
## [1] 151  12
```

The summary data contain 151 genetic instruments (i.e. SNPs).

- SNP: the genetic instrument
- a1: effect allele
- a2: the other allele
- freq: frequency of a1
- bzx: the effect size of a1 on risk factor
- bzx_se: standard error of bzx
- bzx_pval: p value for bzx
- bzx_n: per-SNP sample size of GWAS for the risk factor
- bzy: the effect size of a1 on disease
- bzy_se: standard error of bzy
- bzy_pval: p value for bzy
- bzy_n: per-SNP sample size of GWAS for the disease

1.2 Estimate the LD correlation matrix

```
# Save the genetic variants and coded alleles in R
write.table(gsmr_data[,c(1,2)], "gsmr_example_snps.allele", col.names=F, row.names=F, quote=F)
# Extract the genotype data from a PLINK file using GCTA (command line)
gcta64 --bfile gsmr_example --extract gsmr_example_snps.allele --update-ref-allele gsmr_example_s
nps.allele --out gsmr_example
```

Note: the two steps above guarantee that the LD correlations are calculated based on the coded alleles (sometimes called effect alleles) for the SNP effects.

```
# Estimate LD correlation matrix in R
snp_coeff_id = scan("gsmr_example.xmat.gz", what="", nlines=1)
snp_coeff = read.table("gsmr_example.xmat.gz", header=F, skip=2)
```

```
snp_order = match(gsmr_data[,1], snp_coeff_id)
snp_coeff_id = snp_coeff_id[snp_order]
snp_coeff = snp_coeff[, snp_order]
ldrho = cor(snp_coeff)
colnames(ldrho) = rownames(ldrho) = snp_coeff_id
# Check the size of the correlation matrix and double-check if the order of the SNPs in the LD co
rrelation matrix is consistent with that in the GWAS summary data.
```

```
dim(ldrho)
```

```
## [1] 151 151
```

```
# show the first 5 rows and columns of the matrix
ldrho[1:5,1:5]
```

```
##               rs2419604    rs676385     rs648673   rs17035630     rs646776
## rs2419604   1.000000000  0.01225467  0.003622746 -0.003508759  0.008039383
## rs676385    0.012254667  1.00000000 -0.086363592  0.032564923  0.167010220
## rs648673    0.003622746 -0.08636359  1.000000000 -0.033264311  0.204437659
## rs17035630 -0.003508759  0.03256492 -0.033264311  1.000000000 -0.195795791
## rs646776    0.008039383  0.16701022  0.204437659 -0.195795791  1.000000000
```

Note: all the analyses implemented in this R-package only require the summary data (e.g. “gsmr_data”) and the LD correlation matrix (e.g. “ldrho”) listed above.

2. Standardization

If the risk factor was not standardised in GWAS, we need to re-scale the effect sizes using the method below. This process requires allele frequencies, z-statistics and sample size.

```
snpfreq = gsmr_data$freq             # minor allele frequency of SNPs
bzx = gsmr_data$bzx     # effects of instruments on risk factor
bzx_se = gsmr_data$bzx_se       # standard errors of bzx
bzx_n = gsmr_data$bzx_n          # sample size for GWAS of the risk factor
std_zx = std_effect(snpfreq, bzx, bzx_se, bzx_n)    # perform standardize
gsmr_data$std_bzx = std_zx$b    # standardized bzx
gsmr_data$std_bzx_se = std_zx$se    # standardized bzx_se
head(gsmr_data)
```

```
##          SNP a1 a2      freq     bzx bzx_se   bzx_pval    bzx_n       bzy
## 1  rs2419604  A  G 0.2830715  0.0302 0.0040  7.490e-14 172807.0  0.010183
## 2   rs676385  A  G 0.3116318 -0.0354 0.0043  1.169e-15 171609.0 -0.022094
## 3   rs648673  C  G 0.1315721 -0.0503 0.0057  1.155e-18 163522.0 -0.026150
## 4 rs17035630  A  G 0.1352071  0.0505 0.0061  1.438e-16 167679.5  0.031693
## 5   rs646776  C  T 0.2241335 -0.1602 0.0044 1.630e-272 173021.0 -0.101049
## 6    rs10410  A  G 0.1075555  0.0410 0.0061  6.197e-11 168300.0  0.037495
##      bzy_se     bzy_pval  bzy_n     std_bzx  std_bzx_se
## 1 0.0103044 3.230472e-01 184305  0.02850320 0.003775258
## 2 0.0101795 2.997370e-02 184305 -0.03033421 0.003684664
## 3 0.0141759 6.508460e-02 184305 -0.04563918 0.005171835
## 4 0.0131027 1.557110e-02 184305  0.04179863 0.005048943
## 5 0.0114222 9.010000e-19 184305 -0.14785679 0.004060986
## 6 0.0173902 3.107610e-02 184305  0.03738796 0.005562599
```

3. HEIDI-outlier analysis

The estimate of causal effect of risk factor on disease can be biased by pleiotropy (see Ref 1 for details). This is an analysis to detect and eliminate from the analysis instruments that show significant pleiotropic effects on both risk factor and disease. The HEIDI-outlier analysis requires bzx (effect of genetic instrument on risk factor), bzx_se (standard error of bzx), bzy (effect of genetic instrument on disease), bzy_se (standard error of bzy) and ldrho (LD matrix of instruments). Note that LD matrix can be estimated from a reference sample with individual-level genotype data.

Here is an example to perform a HEIDI-outlier analysis.

```
bzx = gsmr_data$std_bzx     # SNP effects on risk factor
bzx_se = gsmr_data$std_bzx_se       # standard errors of bzx
bzx_pval = gsmr_data$bzx_pval    # p-values for bzx
bzy = gsmr_data$bzy     # SNP effects on disease
bzy_se = gsmr_data$bzy_se       # standard errors of bzy
gwas_thresh = 5e-8    # GWAS threshold to select SNPs as the instruments for the GSMR analysis
heidi_thresh = 0.01    # HEIDI-outlier threshold
filtered_index = heidi_outlier(bzx, bzx_se, bzx_pval, bzy, bzy_se, ldrho, snp_coeff_id, gwas_thre
sh, heidi_thresh) # perform HEIDI-outlier analysis
filtered_gsmr_data = gsmr_data[filtered_index,]   # select data passed HEIDI-outlier filtering
filtered_snp_id = snp_coeff_id[filtered_index]   # select SNPs that passed HEIDI-outlier filterin
g
dim(gsmr_data)
```

```
## [1] 151  14
```

```
dim(filtered_gsmr_data)
```

```
## [1] 138  14
```

There are 13 instruments filtered out by HEIDI-outlier and 138 instruments are retained for further analysis.

4. GSMR analysis

This is the main analysis of this R-package which utilises multiple genetic instruments to test for causal effect of risk factor on disease.

```
bzx = filtered_gsmr_data$std_bzx    # SNP effects on risk factor
bzx_se = filtered_gsmr_data$std_bzx_se    # standard errors of bzx
bzx_pval = filtered_gsmr_data$bzx_pval   # p-values for bzx
bzy = filtered_gsmr_data$bzy    # SNP effects on disease
bzy_se = filtered_gsmr_data$bzy_se    # standard errors of bzy
filtered_ldrho = ldrho[filtered_gsmr_data$SNP,filtered_gsmr_data$SNP]  # LD correlation matrix of
 SNPs
gsmr_results = gsmr(bzx, bzx_se, bzx_pval, bzy, bzy_se, filtered_ldrho, filtered_snp_id)    # GSM
R analysis
cat("Effect of exposure on outcome: ",gsmr_results$bxy)
```

```
## Effect of exposure on outcome:  0.4080517
```

```
cat("Standard error of bxy: ",gsmr_results$bxy_se)
```

```
## Standard error of bxy:  0.02249235
```

```
cat("P-value of bxy: ", gsmr_results$bxy_pval)
```

```
## P-value of bxy:  1.490807e-73
```

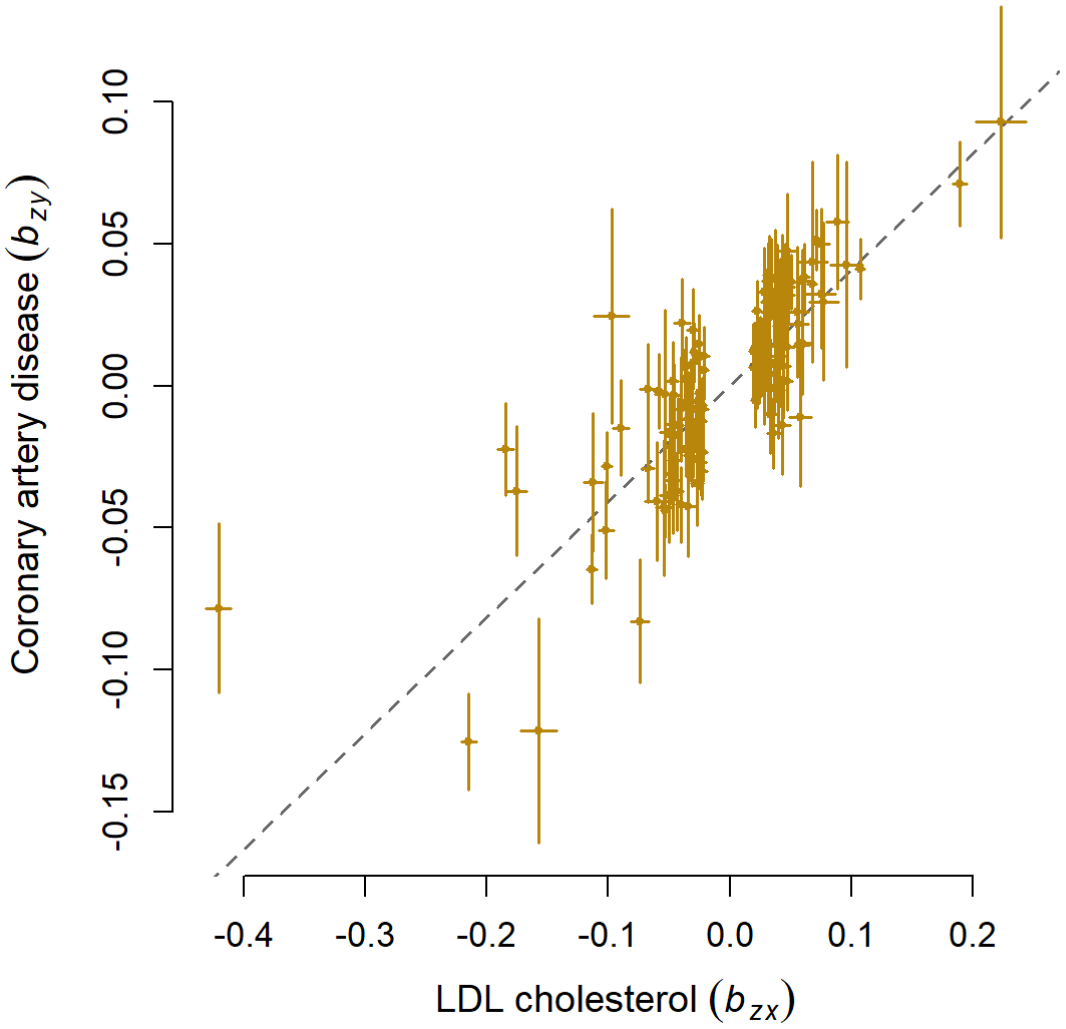
```
cat("Used index to GSMR analysis: ", gsmr_results$used_index[1:5], "...")
```

```
## Used index to GSMR analysis:  1 2 3 4 5 ...
```

5. Visulization

```
effect_col = colors()[75]
vals = c(bzx-bzx_se, bzx+bzx_se)
xmin = min(vals); xmax = max(vals)
vals = c(bzy-bzy_se, bzy+bzy_se)
ymin = min(vals); ymax = max(vals)
par(mar=c(5,5,4,2))
plot(bzx, bzy, pch=20, cex=0.8, bty="n", cex.axis=1.1, cex.lab=1.2,
        col=effect_col, xlim=c(xmin, xmax), ylim=c(ymin, ymax),
        xlab=expression(LDL~cholesterol~(italic(b[zx]))),
        ylab=expression(Coronary~artery~disease~(italic(b[zy]))))
abline(0, gsmr_results$bxy, lwd=1.5, lty=2, col="dim grey")

nsnps = length(bzx)
for( i in 1:nsnps ) {
    # x axis
    xstart = bzx[i] - bzx_se[i]; xend = bzx[i] + bzx_se[i]
    ystart = bzy[i]; yend = bzy[i]
    segments(xstart, ystart, xend, yend, lwd=1.5, col=effect_col)
    # y axis
    xstart = bzx[i]; xend = bzx[i]
    ystart = bzy[i] - bzy_se[i]; yend = bzy[i] + bzy_se[i]
    segments(xstart, ystart, xend, yend, lwd=1.5, col=effect_col)
}
```



Note: The dashed line is not a fitted regression line but a line with slope of bxy and intercept of 0.

Package Document

gsmr

> GSMR (Generalised Summary-data-based Mendelian Randomisation) is a flexible and powerful approach that utilises multiple genetic instruments to test for causal association between a risk factor and disease using summary-level data from independent genome-wide association studies.

heidi_outlier

> An analysis to detect and eliminate from the analysis instruments that show significant pleiotropic effects on both risk factor and disease

std_effect

> Standardization of SNP effect and its standard error using z-statistic, allele frequency and sample size

Package Document

gsmr

> GSMR (Generalised Summary-data-based Mendelian Randomisation) is a flexible and powerful approach that utilises multiple genetic instruments to test for causal association between a risk factor and disease using summary-level data from independent genome-wide association studies.

Usage

```
gsmr(bzx, bzx_se, bzx_pval, bzy, bzy_se, ldrho, snpid, gwas_thresh=5e-8, nsnps_thresh=10)
```

Arguments

`bzx`	vector, SNP effects on risk factor
`bzx_se`	vector, standard errors of bzx
`bzx_pval`	vector, p value for bzx
`bzy`	vector, SNP effects on disease
`bzy_se`	vector, standard errors of bzy
`ldrho`	LD correlation matrix of the SNPs
`snpid`	genetic instruments
`gwas_thresh`	threshold p-value to select instruments from GWAS for risk factor
`nsnps_thresh`	the minimum number of instruments required for the GSMR analysis (we do not recommend users to set this number smaller than 10)

Value

Estimate of causative effect of risk factor on disease (bxy), the corresponding standard error (bxy_se), p-value (bxy_pval) and SNP index (snp_index).

Examples

```
data("gsmr")
gsmr_result = gsmr(gsmr_data$bzx, gsmr_data$bzx_se, gsmr_data$bzx_pval, gsmr_data$bzy,
```

```
gsmr_data$bzy_se, ldrho, gsmr_data$SNP)
```

heidi_outlier

> An analysis to detect and eliminate from the analysis instruments that show significant pleiotropic effects on both risk factor and disease

Usage

```
heidi_outlier(bzx, bzx_se, bzx_pval, bzy, bzy_se, ldrho, snpid, gwas_thresh=5e-8, heidi_thresh=0.01)
```

Arguments

`bzx`	vector, SNP effects on risk factor
`bzx_se`	vector, standard errors of bzx
`bzx_pval`	vector, p value for bzx
`bzy`	vector, SNP effects on disease
`bzy_se`	vector, standard errors of bzy
`ldrho`	LD correlation matrix of the SNPs
`snpid`	genetic instruments
`gwas_thresh`	threshold p-value to select instruments from GWAS for risk factor
`heidi_thresh`	threshold p-value to remove pleiotropic outliers (the default value is 0.01)

Value

Retained index of genetic instruments

Examples

```
data("gsmr")
filtered_index = heidi_outlier(gsmr_data$bzx, gsmr_data$bzx_se, gsmr_data$bzx_pval,
gsmr_data$bzy, gsmr_data$bzy_se, ldrho, gsmr_data$SNP, 5e-8, 0.01)
```

std_effect

> Standardization of SNP effect and its standard error using z-statistic, allele frequency and sample size

Usage

```
std_effect(snp_freq, b, se, n)
```

Arguments

`snp_freq`	vector, allele frequency
`b`	vector, SNP effect on risk factor
`se`	vector, standard error of b
`n`	vector, per-SNP sample size of GWAS for risk factor

Value

Standardised effect (b) and standard error (se)

Examples

```
data("gsmr")
std_effects = std_effect(gsmr_data$freq, gsmr_data$bzx, gsmr_data$bzx_se, gsmr_data$bzx_n)
```