Package 'ldscR'

December 20, 2024

Type Package

Title Heritability and Genetic Correlation Matrix Estimation Using Linkage Disequilibrium Score Regression

Version 0.1.0

Author Yihe Yang

Maintainer Yihe Yang <yxy1234@case.edu>

Description This package provides tools for estimating heritability and genetic correlation matrices using Linkage Disequilibrium Score Regression (LDSC). It automates allele harmonization, merges GWAS results from multiple traits, and estimates the genetic covariance matrix for these traits. The diagonal elements of the genetic covariance matrix represent heritability estimates, while the corresponding correlation matrix represents the genetic correlation coefficients.

License MIT

Encoding UTF-8

LazyData true

Depends R (>= 2.10)

Imports MASS,
nloptr,
data.table,
CppMatrix

Remotes harryyiheyang/CppMatrix

RoxygenNote 7.2.3

R topics documented:

AFRLDSC *African Population LDSC Data*

Description

LDSC data for African populations. It includes LD scores calculated from the sample LD matrix and POET-estimated LD scores.

Usage

```
data(AFRLDSC)
```

Format

A data frame with 2 variables:

SNP SNP identifier.

LDSC LD score calculated from the sample LD matrix.

Source

LD scores derived from UK Biobank samples using PRScsx (https://github.com/getian107/PRScsx).

AMRLDSC *American Admixed Population LDSC Data*

Description

LDSC data for American admixed populations (Hispanic). It includes LD scores calculated from the sample LD matrix and POET-estimated LD scores.

Usage

```
data(AMRLDSC)
```

Format

A data frame with 2 variables:

SNP SNP identifier.

LDSC LD score calculated from the sample LD matrix.

Source

LD scores derived from UK Biobank samples using PRScsx (https://github.com/getian107/PRScsx).

EASLDSC *East Asian Population LDSC Data*

Description

LDSC data for East Asian populations. It includes LD scores calculated from the sample LD matrix and POET-estimated LD scores.

Usage

```
data(EASLDSC)
```

Format

A data frame with 2 variables:

SNP SNP identifier.

LDSC LD score calculated from the sample LD matrix.

Source

LD scores derived from UK Biobank samples using PRScsx (https://github.com/getian107/PRScsx).

EURLDSC *European Population LDSC Data*

Description

LDSC data for European populations. It includes LD scores calculated from the sample LD matrix and POET-estimated LD scores.

Usage

```
data(EURLDSC)
```

Format

A data frame with 2 variables:

SNP SNP identifier.

LDSC LD score calculated from the sample LD matrix.

Source

LD scores derived from UK Biobank samples using PRScsx (https://github.com/getian107/PRScsx).

filter_align *Filter and Align GWAS Data to a Reference Panel*

Description

The `filter_align` function processes a list of GWAS summary statistics data frames, harmonizes alleles according to a reference panel, removes duplicates, and aligns data to common SNPs. It's used to prepare data for further analysis such as LDSC.

Usage

```
filter_align(gwas_data_list, ref_panel, allele_match = T)
```

Arguments

`gwas_data_list`	A list of data.frames where each data.frame contains GWAS summary statistics for a trait. Each data.frame should include columns for SNP identifiers, Z-scores of effect size estimates, sample sizes (N), effect allele (A1), and reference allele (A2).
`ref_panel`	A data.frame containing the reference panel data. It must include columns for SNP, A1, and A2.
`allele_match`	An indicator of whether performing allele harmonization. Default to T.

Details

The function performs several key steps: adjusting alleles according to a reference panel, removing duplicate SNPs, and aligning all GWAS data frames to a set of common SNPs. This is often a necessary preprocessing step before performing genetic correlation and heritability analyses.

Value

A list of data.frames, each corresponding to an input GWAS summary statistics data frame, but filtered, harmonized, and aligned to the common SNPs found across all data frames.

hapmap3 *HapMap3 and UKBB Genotype SNP Data*

Description

A data frame containing combined SNP data from 1000 Genomes Project Phase 3 and UK Biobank (UKBB) genotypes. It includes a total of 1,664,852 SNPs.

Usage

```
data(hapmap3)
```

Format

A data frame with 1,664,852 rows and 3 variables:

SNP SNP identifier.

A1 Effect allele.

A2 Reference allele.

Source

1000 Genomes Project Phase 3 and UK Biobank genotype data.

---

ldsc.bicov *Single-Variate Linkage Disequilibrium Score Regression (LDSC)*

Description

The 'ldsc.bicov' function performs single-variate Linkage Disequilibrium Score Regression (LDSC) analysis. It is designed to estimate genetic covariance from two GWAS summary statistics, accounting for linkage disequilibrium (LD) between SNPs.

Usage

```
ldsc.bicov(
  gwas1,
  gwas2,
  h1,
  h2,
  LDSC,
  zsquare_thresh = 50000,
  nblock = 500,
  sampling.time = 200,
  sampling.ratio = 0.5
)
```

Arguments

gwas1	A data.frame containing GWAS summary statistics for the first trait. The data.frame should include columns for SNP identifiers, Z-scores of effect size estimates, sample sizes (N), effect allele (A1), and reference allele (A2).
gwas2	A data.frame containing GWAS summary statistics for the second trait. The data.frame should include columns for SNP identifiers, Z-scores of effect size estimates, sample sizes (N), effect allele (A1), and reference allele (A2).
h1	The heritability of the first trait.
h2	The heritability of the second trait.
LDSC	A data.frame containing LD Score Regression (LDSC) estimates. It should include LDSC scores and other necessary metrics for the analysis.
zsquare_thresh	A threshold for the squared Z-scores in heritability estimation to control for extreme values.
nblock	The number of blocks for bootstrap-based standard error estimation.
sampling.time	The number of block bootstrap.
sampling.ratio	The sub-sampling ratio in each bootstrap.

Details

The 'ldsc.bicov' function is particularly useful for the estimation of genetic covariance of two traits using LDSC. It requires well-formatted GWAS summary statistics and an understanding of LDSC methodology. The function includes an initial estimation phase followed by a reweighting step for improved efficiency and accuracy. It finally applies a blokc-wise bootstrap to estimate the standard error.

Value

A data.frame containing heritability estimates and their standard errors, along with the intercept and its standard error.

Examples

```
# Example usage
data(hapmap3)
data(EURLDSC)
ref_panel <- hapmap3
LDSC <- EURLDSC
results <- ldsc.univ(gwas1 = GWAS_data1, gwas2 = GWAS_data2, h1=h1, h2=h2, LDSC = LDSC)
```

---

ldsc.univ *Single-Variate Linkage Disequilibrium Score Regression (LDSC)*

Description

The 'ldsc.uni' function performs single-variate Linkage Disequilibrium Score Regression (LDSC) analysis. It is designed to estimate heritability from GWAS summary statistics, accounting for linkage disequilibrium (LD) between SNPs. The function harmonizes GWAS data with LD scores and applies non-linear optimization to estimate heritability.

Usage

```
ldsc.univ(
  gwas,
  LDSC,
  Boundary = F,
  zsquare_thresh = 50000,
  nblock = 500,
  sampling.time = 500,
  sampling.ratio = 0.5
)
```

Arguments

gwas	A data.frame containing GWAS summary statistics for a single trait. The data.frame should include columns for SNP identifiers, Z-scores of effect size estimates, sample sizes (N), effect allele (A1), and reference allele (A2).
LDSC	A data.frame containing LD Score Regression (LDSC) estimates. It should include LDSC scores and other necessary metrics for the analysis.

Boundary	An optional list of parameter estimate boundaries. If not provided, default values are used. It typically includes intercept.lower (the lower boundary of the intercept estimate), intercept.upper (the upper boundary of the intercept estimate), and h2.upper (the upper boundary of the heritability estimate).
zsquare_thresh	A threshold for the squared Z-scores in heritability estimation to control for extreme values.
nblock	The number of blocks for bootstrap-based standard error estimation.
sampling.time	The number of block bootstrap.
sampling.ratio	The sub-sampling ratio in each bootstrap.

Details

The 'ldsc.univ' function is particularly useful for single-trait heritability analysis using LDSC. It requires well-formatted GWAS summary statistics and an understanding of LDSC methodology. The function includes an initial estimation phase followed by a reweighting step for improved efficiency and accuracy. It finally applies a blokc-wise bootstrap to estimate the standard error.

Value

A data.frame containing heritability estimates and their standard errors, along with the intercept and its standard error.

Examples

```
# Example usage
data(hapmap3)
data(EURLDSC)
ref_panel <- hapmap3
LDSC <- EURLDSC
results <- ldsc.univ(gwas = GWAS_data, LDSC = LDSC)
```

---

ldscR *Estimate Heritability and Genetic Correlation Matrix Using LDSC estimated from sample LD matrix estimate*

Description

The ldscR function estimates heritability and genetic correlation matrices using Linkage Disequilibrium Score Regression (LDSC). It processes GWAS summary statistics, harmonizes alleles with a reference panel, and computes genetic covariance and error covariance matrices.

Usage

```
ldscR(
  GWAS_List,
  LDSC,
  Boundary = F,
  zsquare_thresh = 50000,
  cov_thresh = 10000,
  estimate_SE = F,
```

```
  nblock = 500,
  sampling.time = 500,
  sampling.ratio = 0.5
)
```

Arguments

GWAS_List	A list of data.frames where each data.frame contains GWAS summary statistics for a trait. Each data.frame should include columns for SNP identifiers, Z-scores of effect size estimates, sample sizes (N), effect allele (A1), and reference allele (A2).
LDSC	A data.frame containing LD Score Regression (LDSC) estimates. It should include LDSC scores and other necessary metrics for the analysis.
Boundary	A list of the information of upper and lower boundaries of parameter estimates. It typically includes intercept.lower (he lower boundary of the intercept estimate), intercept.upper (the upper boundary of the intercept estimate), h2.upper (the upper boundary of the heritability estimate).
zsquare_thresh	Threshold on the z-score square in heritability estimation.
cov_thresh	Threshold on the abs(z-score1 * z-score2) in genetic covariance estimation.
estimate_SE	If estimating the standard error of the genetic covariance matrix and estimation error covariance matrix. Default to F.
nblock	The number of blocks for bootstrap-based standard error estimation. Default to 200.
sampling.time	time for the estimation of standard errors. Default to 200.
sampling.ratio	The sub-sampling ratio in each bootstrap. Default to 0.5.

Details

The ldscR function is designed for advanced genetic statistics and requires a good understanding of GWAS summary statistics, LDSC methodology, and statistical genetics. Users should ensure that input data is correctly formatted and that they understand the implications of the estimates produced by the function.

Value

A list containing the following elements:

- GCovEstEstimated genetic covariance matrix.
- GCovSEStandard errors of the estimated genetic covariance matrix.
- ECovEstEstimated error covariance matrix.
- ECovSEStandard errors of the estimated error covariance matrix.
- Computing.timeComputing time in each stage.

Examples

```
data(hapmap3)
data(EURLDSC)
ref_panel <- hapmap3
LDSC <- EURLDSC
results <- ldscR(GWAS_List,LDSC)
```

SASLDSC *South Asian LD Score Regression (LDSC) Data*

Description

LDSC data for South Asian populations. It includes LD scores calculated from the sample LD matrix and POET-estimated LD scores.

Usage

```
data(SASLDSC)
```

Format

A data frame with 2 variables:

SNP SNP identifier.

LDSC LD score calculated from the sample LD matrix.

Source

LD scores derived from UK Biobank samples using PRScsx (https://github.com/getian107/PRScsx).

Index