Package 'TGVIS'

December 7, 2024

Type Package

Title Tissue-Gene pairs, direct causal Variants, and Infinitesimal effects Selector (TGVIS)

Version 0.1.0

Author Yihe Yang

Maintainer Yihe Yang <yxy1234@case.edu>

Description Implements the TGVI selector, a multivariable TWAS method that selects causal tissue-gene pairs and direct causal variants using SuSiE, while modeling infinitesimal effects via REML.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.3.1

Imports MASS,
Matrix,
Rcpp (>= 1.0.0),
RcppArmadillo,
susieR,
data.table,
tidyr,
CppMatrix

Remotes harryyiheyang/CppMatrix,
stephenslab/susieR

ORCID 0000-0001-6563-3579

Contents

`allele_harmonise` *allele_harmonise: Function to harmonize allele coding between a reference panel and GWAS/eQTL summary data*

Description

This function harmonizes the allele coding between an LD reference panel and GWAS/eQTL summary data to ensure consistency in the effect allele (A1) and the other allele (A2). The Z-scores in the GWAS/eQTL summary data are adjusted accordingly if the alleles are flipped.

Usage

```
allele_harmonise(ref_panel, gwas_data)
```

Arguments

`ref_panel`	A data.frame or data.table representing an LD reference panel. It must contain columns "SNP", "A1" (effect allele), and "A2" (other allele).
`gwas_data`	A data.frame or data.table representing GWAS/eQTL summary data. It must contain columns "SNP", "A1", "A2", and "Zscore".

Value

A data.table with harmonized alleles and adjusted Z-scores. The output contains the harmonized alleles from the reference panel and the original alleles from the GWAS/eQTL data as separate columns.

`ctwas` *ctwas: Function to perform SuSiE-based causal tissue-gene pair and direct causal variant estimation*

Description

This function performs the SuSiE (Sum of Single Effects) algorithm to estimate the causal tissue-gene pairs and direct causal variants using a provided set of summary statistics and linkage disequilibrium (LD) matrix. It returns a list of estimates and their associated statistics, including credible sets and Pratt estimations.

Usage

```
ctwas(
  by,
  bXest,
  LD,
  Noutcome,
  L.causal = 10,
  pip.thres.cred = 0.95,
  susie.iter = 500
)
```

Arguments

by	A vector of Z-scores of the marginal effects from the outcome GWAS.
bXest	A matrix of direct effect estimates based on the Z-scores of tissue-gene pairs.
LD	The LD matrix of variants.
Noutcome	The sample size of the outcome GWAS.
L.causal	The number of single effects used for tissue-gene pairs and direct causal variants. Default is 10.
pip.thres.cred	The cumulative PIP threshold for variables in a credible set. Default is 0.95.
susie.iter	The maximum number of iterations for 'susie_rss' within the profile-likelihood algorithm. Default is 50.

Value

A list containing:

theta	The estimated effects for tissue-gene pairs, scaled by the outcome GWAS sample size.
gamma	The estimated effects for direct causal variants, scaled by the outcome GWAS sample size.
theta.se	Standard errors of the tissue-gene pair effects.
gamma.se	Standard errors of the direct causal variant effects.
theta.pip	Posterior inclusion probabilities (PIP) for tissue-gene pairs.
gamma.pip	Posterior inclusion probabilities (PIP) for direct causal variants.
theta.pratt	Pratt index for tissue-gene pairs.
gamma.pratt	Pratt index for direct causal variants.
theta.cs	Credible set indicators for tissue-gene pairs.
gamma.cs	Credible set indicators for direct causal variants.
theta.cs.pip	Cumulative PIP within credible sets for tissue-gene pairs.
gamma.cs.pip	Cumulative PIP within credible sets for direct causal variants.
fit.causal	The SuSiE fit object for the causal analysis.
cs.summary	A summary of the credible sets obtained from the analysis.

---

eQTLmapping_susie *eQTLmapping_susie: Function to perform eQTL fine-mapping using SuSiE and optional resampling*

Description

This function performs fine-mapping of eQTLs using the SuSiE (Sum of Single Effects) algorithm. It allows for optional resampling of eQTL effects and returns both the estimated effects and resampled effects.

Usage

```
eQTLmapping_susie(
  bX,
  LD,
  Nvec,
  pip.thres = 0.5,
  pip.min = 0.2,
  L = 5,
  resample = F,
  sampling.time = 100
)
```

Arguments

bX	A matrix of Z-scores of marginal eQTL effect estimates for tissue-gene pairs.
LD	The LD matrix of variants.
Nvec	A vector representing the sample sizes of tissue-gene pair eQTL studies.
pip.thres	A threshold for individual PIP when no credible set is found. Default is 0.2.
pip.min	The minimum individual PIP in each 95% credible set. Used to remove variables with low PIPs within credible sets. Default is 0.05.
L	The number of single effects to be used in the SuSiE model. Default is 5.
resample	A logical value indicating whether to resample eQTL effects. Default is 'FALSE'.
sampling.time	The number of resampling iterations to perform when 'resample' is 'TRUE'. Default is 100.

Value

A list containing:

Estimate	A matrix of estimated eQTL effects for tissue-gene pairs.
Sampling	A matrix of resampled eQTL effects, if resampling is enabled. Otherwise, this will contain zeros.

---

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 = TRUE)
```

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	Logical. Whether to match alleles. Default TRUE.

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.

---

make_design_matrix *make_design_matrix: Constructs a Design Matrix of Z-scores for Gene-Tissue Pairs*

Description

This function constructs a design matrix from a ‘data.table‘ where each row represents a SNP, each column represents a tissue-gene pair, and the values are z-scores. If a tissue-gene pair does not have a z-score for a SNP, the corresponding entry will be ‘NA‘.

Usage

```
make_design_matrix(df)
```

Arguments

df	A ‘data.table‘ with three columns: ‘SNP‘, ‘Variable‘ (representing tissue-gene pairs), and ‘Zscore‘.

Value

A matrix where rows correspond to SNPs, columns correspond to tissue-gene pairs, and values are z-scores. Missing values (‘NA‘) indicate that the tissue-gene pair does not have a z-score for that SNP.

`modified_predixcan` *modified_predixcan: Function to estimate causal effects using a modified PrediXcan approach with BIC-based tuning parameter selection*

Description

This function implements a modified version of the PrediXcan method, using an alternating direction method of multipliers (ADMM) algorithm and BIC for selecting the optimal tuning parameter. It estimates the causal effects of tissue-gene pairs while accounting for direct causal variants.

Usage

```
modified_predixcan(
  by,
  bxest,
  LD,
  tauvec = seq(3, 10, by = 1),
  rho.gamma = 1.5,
  max.iter = 15,
  max.eps = 0.005,
  ebic.factor = 2,
  normmax = 2,
  pleiotropy.rm = NULL
)
```

Arguments

`by`	A vector of Z-scores of marginal effects from the outcome GWAS (same as in ctwas).
`bxest`	A vector of direct xQTL effect estimates for a tissue-gene pair.
`LD`	The LD matrix of variants (same as in ctwas).
`tauvec`	A vector of tuning parameters used in penalizing the direct causal effect. Default is 'seq(3,10,by=1)'.
`rho.gamma`	A parameter set in the ADMM algorithm. Default is 1.5.
`max.iter`	The maximum number of iterations for the ADMM algorithm. Default is 15.
`max.eps`	The convergence tolerance for the ADMM algorithm. Default is 0.005.
`ebic.factor`	The extended BIC factor for model selection. Default is 2.
`pleiotropy.rm`	A vector of indices specifying which variants should not be considered as having direct causal effects.

Value

A list containing:

`theta`	The estimated effect size of the tissue-gene pair.
`gamma`	The estimated effect sizes of the direct causal variants.
`covtheta`	The covariance of the estimated effect size 'theta'.
`Bic`	The BIC values for each tuning parameter.

Btheta	The estimated ‘theta‘ values for each tuning parameter.
Bgamma	The estimated ‘gamma‘ values for each tuning parameter.
Eta	The estimated linear predictor for the tissue-gene pair.

`poet_shrinkage` *poet_shrinkage: Function to perform covariance matrix shrinkage using the POET method*

Description

This function performs shrinkage of the covariance matrix using the POET (Principal Orthogonal complEment Thresholding) method, with an automatic selection of the number of latent factors and the optimal shrinkage parameter.

Usage

```
poet_shrinkage(
  LD,
  KMax = min(15, round(nrow(LD)/2)),
  lamvec = seq(0.025, 0.25, by = 0.025),
  minvalue = 0.001
)
```

Arguments

LD	A covariance matrix of linkage disequilibrium (LD).
KMax	The maximum number of latent factors used in POET. Default is ‘min(15, round(nrow(LD)/2))‘.
lamvec	A vector of candidate shrinkage parameters. Default is ‘seq(0.025, 0.25, by = 0.025)‘.
minvalue	The threshold for the minimum eigenvalues used in determining the optimal shrinkage parameter. Default is ‘1e-3‘.

Value

A covariance matrix that has been shrunk using the POET method.

`R2_partition` *R2_partition: Function to partition the Pratt index contributions of tissue-gene pairs, direct causal variants, and infinitesimal effects*

Description

This function partitions the total Pratt index of all effects into contributions from tissue-gene pairs, direct causal variants, and infinitesimal effects, using vectors of linear predictors and the outcome GWAS Z-scores.

Usage

```
R2_partition(y, eta.theta, eta.gamma, eta.upsilon = NULL, LD)
```

Arguments

y	A vector of Z-scores of marginal effects from the outcome GWAS.
eta.theta	A vector of the linear predictor of tissue-gene causal effects.
eta.gamma	A vector of the linear predictor of direct causal effects.
eta.upsilon	A vector of the linear predictor of infinitesimal effects. Default is 'NULL'.
LD	A LD matrix.

Value

A data frame containing:

r1	Total Pratt index of all effects.
r2	Pratt index contribution of tissue-gene pairs.
r3	Pratt index contribution of direct causal variants.
r4	Pratt index contribution of infinitesimal effects.

---

remove_missing_row_column

remove_missing_row_column: Removes Rows and Columns with Excessive Missing Values from a Matrix.

Description

This function processes a matrix, typically the output from 'make_design_matrix', by removing rows and columns that have a high percentage of missing values ('NA'). The function allows for flexibility in the order of operations, either removing rows first or columns first.

Usage

```
remove_missing_row_column(M, rowthres = 0.95, colthres = 0.95, rowfirst = TRUE)
```

Arguments

M	A matrix, typically the output from 'make_design_matrix', where rows represent SNPs, columns represent tissue-gene pairs, and values are z-scores.
rowthres	A numeric threshold (between 0 and 1). Rows with a proportion of missing values greater than this threshold will be removed.
colthres	A numeric threshold (between 0 and 1). Columns with a proportion of missing values greater than this threshold will be removed.
rowfirst	A logical value. If 'TRUE', rows are processed before columns. If 'FALSE', columns are processed before rows.

Value

A matrix with rows and columns containing excessive missing values removed based on the specified thresholds.

tgfm *tgfm: Function to perform eQTL fine-mapping and causal tissue-gene pair estimation*

Description

This function performs fine-mapping of eQTLs and estimates causal tissue-gene pairs using summary statistics, linkage disequilibrium (LD) matrix, and resampling methods. It returns a list of estimates and associated statistics, including credible sets, PIP resampling, and Pratt estimations.

Usage

```
tgfm(
  by,
  bX,
  LD,
  Nvec,
  L.eqtl = 5,
  L.causal = 10,
  pip.thres.cred = 0.5,
  eqtl.sampling.time = 100,
  causal.sampling.time = 100,
  eqtl.thres = 0.05,
  susie.iter = 1000
)
```

Arguments

by	A vector of Z-scores of the marginal effects from the outcome GWAS.
bX	A matrix of Z-scores of marginal eQTL effect estimates for tissue-gene pairs.
LD	The LD matrix of variants.
Nvec	A vector representing the sample sizes, with the first element for the outcome GWAS and the subsequent elements for eQTL studies.
L.eqtl	The number of single effects used in the eQTL fine-mapping step. Default is 5.
L.causal	The number of single effects used for tissue-gene pairs and direct causal variants. Default is 10.
pip.thres.cred	The cumulative PIP threshold for variables in a credible set. Default is 0.95.
eqtl.sampling.time	The number of resampling iterations for eQTL effect estimation. Default is 100.
causal.sampling.time	The number of resampling iterations for causal effect estimation. Default is 100.
eqtl.thres	A threshold for the minimum individual PIPs in each credible set during eQTL fine-mapping. This threshold is used to pre-remove variables with very low PIPs to avoid overly large credible sets. Default is 0.05.
susie.iter	The maximum number of iterations for ‘susie_rss‘ within the profile-likelihood algorithm. Default is 50.

Value

A list containing:

`theta`	The estimated effects for tissue-gene pairs, scaled by the outcome GWAS sample size.
`gamma`	The estimated effects for direct causal variants, scaled by the outcome GWAS sample size.
`theta.se`	Standard errors of the tissue-gene pair effects.
`gamma.se`	Standard errors of the direct causal variant effects.
`theta.pip`	Posterior inclusion probabilities (PIP) for tissue-gene pairs.
`gamma.pip`	Posterior inclusion probabilities (PIP) for direct causal variants.
`theta.pratt`	Pratt estimations for tissue-gene pairs.
`gamma.pratt`	Pratt estimations for direct causal variants.
`theta.cs`	Credible set indicators for tissue-gene pairs.
`gamma.cs`	Credible set indicators for direct causal variants.
`theta.cs.pip`	PIP within credible sets for tissue-gene pairs.
`gamma.cs.pip`	PIP within credible sets for direct causal variants.
`fit.causal`	The SuSiE fit object for the causal analysis.
`cs.summary`	A summary of the credible sets obtained from the analysis.
`pip.causal.samping`	The PIP resampling results from causal effect estimation.
`estimate.causal.sampling`	The effect size resampling results from causal effect estimation.
`bXest`	The matrix of resampled eQTL effects used in causal effect estimation.

---

tgvis *tgvis: Function to estimate and select the optimal number of single effects using profile-likelihood and BIC*

Description

This function estimates the number of single effects in a locus by combining profile-likelihood methods and Bayesian Information Criterion (BIC) to optimize the model. It includes resampling for estimating standard errors and performs score tests for infinitesimal effects.

Usage

```
tgvis(
  by,
  bXest,
  LD,
  Noutcome,
  L.causal.vec = c(1:8),
  max.iter = 50,
  max.eps = 0.001,
  susie.iter = 500,
```

```
  pip.thres.cred = 0.95,
  eigen.thres = 1,
  varinf.upper.boundary = 0.25,
  varinf.lower.boundary = 0.001,
  ebic.beta = 1,
  ebic.upsilon = 1,
  pip.min = 0.05,
  pv.thres = 0.05,
  pleiotropy.rm = NULL,
  prior.weight.theta = NULL,
  prior.weight.gamma = NULL
)
```

Arguments

by	A vector of Z-scores of the marginal effects from the outcome GWAS.
bXest	A matrix of direct effect estimates based on the Z-scores of tissue-gene pairs.
LD	The LD matrix of variants.
Noutcome	The sample size of the outcome GWAS.
L.causal.vec	A vector of candidate numbers of single effects used in BIC. Default is 'c(1:8)'.
max.iter	The maximum number of iterations for the profile-likelihood algorithm. Default is 50.
max.eps	The convergence tolerance for the profile-likelihood algorithm. Default is 1e-3.
susie.iter	The maximum number of iterations for 'susie_rss' within the profile-likelihood algorithm. Default is 50.
pip.thres.cred	The cumulative PIP threshold for variables in a credible set. Default is 0.95.
eigen.thres	The threshold of eigenvalues for modelling the infinitesimal effect. Default is 1.
varinf.upper.boundary	The upper boundary for the prior variance of infinitesimal effects, multiplied by var(y) to adapt to different locus variances. Default is 0.25.
varinf.lower.boundary	The lower boundary for the prior variance of infinitesimal effects, not multiplied by var(y). Default is 0.001.
ebic.beta	The extended BIC factor for causal effects of tissue-gene pairs and direct causal variants used in BIC computation. Default is 1.
ebic.upsilon	The extended BIC factor for infinitesimal effects used in BIC computation. Default is 1.
pip.min	The minimum PIP threshold for individual causal effects in the profile-likelihood. This is used to specify which tissue-gene pairs and direct causal variants to include in the score test of variance of infinitesimal effects. Default is 0.05.
pv.thres	The p-value threshold for the score test. Default is 0.05.
pleiotropy.rm	A vector of indices specifying which variants should not be considered as having direct causal effects.
prior.weight.theta	A vector of prior weights of gene-tissue pairs, which will be used as input in SuSiE. Default is NULL.
prior.weight.gamma	A vector of prior weights of direct causal variants, which will be used as input in SuSiE. Default is NULL.

Value

A list containing:

`theta`	The estimated effects for tissue-gene pairs, scaled by the outcome GWAS sample size.
`gamma`	The estimated effects for direct causal variants, scaled by the outcome GWAS sample size.
`theta.se`	Standard errors of the tissue-gene pair effects.
`gamma.se`	Standard errors of the direct causal variant effects.
`theta.pip`	Posterior inclusion probabilities (PIP) for tissue-gene pairs.
`gamma.pip`	Posterior inclusion probabilities (PIP) for direct causal variants.
`theta.pratt`	Pratt estimations for tissue-gene pairs.
`gamma.pratt`	Pratt estimations for direct causal variants.
`theta.cs`	Credible set indicators for tissue-gene pairs.
`gamma.cs`	Credible set indicators for direct causal variants.
`theta.cs.pip`	PIP within credible sets for tissue-gene pairs.
`gamma.cs.pip`	PIP within credible sets for direct causal variants.
`upsilon`	The estimated infinitesimal effects.
`var.upsilon`	The estimated variance of infinitesimal effects.
`fit.causal`	The SuSiE fit object for the causal analysis.
`cs.summary`	A summary of the credible sets obtained from the analysis.
`Bicvec`	A vector of BIC values for each candidate number of single effects.

Index