Package 'INTACT'

April 16, 2024

Type Package

Title Integrate TWAS and Colocalization Analysis for Gene Set Enrichment Analysis

Version 1.2.0

Description This package integrates colocalization probabilities from colocalization analysis with transcriptome-wide association study (TWAS) scan summary statistics to implicate genes that may be biologically relevant to a complex trait. The probabilistic framework implemented in this package constrains the TWAS scan z-score-based likelihood using a gene-level colocalization probability. Given gene set annotations, this package can estimate gene set enrichment using posterior probabilities from the TWAS-colocalization integration step.

Depends R (>= 4.2.0)

Imports SQUAREM, bdsmatrix, numDeriv, stats

License GPL-3 + file LICENSE

Encoding UTF-8

RoxygenNote 7.2.3

VignetteBuilder knitr

Suggests BiocStyle, knitr, rmarkdown, testthat

BugReports https://github.com/jokamoto97/INTACT/issues

URL https://github.com/jokamoto97/INTACT

biocViews Bayesian, GeneSetEnrichment

LazyData false

Config/testthat/edition 3

git_url https://git.bioconductor.org/packages/INTACT

git_branch RELEASE_3_18

git_last_commit 0b5c669

git_last_commit_date 2023-10-24

Repository Bioconductor 3.18

Date/Publication 2024-04-15

Author Jeffrey Okamoto [aut, cre] (<https://orcid.org/0000-0001-9988-1618>),
Xiaoquan Wen [aut] (<https://orcid.org/0000-0001-8990-2737>)

Maintainer Jeffrey Okamoto <jokamoto@umich.edu>

R topics documented:

`.em_est` *Compute gene set enrichment estimates.*

Description

Compute gene set enrichment estimates.

Usage

```
.em_est(pprobs, d_vec)
```

Arguments

`pprobs`	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.
`d_vec`	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.

Value

Maximum likelihood estimates for alpha0 and alpha1; convergence indicator.

.enrich_bootstrap_se *Compute bootstrap standard errors for alpha MLEs.*

Description

Compute bootstrap standard errors for alpha MLEs.

Usage

```
.enrich_bootstrap_se(pprobs, d_vec, reps = 100)
```

Arguments

pprobs	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.
d_vec	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.
reps	Number of bootstrap samples.

Value

MLEs for alpha0 and alpha1 from bootstrap samples.

.enrich_res *Compute gene set enrichment estimates with standard errors.*

Description

Compute gene set enrichment estimates with standard errors.

Usage

```
.enrich_res(sig_lev, pprobs, d_vec, SE_type = "NDS", boot_rep = NULL)
```

Arguments

sig_lev	A significance threshold for gene set enrichment hypothesis testing.
pprobs	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.
d_vec	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.
SE_type	A method to compute standard errors of the gene set enrichment estimates. Possible methods are "profile_likelihood," "bootstrap," and "NDS". NDS performs numerical differentiation of the Fisher score vector.
boot_rep	Number of bootstrap samples, if boostrap standard errors are specified for SE_type.

Value

A data frame with the alpha1 estimate, standard error, z-score, p-value, (1-sig_lev)% CI limits, and convergence indicator.

`.logistic_em` *A fixed-point mapping for the expectation-maximization algorithm. Used as an argument for fixptfn in the squarem function.*

Description

A fixed-point mapping for the expectation-maximization algorithm. Used as an argument for fixptfn in the squarem function.

Usage

```
.logistic_em(d_vec, pprobs, alpha)
```

Arguments

`d_vec`	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.
`pprobs`	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.
`alpha`	A vector containing the current estimates of the enrichment parameters alpha0 and alpha1$.

Value

Updated estimates of alpha0 and alpha1.

`.logistic_em_nopseudo` *Similar to logistic_em(), but does not use pseudocounts to stablize the algorithm.*

Description

Similar to logistic_em(), but does not use pseudocounts to stablize the algorithm.

Usage

```
.logistic_em_nopseudo(d_vec, pprobs, alpha)
```

Arguments

d_vec	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.
pprobs	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.
alpha	A vector containing the current estimates of the enrichment parameters alpha0 and alpha1.

Value

Updated estimates of alpha0 and alpha1.

.logistic_loglik *A log likelihood function for the expectation-maximization algorithm. Used as an argument for objfn in the squarem function.*

Description

A log likelihood function for the expectation-maximization algorithm. Used as an argument for objfn in the squarem function.

Usage

```
.logistic_loglik(alpha, d_vec, pprobs)
```

Arguments

alpha	A vector containing the current estimates of the enrichment parameters alpha0 and alpha1.
d_vec	A vector of gene set annotations for the genes of interest. Entries should be integer(1) if the gene is annotated and integer(0) otherwise.
pprobs	A vector of posterior probabilities for each gene estimated from the intact function. Gene order should match d_vec.

Value

Log likelihood evaluated at the current estimates of alpha0 and alpha1.

.pi1_fun *Estimate pi1 from TWAS scan z-scores.*

Description

Estimate pi1 from TWAS scan z-scores.

Usage

```
.pi1_fun(z_vec, lambda = 0.5)
```

Arguments

z_vec	A vector of TWAS scan z-scores.
lambda	A value between 0 and 1. The density of TWAS scan z-scores should be flat at lambda. Set to 0.5 as default.

Value

A scalar estimate for pi1.

expit *Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.*

Description

Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.

Usage

```
expit(GLCP, t = 0.05, D = 0.1, u = 1, thresholding = "hard")
```

Arguments

GLCP	A gene colocalization probability
t	A hard threshold for the GLCP. Values below this number will be shrunk to zero. Default is 0.05.
D	A curvature shrinkage parameter. Lower values of D will result in a steeper curve. Default is 0.1
u	A factor between 0 and 1 by which the prior function is scaled.
thresholding	An option to use hard thresholding or soft thresholding for the prior function. Default is "hard". For soft thresholding, set to "soft".

Value

The value of the prior.

Examples

```
expit(0.2, 0.05, 1)
```

`fdr_rst` *Bayesian FDR control for INTACT output*

Description

Bayesian FDR control for INTACT output

Usage

```
fdr_rst(posterior, alpha = 0.05)
```

Arguments

`posterior`	A vector of posterior probabilities for each gene estimated from the intact function.
`alpha`	A numeric target FDR control level.

Value

An n x 2 data frame where the first column is the inputted posterior probabilities, and the second is a Boolean vector denoting significance at the specified target control level.

Examples

```
data(simdat)
fdr_rst(simdat$GLCP)
```

gene_set_list *Simulated gene set list.*

Description

A list object containing two elements. Each is a character list of gene names.

Usage

```
gene_set_list
```

Format

A list with two items:

gene set 1 gene set with 503 gene members. Significantly enriched in simdat.

gene set 2 gene set with 200 members. ...

Examples

```
data(gene_set_list)
```

hybrid *Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.*

Description

Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.

Usage

```
hybrid(GLCP, t = 0.05, D = 0.1, u = 1, thresholding = "hard")
```

Arguments

GLCP	A gene colocalization probability
t	A hard threshold for the GLCP. Values below this number will be shrunk to zero. Default is 0.05.
D	A curvature shrinkage parameter. Lower values of D will result in a steeper curve. Default is 0.1
u	A factor between 0 and 1 by which the prior function is scaled.
thresholding	An option to use hard thresholding or soft thresholding for the prior function. Default is "hard". For soft thresholding, set to "soft".

Value

The value of the prior.

Examples

```
hybrid(0.2, 0.05, 1)
```

`intact` *Compute the posterior probability that a gene may be causal, given a gene's TWAS scan z-score (or Bayes factor) and colocalization probability.*

Description

Compute the posterior probability that a gene may be causal, given a gene's TWAS scan z-score (or Bayes factor) and colocalization probability.

Usage

```
intact(
  GLCP_vec,
  prior_fun = linear,
  z_vec = NULL,
  t = NULL,
  D = NULL,
  K = c(1, 2, 4, 8, 16),
  twas_priors = .pi1_fun(z_vec = z_vec, lambda = 0.5),
  twas_BFs = NULL
)
```

Arguments

`GLCP_vec`	A vector of colocalization probabilities for the genes of interest
`prior_fun`	A function to transform a colocalization probability into a prior. Options are linear, step, expit, and hybrid.
`z_vec`	A vector of TWAS scan z-scores for the genes of interest. The order of genes must match GLCP_vec.
`t`	A hard threshold for the GLCP. Values below this number will be shrunk to zero. This argument is used in the user-specified prior function. Default value for the step prior is 0.5. Default value is 0.05 for all other prior functions.
`D`	A curvature shrinkage parameter. Lower values of D will result in a steeper curve. Default is 0.1. This parameter should only be specified if the user selects the expit or hybrid prior function and does not wish to use the default value.
`K`	A vector of values over which Bayesian model averaging is performed.

twas_priors — An optional vector of user-specified gene-specific TWAS priors. If no input is supplied, INTACT computes a scalar prior using the TWAS data (see the corresponding manuscript for more details).

twas_BFs — A vector of TWAS Bayes factors for the genes of interest. This is an alternative option if the user wishes to directly specify Bayes factors instead of computing them from TWAS scan z-scores.

Value

The vector of posteriors.

Examples

```
data(simdat)
intact(GLCP_vec=simdat$GLCP, z_vec = simdat$TWAS_z)
intact(GLCP_vec=simdat$GLCP, prior_fun=expit, z_vec = simdat$TWAS_z,
t = 0.02,D = 0.09)
intact(GLCP_vec=simdat$GLCP, prior_fun=step, z_vec = simdat$TWAS_z,
t = 0.49)
intact(GLCP_vec=simdat$GLCP, prior_fun=hybrid, z_vec = simdat$TWAS_z,
t = 0.49,D = 0.05)
```

---

intactGSE *Perform gene set enrichment estimation and inference, given TWAS scan z-scores and colocalization probabilities.*

Description

Perform gene set enrichment estimation and inference, given TWAS scan z-scores and colocalization probabilities.

Usage

```
intactGSE(
  gene_data,
  prior_fun = linear,
  t = NULL,
  D = NULL,
  gene_sets,
  sig_lev = 0.05,
  SE_type = "NDS",
  boot_rep = NULL
)
```

Arguments

gene_data	A data frame containing gene names and corresponding colocalization probabilities and TWAS z-scores for each gene. Column names should be "gene", "GLCP", and "TWAS_z'. If the user wishes to specify TWAS Bayes factors instead of z-scores, use the column name "TWAS_BFs". If the user wishes to specify gene-specific TWAS priors, use the column name "TWAS_priors".
prior_fun	A function to transform a colocalization probability into a prior. Options are linear, step, expit, and hybrid.
t	A hard threshold for the GLCP. Values below this number will be shrunk to zero. This argument is used in the user-specified prior function. Default value for the step prior is 0.5. Default value is 0.05 for all other prior functions.
D	A curvature shrinkage parameter. Lower values of D will result in a steeper curve. Default is 0.1. This parameter should only be specified if the user selects the expit or hybrid prior function and does not wish to use the default value.
gene_sets	A named list of gene sets for which enrichment is to be estimated. List items should be character vectors of gene IDs. Gene ID format should match the gene column in gene_data.
sig_lev	A significance threshold for gene set enrichment hypothesis testing.
SE_type	A method to compute standard errors of the gene set enrichment estimates. Possible methods are "profile_likelihood" and "bootstrap."
boot_rep	Number of bootstrap samples.

Value

A data frame with the alpha1 estimate, standard error, z-score, p-value, (1-sig_lev)% CI limits, and convergence indicator for each gene set in gene_sets.

Examples

```
data(simdat)
data(gene_set_list)
intactGSE(gene_data = simdat,gene_sets = gene_set_list)
intactGSE(gene_data = simdat,prior_fun = step,t = 0.45,
gene_sets = gene_set_list)
intactGSE(gene_data = simdat,prior_fun = expit,t = 0.08,D = 0.08,
gene_sets = gene_set_list)
intactGSE(gene_data = simdat,prior_fun = hybrid,t = 0.08,D = 0.08,
gene_sets = gene_set_list)
```

`linear` *Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.*

Description

Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.

Usage

```
linear(GLCP, t = 0.05, u = 1, thresholding = "hard")
```

Arguments

GLCP	A gene colocalization probability
t	A hard threshold for the GLCP. Values below this number will be shrunk to zero. Default is 0.05.
u	A factor between 0 and 1 by which the prior function is scaled.
thresholding	An option to use hard thresholding or soft thresholding for the prior function. Default is "hard". For soft thresholding, set to "soft".

Value

The value of the prior.

Examples

```
linear(0.2, 0.05, 1)
linear(c(0.01,0.2,0.9))
```

---

simdat *Simulated TWAS and colocalization summary data.*

Description

A data set containing GLCP and TWAS z-score for 1197 simulated genes.

Usage

```
simdat
```

Format

A data frame with 1197 rows and 3 variables:

gene gene Ensembl ID

GLCP colocalization probability

TWAS_z TWAS z-score ...

Examples

```
data(simdat)
```

`step` *Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.*

Description

Transform a gene colocalization probability (GLCP) to a prior to be used in the evidence integration procedure. There are four prior function options, including expit, linear, step, and expit-linear hybrid.

Usage

```
step(GLCP, t = 0.5, u = 1)
```

Arguments

`GLCP`	A gene colocalization probability
`t`	A hard threshold for the GLCP. Values below this number will be shrunk to zero. Default is 0.5.
`u`	A factor between 0 and 1 by which the prior function is scaled.

Value

The value of the prior.

Examples

```
step(0.2, 0.05, 1)
```

Index