Package ‘omicwas’

February 19, 2020

Type Package

Title Cell-Type-Specific Association Testing in Bulk Omics Experiments

Version 0.3.1

Description In bulk epigenome/transcriptome experiments, molecular expression is measured in a tissue, which is a mixture of multiple types of cells. This package tests association of a disease/phenotype with a molecular marker for each cell type. The proportion of cell types in each sample needs to be given as input. The package is applicable to epigenome-wide association study (EWAS) of microarray experiments. Takeuchi and Kato (submitted)
``omicwas: cell-type-specific epigenome-wide and transcriptome association study``.

URL https://github.com/fumi-github/omicwas

BugReports https://github.com/fumi-github/omicwas/issues

Depends R (>= 3.6.0)

biocViews

License GPL-3

Encoding UTF-8

LazyData true

Imports broom, data.table, dplyr, ff, glmnet, magrittr, MASS, matrixStats, parallel, PCDimension, ridge, rlang, R.utils, sva, tidyr

RoxygenNote 7.0.2

Suggests testthat

NeedsCompilation no

Author Fumihiko Takeuchi [aut, cre] (<https://orcid.org/0000-0003-3185-5661>)

Maintainer Fumihiko Takeuchi <fumihiko@takeuchi.name>

Repository CRAN

Date/Publication 2020-02-19 09:42:00 UTC
R topics documented:

ctassoc ................................................................. 2
cctisQTL ................................................................. 4
ctrUV ................................................................. 6
GSE42861small ............................................................ 7
GSE79262small ............................................................ 7
rrs.fit ................................................................. 8

Index 10

ctassoc  Cell-Type-Specific Association Testing

Description

Cell-Type-Specific Association Testing

Usage

ctassoc(
  X,
  W,
  Y,
  C = NULL,
  test = "ridge",
  num.cores = 1,
  chunk.size = 1000,
  seed = 123
)

Arguments

X  Matrix (or vector) of traits; samples x traits.
W  Matrix of proportion of cell types; samples x cell types.
Y  Matrix (or vector) of bulk omics measurements; markers x samples.
C  Matrix (or vector) of covariates; samples x covariates. X, W, Y, C should be numeric.
test  Statistical test to apply; either "reducedrankridge", "ridge", "full" or "marginal".
num.cores  Number of CPU cores to use. Full and marginal tests are run in serial, thus num.cores is ignored.
chunk.size  The size of job for a CPU core in one batch. If you have many cores but limited memory, and there is a memory failure, decrease num.cores and/or chunk.size.
seed  Seed for random number generation.
Details

Let the indexes be \( h \) for cell type, \( i \) for sample, \( j \) for marker (CpG site or gene), \( k \) for each trait that has cell-type-specific effect, and \( l \) for each trait that has bulk tissue effect. The input data are \( X_{ik}, C_{il}, W_{ih} \) and \( Y_{ji} \), where \( C_{il} \) can be omitted. \( X_{ik} \) and \( C_{il} \) are the two types of traits, showing effects that are cell-type-specific or not, respectively. Thus, calling \( X_{ik} \) and \( C_{il} \) as "traits" and "covariates" gives a rough idea, but is not strictly correct. \( W_{ih} \) represents the cell type proportion and \( Y_{ji} \) represents the marker level, such as methylation or gene expression. For each tissue sample, the cell type proportion \( W_{ih} \) is the proportion of each cell type in the bulk tissue, which is measured or imputed beforehand. The marker level \( Y_{ji} \) in bulk tissue is measured and provided as input.

The cell-type-specific marker level \( Z_{hji} \) is not observed and is treated as a hidden variable. The parameters we estimate are the effect of cell-type-specific traits \( \beta_{hjk} \), the effect of non-specific traits \( \gamma_{jl} \), and the cell-type-specific basal marker level \( \mu_{hj} \).

We assume normal distribution for the cell-type-specific marker level,

\[
Z_{hji} \sim N(\mu_{hj} + \sum_k \beta_{hjk} * X_{ik}, \sigma_{hj}^2).
\]

Since the bulk tissue marker level is the sum weighted by \( W_{ih} \),

\[
Y_{ji} \sim N(\sum_h W_{ih} [\mu_{hj} + \sum_k \beta_{hjk} * X_{ik}] + \sum_l \gamma_{jl} * C_{il}, \tau_{j}^2).
\]

Although formally, the variance comprises of components of cell type level and tissue level, we approximate and unify into \( \tau_{j}^2 \).

The full model is the linear regression

\[
x_{ji} (\sum_h \mu_{hj} * W_{ih}) + (\sum_{hk} \beta_{hjk} * W_{ih} * X_{ik}) + (\sum_l \gamma_{jl} * C_{il}) + error.
\]

The ridge model aims to cope with multicollinearity of the interacting terms \( W_{ih} * X_{ik} \). It first adjusts for \( \mu_{hj} \) and \( \gamma_{jl} \) by fitting linear regression and taking the residuals. Afterwards, ridge regression is used to fit \( \beta_{hjk} \). We use the linearRidge function of the ridge package. The marginal model tests the trait association only in one cell type \( h \), under the linear regression,

\[
x_{ji} (\sum_{h'} \mu_{h'j} * W_{ih'}) + (\sum_k \beta_{hjk} * W_{ih} * X_{ik}) + (\sum_l \gamma_{jl} * C_{il}) + error.
\]

Value

A list with one element, which is named "coefficients". The element gives the estimate, statistic, p.value in tibble format.

See Also

ctrUV
**Examples**

data(GSE42861small)
X = GSE42861small$X
Y = GSE42861small$Y
W = GSE42861small$W
C = GSE42861small$C
Y = ctRUV(X, W, Y, C = C)
result = ctassoc(X, W, Y, C = C)
result$coefficients

---

**ctcisQTL**  
*Cell-Type-Specific QTL analysis*

---

**Description**

Cell-Type-Specific QTL analysis

**Usage**

```r
ctcisQTL(  
  X,  
  Xpos,  
  W,  
  Y,  
  Ypos,  
  C = NULL,  
  max.pos.diff = 1e+06,  
  nPC = NULL,  
  outdir = tempdir(),  
  outfile = "ctcisQTL.out.txt",  
  num.cores = 1,  
  seed = 123  
)
```

**Arguments**

- **X**  
  Matrix (or vector) of SNP genotypes; SNPs x samples.

- **Xpos**  
  Vector of the physical position of X

- **W**  
  Matrix of proportion of cell types; samples x cell types.

- **Y**  
  Matrix (or vector) of bulk omics measurements; markers x samples.

- **Ypos**  
  Vector of the physical position of Y

- **C**  
  Matrix (or vector) of covariates; samples x covariates. X, Xpos, W, Y, Ypos, C should be numeric.
**max.pos.diff** Maximum positional difference to compute cis-QTL. Association is computed between a row of X and a row of Y, only when they are within this limit. Since the limiting is only by position, the function needs to be run separately for each chromosome.

**nPC** A hyperparameter that (indirectly) specifies the penalty for ridge regression. If NULL, it is inferred from the data. A unique hyperparameter is applied to all SNP-marker pairs.

**outdir** Output directory.

**outfile** Output file.

**num.cores** Number of CPU cores to use. Full and marginal tests are run in serial, thus num.cores is ignored.

**seed** Seed for random number generation.

**Details**

The functionality is almost the same as the ridge test in ctaassoc. Association is tested between each row of Y and each row of X. Usually, the former will be a methylation/expression marker, and the latter will be a SNP. To cope with the large number of combinations, the testing is limited to pairs whose position is within the difference specified by max.pos.diff In addition, a uniform hyperparameter of ridge regression is used for all pairs. We use the linearRidge function of the ridge package.

**Value**

The estimate, statistic, p.value are written to the specified file.

**See Also**

cRUV

**Examples**

data(GSE79262small)
X = GSE79262small$X
Xpos = GSE79262small$Xpos
W = GSE79262small$W
Y = GSE79262small$Y
Ypos = GSE79262small$Ypos
C = GSE79262small$C
Y = Y[seq(1, 601, 20), ] # for brevity
Ypos = Ypos[seq(1, 601, 20)]
ctcisQTL(X, Xpos, W, Y, Ypos, C = C)
ctRUV

Remove Unwanted Variations prior to applying ctassoc

Description

Remove Unwanted Variations prior to applying ctassoc

Usage

ctrUV(X, W, Y, C = NULL, method = "PCA", nPC = NULL)

Arguments

X Matrix (or vector) of traits; samples x traits.
W Matrix of proportion of cell types; samples x cell types.
Y Matrix (or vector) of bulk omics measurements; markers x samples.
C Matrix (or vector) of covariates; samples x covariates. X, W, Y, C should be numeric.
method "PCA" or "SVA"
nPC Number of PCs to be regarded as unwanted variation. If NULL, automatically computed by the Auer-Gervini approach.

Details

First, for each marker, the full linear model of the ctassoc function is fitted, and the residual is computed. For the residuals over all markers, the principal components (PCs) are computed. The top PCs are regarded as the unwanted variations, and subtracted from Y.

Value

Y adjusted for the unwanted variations.

See Also

tassoc
Small Subset of GSE42861 Dataset From GEO

Description
The dataset includes 336 rheumatoid arthritis cases and 322 controls. A subset of 500 CpG sites were randomly selected from the original EWAS dataset.

Usage
data(GSE42861small)

Format
An object of class list of length 4.

Source
GEO

See Also
cassoc

Examples
data(GSE42861small)
X = GSE42861small$X
Y = GSE42861small$Y
Y = Y[seq(1, 20), ] # for brevity
W = GSE42861small$W
C = GSE42861small$C
result = cassoc(X, W, Y, C = C)
result$coefficients

Small Subset of GSE79262 Dataset From GEO

Description
The dataset includes 53 samples. A subset of 737 CpG sites and 3624 SNPs within Chr1:100,000,000-110,000,000 were selected from the original EWAS dataset. DNA methylation was measured in T cells. The estimated proportion of CD4T, CD8T, NK cells are saved in W.

Usage
data(GSE79262small)
rrs.fit

Fitting reduced-rank ridge regression with given rank and shrinkage penalty

Description

Fitting reduced-rank ridge regression with given rank and shrinkage penalty. This is a modification of rrs.fit in rrpack version 0.1-6. In order to handle extremely large \( q = \text{ncol}(Y) \), generation of a \( q \) by \( q \) matrix is avoided.

Usage

\[
\text{rrs.fit}(Y, X, \text{nrank} = \text{min}(\text{ncol}(Y), \text{ncol}(X)), \text{lambda} = 1, \text{coefSVD} = \text{FALSE})
\]

Arguments

- **Y**: a matrix of response (\( n \) by \( q \))
- **X**: a matrix of covariate (\( n \) by \( p \))
- **nrank**: an integer specifying the desired rank
- **lambda**: tunning parameter for the ridge penalty
- **coefSVD**: logical indicating the need for SVD for the coeffient matrix int the output
Value

S3 `rrr` object, a list consisting of

- `coef` : coefficient of rrs
- `coef.ls` : coefficient of least square
- `fitted` : fitted value of rrs
- `fitted.ls` : fitted value of least square
- `A` : right singular matrix
- `Ad` : singular value vector
- `nrank` : rank of the fitted rrr

References


Examples

```r
Y <- matrix(rnorm(400), 100, 4)
X <- matrix(rnorm(800), 100, 8)
rfit <- rrs.fit(Y, X)
```
Index

+Topic **datasets**
  GSE42861small, 7
  GSE79262small, 7

ctassoc, 2
ctcisQTL, 4
ctrUV, 6

GSE42861small, 7
GSE79262small, 7

linearRidge, 3, 5

rrs.fit, 8