Package 'MultiABEL'

June 15, 2017

Type Package

Title Multi-Trait Genome-Wide Association Analysis

Version 1.1-6

Date 2017-06-14

Author Xia Shen

Maintainer Xia Shen <xia.shen@ed.ac.uk>

Description Multivariate genome-wide association analyses. The analysis can be performed on individual-level data or multiple single-trait genome-wide summary statistics.

Depends R (>= 2.10), svMisc, data.table

Suggests GenABEL, DatABEL

License GPL (>= 2)

URL https://github.com/xiashen/MultiABEL

LazyLoad yes

RoxygenNote 6.0.1

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-06-15 17:40:16 UTC

R topics documented:

load.summary *Loading multiple summary statistics from genome-wide association studies*

Description

The function loads multiple meta-GWAS summary statistics, for subsequent multi-trait GWAS. Currently, the package only analyzes summary statistics from inverse-Gaussianized continuous traits.

Usage

```
load.summary(files, cor.pheno = NULL, indep.snps = NULL, est.var = FALSE,
  columnNames = c("snp", "a1", "a2", "freq", "beta", "se", "n"),
  fixedN = NULL)
```

Arguments

files	A vector of file names as strings. Each file name should contain summary statistics of one trait to be included in the multi-trait analysis. The columns of the summary statistics have to contain (uppercase or lowercase does not matter) 'snp' (marker ID), 'a1' (the first allele), 'a2' (the second allele), 'freq', (frequency of the first allele), 'beta' (effect size), 'se' (standard error), and 'n' (sample size).
cor.pheno	A #traits x #traits matrix of correlation matrix of the phenotypes, to be used to construct the multi-trait test statistic. If NULL, this matrix will be estimated from genome-wide summary statistics. If you have partially overlapping samples for different traits, shrinkage correlation matrix is recommended (see reference), so in that case, unless you know what you are doing, leave this argument as default, i.e. NULL.
indep.snps	A vector of strings containing the names of a set of independent SNPs. This is recommended to be generated by LD-pruning the genotype data in a certain cohort. Typically the number of SNPs should be more than 10,000 in order to obtain a good estimate of cor.pheno. If cor.pheno = NULL, this argument cannot be NULL.
est.var	A logical value. If FALSE, each phenotypic variance is assumed to be known as 1. If TRUE, each phenotypic variance will be estimated to adjust the summary statistics, so that the corresponding phenoypic variance is 1.
columnNames	A vector with names of columns containing necessary information in the input file; default values are c('snp','a1','a2','freq','beta','se','n'). The values are case-insensitive. Note: check your allele definitions for different traits are based on the same strand!
fixedN	sample size to assume across all analyses, when provided, this number will be used (instead of the ones specified in the input files)

Value

The function returns a list of class `multi.summary`, containing two elements: `gwa` (the cleaned data to be processed in multi-trait GWAS), `cor.pheno` (user input or estimated), and `var.pheno` (default or estimated).

Author(s)

Xia Shen, Yakov, Tsepilov, Yurii S. Aulchenko

References

Xia Shen, Yakov Tsepilov, ..., Yudi Pawitan, Chris S. Haley, Yurii S. Aulchenko (2017). Discovery, replication, and in silico functional investigation of 22 new pleiotropic anthropometric loci. *Submitted*.

See Also

`MultiSummary`

Examples

```
## Not run:
## download the six example files from:
## https://www.dropbox.com/sh/hhta45cewvvea2s/AADfj4OXlbroToZAwIii2Buha?dl=0
## the summary statistics from Randall et al. (2013) PLoS Genet
## for males only
## bmi: body mass index
## hip: hip circumference
## wc: waist circumference
## whr: waist-hip ratio

## load the prepared set of independent SNPs
indep.snps <- as.character(read.table('indep.snps')$V1)

## load summary statistics of the six traits
stats.male <- load.summary(files = list.files(pattern = '*.txt'), indep.snps = indep.snps)

## perform multi-trait meta-GWAS
result <- MultiSummary(stats.male)
head(result)

## End(Not run)
```

`MultiABEL` *Multivariate GWAS in R*

Description

MultiABEL: Multivariate Genome-Wide Association Analyses

Details

Performing multivariate genome-wide association (MVGWA) analyses. The modules are compatible with existing *ABEL data formats. The GWA analyses can be done on individual level data or on single-trait GWA summary statistics only.

For converting data from other formats, see

`{convert.snp.illumina}` (Illumina/Affymetrix-like format). This is our preferred converting function, very extensively tested. Other conversion functions include: `{convert.snp.text}` (conversion from human-readable GenABEL format), `{convert.snp.ped}` (Linkage, Merlin, Mach, and similar files), `{convert.snp.mach}` (Mach-format), `{convert.snp.tped}` (from PLINK TPED format), `{convert.snp.affymetrix}` (BRML-style files).

For converting of GenABEL's data to other formats, see `{export.merlin}` (MERLIN and MACH formats), `{export.impute}` (IMPUTE, SNPTEST and CHIAMO formats), `{export.plink}` (PLINK format, also exports phenotypic data).

To load the data, see `{load.gwaa.data}`.

For conversion to DatABEL format (used by ProbABEL and some other GenABEL suite packages), see `{impute2databel}`, `{impute2mach}`, `{mach2databel}`.

For data managment and manipulations see `{merge.gwaa.data}`, `{merge.snp.data}`, `{gwaa.data-class}`, `{snp.data-class}`, `{snp.names}`, `{snp.subset}`.

Author(s)

Xia Shen

References

If you use the MultiABEL package in your analysis, please cite the papers in `citation("MultiABEL")`.

See Also

`GenABEL`, `DatABEL`

`MultiLoad` *Load individual-level data for multivariate GWA analysis*

Description

The function imports GenABEL (gwaa.data class) or DatABEL (.fv*) data formats to perform multivariate test for each genetic variant.

Usage

```
MultiLoad(gwaa.data = NULL, phenofile = NULL, genofile = NULL, trait.cols,
  covariate.cols = NULL, cuts = 20, impute = TRUE, gaussianize = TRUE,
  ...)
```

Arguments

gwaa.data	An (optional) object of {gwaa.data-class}.
phenofile	An (optional) plain text file contains phenotypic outcomes and covariates.
genofile	An (optional) object of {databel-class} containing genotype data.
trait.cols	A vector (length > 1) giving the column indices of the phenotypes to be analyzed.
covariate.cols	An (optional) vector giving the column indices of the covariates to be included.
cuts	An integer telling how many pieces the genotype data matrix will be splitted for analyze. The value can be set depending on the memory size. The smaller the value is, potentially the faster the analysis will be.
impute	An (optional) logical argument telling whether missing genotypes should be imputed.
gaussianize	An (optional) logical argument telling whether the phenotypes should be gaussianized via inverse-Gaussian transformation.
...	not used.

Value

The function returns a list of cleaned statistics for subsequent, with class multi.loaded.

Note

Either gwaa.data (for GenABEL data format) or the combination of phenofile and genofile (for DatABEL data format) has to be provided. If all are provided, only phenofile and genofile will be used. When using DatABEL format input, individual IDs in phenofile and genofile have to match!

Author(s)

Xia Shen

References

Xia Shen, ..., Jim Wilson, Gordan Lauc, Yurii Aulchenko (2015). Multi-omic-variate analysis identified novel loci associated with compound N-Glycosylation of human Immunoglobulin G. *Submitted*.

See Also

Multivariate

Examples

```
## loading example gwaa.data in GenABEL
require(GenABEL)
data(ge03d2ex.clean)
```

```
## running multivariate GWAS for 3 traits: height, weight, bmi
loaded <- MultiLoad(gwaa.data = ge03d2ex.clean, trait.cols = c(5, 6, 8),
                    covariate.cols = c(2, 3))

## Not run:
## converting the same dataset into DatABEL format files
require(DatABEL)
write.table(phdata(ge03d2ex.clean), 'pheno.txt', col.names = TRUE, row.names = TRUE,
            quote = FALSE, sep = '\t')
geno <- as.double(ge03d2ex.clean)
matrix2databel(geno, 'geno')

## running the multivariate GWAS again
loaded <- MultiLoad(phenofile = 'pheno.txt', genofile = 'geno', trait.cols = c(5, 6, 8),
                    covariate.cols = c(2, 3))

## End(Not run)
```

`MultiMeta` *Meta-analysis for multivariate genome-wide association scan*

Description

The function performs meta-analysis for multiple multivariate GWA analyses

Usage

```
MultiMeta(reslist, outfile = "Multivariate_meta-analysis_results.txt")
```

Arguments

`reslist`	A list where each element is a multivariate GWA result of class "MultiRes".
`outfile`	A string giving the path and file name of the output file. By default, a file named `'Multivariate_meta-analysis_results.txt'` will be written into the current working directory.

Value

The function returns a matrix containing the meta-analysis results, where the row names are the variants names, and the column names are the names of the studies provided in `reslist` or generated by the program if no names are given, with an extra column `"p.meta"` containing the meta-analysis P-values. The results are also written into `outfile`.

Author(s)

Xia Shen

References

Xia Shen, ..., Gordan Lauc, Jim Wilson, Yurii Aulchenko (2014). Multi-omic-variate analysis identified the association between 14q32.33 and compound N-Glycosylation of human Immunoglobulin G *Submitted*.

See Also

`Multivariate`

Examples

```
## Not run:
## loading two gwaa.data sets in GenABEL
data(ge03d2)
data(ge03d2ex)

## in each dataset, running multivariate GWAS for 3 traits: height, weight, bmi
res1 <- Multivariate(gwaa.data = ge03d2, trait.cols = c(5, 6, 8),
                     covariate.cols = c(2, 3))
res2 <- Multivariate(gwaa.data = ge03d2ex.clean, trait.cols = c(5, 6, 8),
                     covariate.cols = c(2, 3))

## running meta-analysis by combining the P-values
meta <- MultiMeta(list(res1, res2))

## End(Not run)
```

`MultiRep` *Replication analysis of multivariate genome-wide association signal*

Description

The function performs replication analysis of multivariate GWA signals.

Usage

```
MultiRep(training.pheno = NULL, training.phenofile = NULL,
  test.pheno = NULL, test.phenofile = NULL, pheno.names = NULL,
  training.geno, test.geno)
```

Arguments

`training.pheno`	An (optional) matrix or data frame contains the phenotype data for the discovery sample, preferrably adjusted for fixed effects and population structure before multivariate GWA analysis.
`training.phenofile`	An (optional) plain text file contains phenotypes for the discovery sample. If this is provided, it will serve as `training.pheno`.

test.pheno	An (optional) matrix or data frame contains the phenotype data for the replication sample, preferrably adjusted for fixed effects and population structure.
test.phenofile	An (optional) plain text file contains phenotypes of the replication sample. If this is provided, it will serve as test.pheno.
pheno.names	A vector (length > 1) giving the column names of the phenotypes to be analyzed.
training.geno	A matrix or data.frame that contains the discovery sample genotype dosages of the variants to replicate.
test.geno	A matrix or data.frame that contains the replication sample genotype dosages of the variants to replicate. This object should have the same column names and order as training.geno.

Value

The function returns a list of 3 matrices. $replication contains the estimate of variant effect on the corresponding compound phenotype (beta_c), standard error (s.e.), replication P-value (P), and proportion of phenotypic variance explained (R-squared). $training.coef contains the estimated coefficients in the discovery sample of each phenotype for each variant to construct the compound phenotype. $test.coef contains similar coefficients as in $training.coef but estimated in the replication sample, but these are just for the record, NOT used in the replication procedure.

Note

Either .pheno or .phenofile has to be provided. If both are provided, only phenofile will be used. Individual IDs in .pheno or .phenofile and .geno have to match!

Author(s)

Xia Shen

References

Xia Shen, ..., Gordan Lauc, Jim Wilson, Yurii Aulchenko (2014). Multi-omic-variate analysis identified the association between 14q32.33 and compound N-Glycosylation of human Immunoglobulin G *Submitted*.

See Also

Multivariate

Examples

```
## Not run:
## loading example discovery sample gwaa.data in GenABEL
data(ge03d2)

## running multivariate GWAS for 3 traits: height, weight, bmi
res <- Multivariate(gwaa.data = ge03d2, trait.cols = c(5, 6, 8),
                    covariate.cols = c(2, 3))
```

```
## extracting 5 significant variants
(top <- res[order(res[,'P.F']),][2:6,])
snps <- rownames(top)
training.geno <- as.double(gtdata(ge03d2)[,snps])

## loading example test sample gwaa.data in GenABEL
data(ge03d2c)

## extracting genotypes of the 5 variants
test.geno <- as.double(gtdata(ge03d2c)[,snps])

## try replication
rep <- MultiRep(training.pheno = phdata(ge03d2), test.pheno = phdata(ge03d2c),
                pheno.names = c('height', 'weight', 'bmi'),
                   training.geno = training.geno, test.geno = test.geno)

## End(Not run)
```

`MultiSummary` *Multivariate genome-wide association scan using summary statistics*

Description

This function performs multivariate GWA analysis using meta-GWAS summary statistics

Usage

```
MultiSummary(x, index = NULL, type = "direct", vars = NULL)
```

Arguments

`x`	A data object of class `multi.summary` loaded by the function `load.summary`.
`index`	A numeric vector that gives the indices of the traits to be analyzed jointly.
`type`	A string gives the type of analysis. Default is `"outbred"`, referring to general outbred populations, following Hardy-Weinberg equilibrium. `"inbred"` refers to inbred populations, where no heterzygotes exists, namely, allele frequency = genotype frequency. `"precise"` refers to precise test statistics, especially when the individual-level data are available, for which the argument `vars` has to be given. `"direct"` refers to test statistics directly constructed from the T-statistics in univariate GWAS, this provides a scale-invariant test most similar to the direct MANOVA, but may be less powerful in some scenarios.
`vars`	A numeric vector gives the variance of the genotypes at each SNP, e.g. coded as 0, 1 and 2. Only used when `type = "precise"`.

Value

The function returns a data frame containing the multi-trait GWAS results, where the row names are the variants names. The column names are: variant name (`Marker`), allele frequency (`Freq`), the smallest sample size of the traits (`N`), effect on the phenotype score (`Beta.S`, see reference), standard error (`SE`), p-value (`P`), and the rest the coefficients to construct the phenotype score (see reference).

Author(s)

Xia Shen

References

Xia Shen, Zheng Ning, Yakov Tsepilov, Peter K. Joshi, James F. Wilson, Yudi Pawitan, Chris S. Haley, Yurii S. Aulchenko (2016). Fast pleiotropic meta-analysis for genetic studies. *Submitted*.

See Also

`load.summary`

Examples

```
## Not run:
## download the six example files from:
## https://www.dropbox.com/sh/hhta45cewvvea2s/AADfj4OXlbroToZAwIii2Buha?dl=0
## the summary statistics from Randall et al. (2013) PLoS Genet
## for males only
## bmi: body mass index
## hip: hip circumference
## wc: waist circumference
## whr: waist-hip ratio

## load the prepared set of independent SNPs
indep.snps <- as.character(read.table('indep.snps')$V1)

## load summary statistics of the six traits
stats.male <- load.summary(files = c('bmi.txt', 'height.txt',
                           'weight.txt', 'hip.txt', 'wc.txt',
                           'whr.txt'), indep.snps = indep.snps)

## perform multi-trait meta-GWAS
result <- MultiSummary(stats.male)
head(result)

## End(Not run)
```

`Multivariate` *Multivariate genome-wide association scan*

Description

The function imports GenABEL (gwaa.data class) or DatABEL (.fv*) data formats and performs multivariate test for each genetic variant using multivariate analysis of variance (MANOVA).

Usage

```
Multivariate(x, trait.idx = NULL, ...)
```

Arguments

`x`	An object created by `MultiLoad`.
`trait.idx`	A vector giving the indices of traits to be analyzed.
`...`	not used.

Value

The function returns a data frame containing the multi-trait GWAS results, where the row names are the variants names. The column names are: variant name (`Marker`), allele frequency (`Freq`), the smallest sample size of the traits (`N`), effect on the phenotype score (`Beta.S`, see reference), standard error (`SE`), p-value (`P`), and the rest the coefficients to construct the phenotype score (see reference).

Note

Either `gwaa.data` (for GenABEL data format) or the combination of `phenofile` and `genofile` (for DatABEL data format) has to be provided. If all are provided, only `phenofile` and `genofile` will be used. When using DatABEL format input, individual IDs in `phenofile` and `genofile` have to match!

Author(s)

Xia Shen

References

Xia Shen, ..., Jim Wilson, Gordan Lauc, Yurii Aulchenko (2015). Multi-omic-variate analysis identified novel loci associated with compound N-Glycosylation of human Immunoglobulin G. *Submitted*.

See Also

`MultiLoad`

Examples

```
## loading example gwaa.data in GenABEL
require(GenABEL)
data(ge03d2ex.clean)

## running multivariate GWAS for 3 traits: height, weight, bmi
loaded <- MultiLoad(gwaa.data = ge03d2ex.clean, trait.cols = c(5, 6, 8),
                    covariate.cols = c(2, 3))

## running the multivariate GWAS
res <- Multivariate(loaded)
```

Index