Package 'MPGE'

January 20, 2025

Type Package

Title A Two-Step Approach to Testing Overall Effect of
Gene-Environment Interaction for Multiple Phenotypes

Version 1.0.0

Date 2020-10-14

Description Interaction between a genetic variant (e.g., a single nucleotide polymorphism) and an environmental variable (e.g., physical activity) can have a shared effect on multiple phenotypes (e.g., blood lipids). We implement a two-step method to test for an overall interaction effect on multiple phenotypes. In first step, the method tests for an overall marginal genetic association between the genetic variant and the multivariate phenotype. The genetic variants which show an evidence of marginal overall genetic effect in the first step are prioritized while testing for an overall gene-environment interaction effect in the second step. Methodology is available from: A Majumdar, KS Burch, S Sankararaman, B Pasaniuc, WJ Gauderman, JS Witte (2020) <doi:10.1101/2020.07.06.190256>.

Depends R (>= 3.5.0)

License GPL-3

Encoding UTF-8

LazyData true

URL `https://github.com/ArunabhaCodes/MPGE`

BugReports `https://github.com/ArunabhaCodes/MPGE/issues`

RoxygenNote 7.1.1

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

Imports car, purrr, stats, utils

NeedsCompilation no

Author Arunabha Majumdar [aut, cre],
Tanushree Haldar [aut]

Maintainer Arunabha Majumdar <statgen.arunabha@gmail.com>

Repository CRAN

Date/Publication 2020-10-23 15:40:07 UTC

Contents

`environment_data` *An example of data of the environmental variable (e.g., smoking status). Here, environment_data is a data frame with single column for the environmental variable. The order of the 500 individuals in the row must be the same as provided in the phenotype and genotype data. Here, the environmental variable has two categories which were coded as 1 and 0 (e.g., smokers and non-smokers). Instead of numeric values, these can also be considered to be factors in the absence of a defined order in the categories.*

Description

An example of data of the environmental variable (e.g., smoking status). Here, environment_data is a data frame with single column for the environmental variable. The order of the 500 individuals in the row must be the same as provided in the phenotype and genotype data. Here, the environmental variable has two categories which were coded as 1 and 0 (e.g., smokers and non-smokers). Instead of numeric values, these can also be considered to be factors in the absence of a defined order in the categories.

Usage

```
data(environment_data)
```

Format

A data.frame with single column for the environmental variable. The order of the 500 individuals in the row must be the same as provided in the phenotype and genotype data:

Examples

```
data(environment_data)
geno <- environment_data
```

genotype_data *An example of genotype data for two genetic variants (SNPs). Here, genotype_data is a data.frame with the columns as SNPs (e.g., rs1 and rs2 here). The rows correspond to the 500 individuals in the same order as in the phenotype data.*

Description

An example of genotype data for two genetic variants (SNPs). Here, genotype_data is a data.frame with the columns as SNPs (e.g., rs1 and rs2 here). The rows correspond to the 500 individuals in the same order as in the phenotype data.

Usage

```
data(genotype_data)
```

Format

A data.frame with the columns as SNPs (e.g., rs1 and rs2 here) and individuals in the rows with the same order as in the phenotype data:

Examples

```
data(genotype_data)
geno <- genotype_data
```

MPGE *MPGE: an R package to implement a two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes.*

Description

Interaction between a genetic variant (e.g., a SNP) and an environmental variable (e.g., physical activity) can have a shared effect on multiple phenotypes (e.g., LDL and HDL). MPGE is a two-step method to test for an overall interaction effect on multiple phenotypes. In first step, the method tests for an overall marginal genetic association between the genetic variant and the multivariate phenotype. In the second step, SNPs which show an evidence of marginal overall genetic effect in the first step are prioritized while testing for an overall GxE effect. That is, a more liberal threshold of significance level is considered in the second step while testing for an overall GxE effect for these promising SNPs compared to the other SNPs.

Details

The package consists of following functions: `mv_G_GE`, `WHT`; `SST`.

Functions

- mv_G_GE for a batch of genetic variants, this function provides two different p-values for each genetic variant, one from the test of marginal overall genetic association with multiple phenotypes , and the other from the test of overall GxE effect on multivariate phenotype allowing for a possible marginal effect due to the genetic variant and a marginal effect due to the environmental variable.
- WHT this function implements the weighted multiple hypothesis testing procedure to adjust for multiple testing while combining the two steps of testing gene-environment interaction in the two-step GxE testing procedure, given two sets of p-values obtained using the previous function mv_G_GE for genome-wide genetic variants.
- SST this function implements the subset multiple hypothesis testing procedure to adjust for multiple testing while combining the two steps of testing gene-environment interaction based on the same two sets of p-values described above.

References

A Majumdar, KS Burch, S Sankararaman, B Pasaniuc, WJ Gauderman, JS Witte (2020) A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. bioRxiv, doi: https://doi.org/10.1101/2020.07.06.190256

mv_G_GE *Test for marginal overall genetic association with multivariate phenotype, and test for overall GxE effect on the multivariate phenotype in presence of marginal effect due to the genetic variant and a marginal effect due to the environmental variable.*

Description

Run mv_G_GE to obtain two different sets of p-values, one from the test for marginal overall genetic association with multiple phenotypes (using multivariate linear regression), and the other from the test of overall GxE effect on multivariate phenotype allowing for a possible genetic effect due to the genetic variant and an effect due to the environmental variable.

Usage

```
mv_G_GE(pheno, geno, env)
```

Arguments

pheno A numeric matrix or data.frame with the number of individuals (n) as the number of rows and the number of phenotypes (k) as the number of columns. It contains the values of k phenotypes (e.g. LDL and HDL) across the individuals. Each phenotype (e.g. LDL) must be individually adjusted for relevant covariates (age, sex, principal components of genetic ancestries, etc) beforehand. Therefore, each column of pheno matrix should be the adjusted residuals of the corresponding phenotype. Each final phenotype (column) should be continuous and normally distributed. No default.

`geno` A numeric matrix/data.frame (for a batch of genetic variants), or a numeric vector (for a single genetic variant). It contains the genotype values of the genetic variants/variant across the individuals. For a batch of variants, columns correspond to variants, and rows correspond to n individuals. For a SNP, three different ways of genotype coding are possible: additive, dominant and recessive, where additive coding is more common. No default.

`env` A vector of length n (number of individuals). It contains the values of the environmental variable (e.g., frequency of alcohol consumption). It can also contain factors, e.g., "yes" or "no" smoking status.

Value

The output is a data.frame with three columns. First column is the name of the SNPs or genetic variants. The main columns are as follows:

`G.P` P value of testing multivariate marginal genetic association between the genetic variant and the vector of phenotypes.

`GE.P` P value of testing multivariate overall GxE effect in presence of possible marginal effect due to the genetic variant and marginal effect due to the environmental variable.

References

A Majumdar, KS Burch, S Sankararaman, B Pasaniuc, WJ Gauderman, JS Witte (2020) A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. bioRxiv, doi: https://doi.org/10.1101/2020.07.06.190256

See Also

`WHT`, `SST`

`mv_G_GxE_pvalues` *An example of step 1 (marginal genetic association) and step 2 (GxE interaction) p-values across genetic variants (SNPs). Here, mv_G_GxE_pvalues is a data.frame with three columns. First column lists the set of 1000 genetic variants. Second column provides the vector of p-values obtained from testing the marginal multivariate genetic association for these SNPs. And the third column provides the vector of p-values obtained from testing the overall GxE effect in presence of possible marginal genetic effect and marginal environmental effect.*

Description

An example of step 1 (marginal genetic association) and step 2 (GxE interaction) p-values across genetic variants (SNPs). Here, mv_G_GxE_pvalues is a data.frame with three columns. First column lists the set of 1000 genetic variants. Second column provides the vector of p-values obtained from testing the marginal multivariate genetic association for these SNPs. And the third column provides the vector of p-values obtained from testing the overall GxE effect in presence of possible marginal genetic effect and marginal environmental effect.

Usage

```
data(mv_G_GxE_pvalues)
```

Format

A data.frame with three columns. First column lists the set of 1000 genetic variants. Second column provides the vector of p-values obtained from testing the marginal multivariate genetic association for these SNPs. And the third column provides the vector of p-values obtained from testing the overall GxE effect in presence of possible marginal genetic effect and marginal environmental effect:

Examples

```
data(mv_G_GxE_pvalues)
geno <- mv_G_GxE_pvalues
```

phenotype_data *An example of phenotype data.*

Description

Here phenotype_data is a data.frame with three columns for three phenotypes and the number of rows to be the number of individuals in the sample (500 in this toy data). Data for each phenotype provided must be adjusted individually for relevant covariates (e.g., age, sex, genetic ancestry) beforehand, and should follow a normal distribution.

Usage

```
data(phenotype_data)
```

Format

A numeric matrix or data.frame with three columns for three phenotypes and 500 rows for the individuals in the sample.

Examples

```
data(phenotype_data)
pheno <- phenotype_data
```

SST *Subset multiple hypothesis testing procedure to combine two steps of testing gene-environment interaction in a two-step procedure.*

Description

Run `SST` to adjust for multiple testing while combining two steps of the GxE interaction testing procedure. The procedure is applicable for a multivariate phenotype, as well as a univariate phenotype.

Usage

```
SST(PVAL, Pg_thr_step1 = 0.005, FWER_step2 = 0.05)
```

Arguments

`PVAL` A data.frame with three columns. The first column (PVAL$SNP) provides the name of all SNPs or genetic variants tested. Second column (PVAL$G.P) contains the p-values of the variants obtained from testing an overall marginal genetic association between the multivariate phenotype and each genetic variant individually. And the third column (PVAL$GE.P) contains the p-values obtained from testing overall GxE effect on the multivariate phenotype in presence of possible marginal effect due to the genetic variant and a marginal effect due to the environmental variable. Number of rows in PVAL is the same as the number of genetic variants, and it has the same structure as in the output of `mv_G_GE`. No default.

`Pg_thr_step1` A positive real number between 0 and 1 providing the p-value threshold to select the set of promising SNPs in step 1. These selected SNPs will be tested for GxE effect in the second step. Default is 0.005.

`FWER_step2` A positive real number between 0 and 1 specifying the family-wise error rate to be maintained in the second step while identifying the genetic variants having a genome-wide significant overall GxE effect on the multivariate phenotype. Default is 0.05.

Value

The output is a vector of SNPs identified to have a genome-wide significant overall GxE effect.

References

A Majumdar, KS Burch, S Sankararaman, B Pasaniuc, WJ Gauderman, JS Witte (2020) A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. bioRxiv, doi: https://doi.org/10.1101/2020.07.06.190256

See Also

`WHT`, `mv_G_GE`

WHT *Weighted multiple hypothesis testing procedure to combine two steps of testing gene-environment interaction in a two-step procedure.*

Description

Run `WHT` to adjust for multiple testing while combining two steps of the GxE interaction testing procedure. The procedure is applicable for a multivariate phenotype, as well as a univariate phenotype.

Usage

```
WHT(PVAL, first_bin_size = 5, FWER = 0.05)
```

Arguments

`PVAL` A data.frame with three columns. The first column (PVAL$SNP) provides the name of all SNPs or genetic variants tested. Second column (PVAL$G.P) contains the p-values of the variants obtained from testing an overall marginal genetic association between the multivariate phenotype and each genetic variant individually. And the third column (PVAL$GE.P) contains the p-values obtained from testing overall GxE effect on the multivariate phenotype in presence of possible marginal effect due to the genetic variant and a marginal effect due to the environmental variable. Number of rows in PVAL is the same as the number of genetic variants, and it has the same structure as in the output of `mv_G_GE`. No default.

`first_bin_size` A positive integer providing the number of SNPs in the top bin while ranking the SNPs or genetic variants according to their relative importance in the first step, which is evaluated with respect to the strength of overall marginal genetic association with the multivariate phenotype. Default is 5.

`FWER` A positive real number between 0 and 1 providing the overall family wise error rate to be maintained while identifying the genetic variants having a genome-wide significant overall GxE effect on the multivariate phenotype. Default is 0.05.

Value

The output produced by the function is a list consisting of:

`GEsnps` Vector of SNPs/genetic variants identified to have a genome-wide significant overall GxE effect.

`adjusted.PV` A data.frame providing the adjusted p-values with the corresponding genetic variants obtained by the weighted multiple hypothesis testing procedure.

References

A Majumdar, KS Burch, S Sankararaman, B Pasaniuc, WJ Gauderman, JS Witte (2020) A two-step approach to testing overall effect of gene-environment interaction for multiple phenotypes. bioRxiv, doi: https://doi.org/10.1101/2020.07.06.190256

See Also

`SST`, `mv_G_GE`

Index