Package ‘GLM2020’

March 30, 2020

Type Package

Title GWAS using GLM

Version 0.1.0

Author Rachael A. Kane, Marc A. Beer

Maintainer Rachael A. Kane <rachael.kane@wsu.edu>, Marc A. Beer <marc.beer@wsu.edu>

Description Toolbox for carrying out genome-wide association studies (GWAS) using a general linear model (GLM). In addition to carrying out GWAs using GLM, GLM2020 includes functions for carrying out pre-processing of cofactors prior to conducting GWAS by GLM, including the calculation of principal components (PCs) from genotypic data to assess population structure. Specifically, the package provides functionality for identifying correlations between genotypic PCs and cofactors specified by users, as well as automatically removing those PCs that exhibit correlation. A vignette using simulated data is provided to depict package functionality.

License CC0

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Suggests knitr, rmarkdown

Depends psych

VignetteBuilder knitr

R topics documented:

`GLM2020-package` *A tool for implementing genome-wide association analysis using a general linear model.*

Description

This package offers several features for efficiently carrying out genome-wide association studies (GWAS) using general linear models (GLM). Users may input phenotypic and genotypic data, as well as user-specified cofactors (e.g., location collected) and principal components calculated from genotype data. In addition to implementing GWAS using a GLM, this package offers functionality for automatically removing genotypic principal components that exhibit correlation with user-specified cofactors.

Usage

```
GLM2020-package
```

Format

An object of class `character` of length 1.

Details

Type vignette("GLM2020_tutorial") for example use.

Common questions and answers:

Function cofactor.pca.cor(U, G)

-Q: What object types are acceptable as inputs for U and G?

-A: Both U and G should be numeric matrices.

-Q: Should user-specified cofactors (U) and genotype data (G) include a column for taxa ID?

-A: Users should exclude taxa IDs in both matrices.

-Q: What if I do not have user-specified cofactors (U)?

-A: The function works without specifying U. In this case, the function will return a numeric matrix containing all principal components and individual scores. When users do not have user-specified cofactors (U), it is recommended that users simply use the native R function prcomp() in order to retain all information offered by prcomp().

Function GWASbyGLM(y, G, C, NC)

-Q: How should I chose the number of covariates to retain (NC)?

-A1: Choosing a value for NC is highly dependent on what the user's covariate (C) data contain. If the user does not have user-specified covariates (U), the covariate matrix (C) will likely contain only principal components (PCs) calculated from genotypic data. When this is the case, users should assess the proportion of variance explained by each PC. Looking across PCs, often the proportion of variance explained becomes less appreciable after the first several PCs (creating what is sometimes referred to as a "hockey stick" shaped graph). There are many approaches for selecting the number of PCs to retain, and one approach is to retain PCs until the "elbow" of the "hockey stick."

-A2: When covariates include both user-specified cofactors (U) and principal components (PCs) calculated from genotypic data, users should remove the PCs that are correlated with user-specified cofactors - the function cofactor.pca.cor() included in this package automatically removes those PCs correlated with user-specified cofactors. Users should check which PCs were removed and assess whether the remaining PCs explain substantial variation in the genotypic data. If the remaining PCs do not explain substantial variance, they likely poorly account for population structure, and the user may opt to exclude them.

-Q: How do I know which p-values correspond to which genetic marker?

-A: The p-values returned by the GWASbyGLm function are ordered in the same way as the genetic markers included in the original genotype data matrix (G). Users should refer to their genotype data matrix, and if applicable, a marker map, when interpreting p-values.

`cofactor.pca.cor` *Correlation between cofactors and principal components.*

Description

Test for correlations between user-specified cofactors and principal components calculated from genotype data. Automatically remove principal components linearly dependent (correlated) with user-specified cofactors.

Usage

```
cofactor.pca.cor(U, G)
```

Arguments

`U`	A numeric matrix containing user-specified cofactors. Dimensions are n rows (individuals) by t columns (cofactors).
`G`	A numeric matrix containing genotype data. Dimensions are n rows (individuals) by m columns (genetic markers).

Details

When U is unspecified, cofactor.pca.cor will return a list of 1 object. With U unspecified, function will carry out principal components analysis identically to the native R function prcomp(), and cofactor.pca.cor will return principal components scores in $cov. $cov is a numeric matrix containing all principal components and individual scores. Dimensions are n rows (individuals) by t columns (principal components).

When U is specified, cofactor.pca.cor will return a list of 3 objects. $orig_pc is a numeric matrix containing all original principal components and individual scores. $cov is a numeric matrix containing user-specified cofactors and all principal components not correlated with the user-specified cofactors. Dimensions are n rows (individuals) by t columns (cofactors). $removed is a character matrix indicating which principal components were removed.

The $cov matrix is intended for use as the "C" argument in the GWASbyGLM function included in this package.

Type vignette("GLM2020_tutorial") for example use.

Value

A list of 1 or 3 objects.

U unspecified: 1 object. $cov, a numeric matrix containing all principal components and individual scores.

U specified: 3 objects. $orig_pc, a numeric matrix containing all original principal components $cov, a numeric matrix containing user-specified cofactors and retained principal components. $removed, a matrix indicating which principal components were removed.

GWASbyGLM — *Genome-wide association analysis using a general linear model.*

Description

Genome-wide association analysis using a general linear model.

Usage

```
GWASbyGLM(y, G, C, NC)
```

Arguments

`y`	A numeric matrix containing phenotype data. Dimensions are n rows (individuals) by 1 column.
`G`	A numeric matrix containing genotype data. Dimensions are n rows (individuals) by m columns (genetic markers).
`C`	A numeric matrix containing covariate data. Dimensions are n rows (individuals) by t columns (covariates). The expected input for this parameter is the $cov numeric matrix returned from the cofactor.pca.cor function included in this package.
`NC`	An integer specifying the number of covariates to retain for analysis.

Details

Numeric matrices should contain only phenotype/genotype/covariate values, no accessory information like taxa ID.

Type vignette("GLM2020_tutorial") for example use.

Value

A numeric matrix containing a p-value for each genetic marker. Dimensions are 1 row by m columns (genetic markers).

Index