Package 'lmem.gwaser'

May 6, 2016

Type Package

Title Linear Mixed Effects Models for Genome-Wide Association Studies

Version 0.1.0

Author Lucia Gutierrez [aut],
Gaston Quero [aut, cre],
Schubert Fernandez [aut],
Sofia Brandariz [aut],
Sebastian Simondi [ctb]

Maintainer Gaston Quero <gastonquero@gmail.com>

Description Performs Genome-Wide Association analysis for diverse populations and for multi-environment and multi-trait analysis using linear mixed models.

License GPL-3

Depends R (>= 2.10)

LazyData TRUE

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

Imports qtl, lattice, pastecs, stringr, fdrtool, LDheatmap, genetics, graphics, utils, grDevices, stats, lme4

Encoding UTF-8

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-05-06 23:00:33

R topics documented:

`gwas.analysis` *Performs GWAS analysis with five optional models and different optional thresholds.*

Description

GWAS analysis with models: naive: y= x + e, fixed: y = x + q + e, kinship: y=x+z+e (Pariseaux and Bernardo, 2004), QK: y = x + q + z + e (Yu et al., 2006), and eigenstrat: y = x + q + e (Price et al., 2006; Malosetti et al., 2007).

Usage

```
gwas.analysis (crossobj, method, provide.K,
covariates, trait, threshold, p,out.file)
```

Arguments

`crossobj` An object of class = cross obtained from the gwas.cross function from this package, or the read.cross function from r/qtl package (Broman and Sen, 2009). This file contains phenotypic means, genotypic marker score, and genetic map.

`method` Methods to perform GWAS analysis. Options are naive, fixed, kinship, QK and egeinstrat. The general Mixed Model equation used is:

$$Y = X\beta + Q\nu + Zu + e$$

, where Y is the phenotypic vector, X is the molecular marker matrix,

$$\beta$$

is the unknown vector of allelic effects to be estimated, Q is the population structure,

$$\nu$$

is the vector of population effects (parameters), Z is a matrix that relates each measurement to the individual from which it was obtained, u is the vector of random background polygenic effects, and e is the residual errors. Random effects are underlined.

The following mainstream models are available with the package: 1) naive; a simple test of association (Kruskal-Wallis) with no correction for population structure

$$Y = X\beta + e$$

,
2) fixed; a fixed-effects model using populations structure as fixed covariate

$$Y = X\beta + Q\nu + e$$

,
3) kinship; a mixed model including the coancestry matrix among genotypes as a random effect following Parisseaux and Bernardo 2004

$$Y = X\beta + Zu + e$$

4) eigenstrat; a mixed-effects model including population structure but as a random effect following Price et al. 2006 and Malosetti et al. 2007

$$Y = X\beta + Q\nu + e$$

5) QK; a mixed-effects model including both population structure and coancestry among genotypes following Yu et al. 2006.

$$Y = X\beta + Q\nu + Zu + e$$

Principal component analysis (PCA) is used as a random effect in the Price model including all significant axes, following Patterson et al. (2006). When used in the Fixed or QK model, PCA, or another population structure is included as a fixed effect.

`provide.K` K is the kinship matrix. If pedigree kinship is available, or a specific kinship matrix is desired, set provide.k=TRUE. Otherwise, a realized kinship matrix is estimated if needed for the model. Indicates whether a qqplot shouldo be performed. TRUE/FALSE term. FALSE is set as default.

`covariates` A vector of structure covariates. Can be pca$scores for eigenstrat or any group for the fixed model. Indicates whether a scatterplot should be performed.

`trait` Indicates the trait to be analyzed.

`threshold` Thresholds options are: Li&Ji (Li and Ji, 2005), FDR (Benjamini and Hochberg, 1995), and set alpha levels (p.values)

`p` Alpha level (numeric) for test of marker-trait hypothesis.

`out.file` Name of the file to be written. Example: 'GWAS fixed Groups model'.

Details

This analysis is performed with adjusted means of the field.

Value

The function return p.values tested on the GWAS analyses saved to gwas_reports, and Manhattan plots.

Note

For multi-trait or multi-environment see GWAS.MEMQ

Author(s)

Lucia Gutierrez

References

Benjamini and Hochberg (1995) Controlling the false discovery rate: a b practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57, 289-300.

Comadran J, Thomas W, van Eeuwijk F, Ceccarelli S, Grando S, Stanca A, Pecchioni N, Akar T, Al-Yassin A, Benbelkacem A, Ouabbou H, Bort J, Romagosa I, Hackett C, Russell J (2009) Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare associatio-mapping population for the Mediterranean basin. Theor Appl Genet 119:175-187

Li J, Ji L (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity:1-7.

Yu et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Genetics 38:203-208.

Malosetti et al. (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. Genetics 175:879-889.

Parisseaux B, Bernardo R (2004) Insilico mapping of quantitative trait loci in maize. Theor. Appl. Genet. 109:08-514.

Peterson RF, Campbell AB, Han_nah AE, 1948. A diagrammatic scale for estimating rust intensity on leaves and stems of cereals. Canadian Journal of Genetics and Cytology C. 26:496-500

Price et al. (2006) Principal components analysis corrects for stratificat on in genome-wide association studies, Nat. Genet. 38:904-909

Turner, S. (2014). qqman: Q-Q and manhattan plots for GWAS data R package 0.1.2 https://CRAN.R-project.org/package=qqman

See Also

gwas.cross mq.g.diagnostics

Examples

```
## Not run:
data (QA_geno)
data (QA_map)
data (QA_pheno)

P.data <- QA_pheno
G.data <- QA_geno
map.data <- QA_map

cross.data <- gwas.cross (P.data, G.data, map.data,
cross='gwas', heterozygotes=FALSE)
summary (cross.data)


#PCA
```

```
pca <- pca.analysis (crossobj=cross.data, p.val=0.05)

#LD.plots
linkdis.plots(crossobj = cross.data, heterozygotes = FALSE, chr = c('1'))

#Mixed model: Q+K
(qk.GWAS <- gwas.analysis (crossobj=cross.data4, method="QK", provide.K=FALSE,
covariates=pca$scores, trait="yield", threshold="Li&Ji", p=0.05,
out.file="GWAS Q + K model"))$selected

#Mixed model: Eigenanalysis (PCA as random component)
(pcaR.GWAS <- gwas.analysis(crossobj=cross.data4, method="eigenstrat",
provide.K=FALSE, covariates=pca$scores, trait="yield", threshold="Li&Ji",
 p=0.05, out.file="GWAS PCA as Random model"))$selected

#Mixed model: Kinship model
 (k.GWAS <- gwas.analysis(crossobj=cross.data4, method="kinship",
 provide.K=FALSE, covariates=FALSE, trait="yield",
 threshold="Li&Ji", p=0.05, out.file =" GWAS K as Random model "))$selected

#Fixed effects: Groups
 data (QA_pheno2)
 P.data.1 <- QA_pheno2
 covariate <- P.data.1 [,2]

 (g.GWAS <- gwas.analysis (crossobj=cross.data4,
 method="fixed", provide.K=FALSE, covariates=covariate,
 trait="yield", threshold="Li&Ji", p=0.05,
 out.file="GWAS fixed Groups model"))$selected

# Naive
 (naive.GWAS <- gwas.analysis(crossobj=cross.data4, method="naive",
  provide.K=FALSE, covariates=FALSE, trait="yield", threshold="Li&Ji",
  p=0.05, out.file="GWAS naive model"))$selected

## End(Not run)
```

`gwas.cross` *Read genomic data to perform GWAS analyses.*

Description

This function reads genomic data and is similar to the read.cross function from r/qtl package (Broman and Sen, 2009) but allows importing data from a flapjack format (Milne et al., 2010). Additionally, it loads diverse populations for GWAS analysis into r/qtl format. The files required include a file containing phenotypic information (P.data), a file containing genotypic information (G.data), and a file containing map information (map.data) for all markers.

Usage

```
gwas.cross(P.data = NULL, G.data, map.data, cross = "gwas",
  heterozygotes = TRUE, sep = "\t")
```

Arguments

P.data	Name of the file containing phenotypic information. Each row represents the individuals while each column represents the phenotypic traits. The first column should be labeled as 'genotype' and should contain identification name for each individual. The name of each trait should also be included.
G.data	Name of the file containing genotypic (marker scores) information. Each row represents the individuals while each column represents the markers. Headers for markers should be included, but not for genotypes. The first column contains the names of the genotypes. The first row contains the names of the markers. The marker genotypes are coded by two characters corresponding to the alleles using a separator between alleles (by default a slash /). If a single character is given, the genotype is assumed to be homozygous. Missing values are indicated by default with '-'. In the example below, the two alleles have been called 1 and 2 because it is useful to link alleles to their origin, i.e. parent 1 or parent 2. Therefore, 1 corresponds to homozygous for allele 1 (synonymous to 1/1), 1/2 corresponds to heterozygous, and 2 corresponds to homozygous for allele 2 (synonymous to 2/2). In the case of partially informative markers (e.g. dominant markers) genotypes are coded as 1/- or 2/-, depending on whether the dominant allele originated from parent 1 or parent 2.
map.data	Name of the file containing marker map information (i.e. linkage group and position within linkage group). The file is a text tab delimited file. Each row represents markers. The file consists of three columns. Column 1 gives the marker names, column 2 the chromosome on which the marker has been mapped, and column 3 indicates the position of the marker within the chromosome.
cross	The type of population studied. gwas is set as default Diverse population panel for GWAS.
heterozygotes	It indicates whether there are heterozygotes or not in the association mapping population. TRUE is set as default.
sep	To define the espace between the data.

Details

The function creates an intermediate file called 'temp.csv' and then uses the read.cross from r/qtl to read it. The output object is an object of class=cross, the same as the one produced by the function read.cross in r/qtl (Broman and Sen, 2009)

Value

Creates an object of class cross to be used in GWAS analysis. The component are the same as r/qtl (Broman and Sen, 2009): geno

Note

All functions in this package uses cross data style.

Author(s)

Lucia Gutierrez, Gaston Quero.

References

Broman KW, Sen S (2009) A Guide to QTL Mapping with R/qtl. Springer, NewYork Comadran J, Thomas W, van Eeuwijk F, Ceccarelli S, Grando S, Stanca A, Pecchioni N, Akar T, Al-Yassin A, Benbelkacem A, Ouabbou H, Bort J, Romagosa I, Hackett C, Russell J (2009) Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association-mapping population for the Mediterranean basin. Theor Appl Genet 119:175-187 Milne et al., (2010) Flapjack - graphical genotype visualization. Bioinformatics 26(24), 3133-3134.

See Also

gwas.analysis

Examples

```
data (QA_geno)
data (QA_map)
data (QA_pheno)

P.data <- QA_pheno
G.data <- QA_geno
map.data <- QA_map

cross.data <- gwas.cross (P.data, G.data, map.data,
cross='gwas', heterozygotes=FALSE)

summary (cross.data)
```

---

linkdis.plots *Linkage Disequilibrium heatmap plot*

Description

Performs a Linkage Disequilibrium heatmap plot for the GWAS analysis. Non-random association of markers (linkage disequilibrium) are estimated as Lewont in $'/c$s D'/c (Lewontin$'/c$s, 1964) with the LD function of the genetics package (Warnes and Leisch, 2005) and isualized with the LD.heatmap package (Shin et al., 2015). D'/c is estimated as: $D'/c = D/DMax$ where D= pAB - pApB, DMax=Min (pApb, papB), and pA is the probability of the A allele for marker 1, pa=1-pA, pB is the probability of the B allele for marker 2, pb=1-pB, and pAB is the probability of AB alleles.

Usage

```
linkdis.plots (crossobj, heterozygotes, chr)
```

Arguments

crossobj	An object of class = cross obtained from the gwas.cross function from this package, or the read.cross function from r/qtl package (Broman and Sen, 2009). This file contains phenotypic means, genotypic marker score, and genetic map.
heterozygotes	Logical value indicating whether heterozygotes are present.
chr	A vector containing chromosome number to use.

Details

The function returns the LD.heatmap for the chromosomes selected.

Value

Return a Linkage Disequilibrium heatmap plot.

Note

When large data sets are being used, linkdis.plots is encourage to be performed for each chromosome separately.

Author(s)

Lucia Gutierrez

References

Comadran J, Thomas W, van Eeuwijk F, Ceccarelli S, Grando S, Stanca A, Pecchioni N, Akar T, Al-Yassin A, Benbelkacem A, Ouabbou H, Bort J, Romagosa I, Hackett C, Russell J (2009) Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association-mapping population for the Mediterranean basin. Theor Appl Genet 119:175-187

Warnes, G; Leisch, F. 2005. Genetics: Population genetics R package 1.2.0. Lewontin, R. 1964. The interaction of selection and linkage. I. General Considerations: Heterotic models. Genetics 49: 49-67.

See Also

gwas.analysis

Examples

```
## Not run:
data (QA_geno)
data (QA_map)
data (QA_pheno)
```

```
P.data <- QA_pheno
G.data <- QA_geno
map.data <- QA_map

cross.data <- gwas.cross (P.data, G.data, map.data,
cross='gwas', heterozygotes=FALSE)
summary (cross.data)

LD.plots

linkdis.plots(crossobj = cross.data, heterozygotes = FALSE, chr = c('1', '7'))

## End(Not run)
```

`mq.g.diagnostics` *Performs molecular markers quality diagnostics.*

Description

Performs molecular markers quality diagnostic of an object of class cross created by the gwas.cross function, including summary description for marker distribution and coverage, evaluating the map quality, the presence of identical individuals, visualizing marker alleles and missing marker scores for all individuals across the genome, the pairwise number of alleles shared by each pair of individuals, the pairwise recombination fraction among each pair of markers, and a test for segregation distortion for each marker in linkage analysis.

Usage

```
mq.g.diagnostics(crossobj, I.threshold = 0.1, I.quant = FALSE,
  p.val = 0.01, na.cutoff = 0.1)
```

Arguments

`crossobj`	An object of class = cross obtained from the gwas.cross function from this package, or the read.cross function from r/qtl package (Broman and Sen, 2009). This file contains phenotypic means, genotypic marker score, and genetic map.
`I.threshold`	Threshold for proportion of allelic differences below which individuals are marked as too similar, pairs that differ more than (1-threshold) are marked as exceptioally different. Default is set to 10 per cent (I.threshold = 0.1).
`I.quant`	Threshold indicating the quantile to identify the most similar individuas. Default is set to FALSE.
`p.val`	Significance level for the chi-square test for segregation distortion. This function is only used for balanced populations, and is not used in GWAS populations. The default is set to p<0.01. No multiple comparison correction is performed here.
`na.cutoff`	Proportion of missing data above which individuals and markers are reported . Default is set to 10 per cent (na.cutoff = 0.1).

Details

Performs plots in the work directory.

Value

The following reports are written to mq_reports: 1) mq_g_summary_markers, reports on missing data and segregation distortion.

2) mq_g_problems_markers, reports on duplicate or outlier genotypes.

Additionally, several diagnostic plots are performed:

1) mq_g_markermap_plot, this figure shows the position of all markers across the genome (equivalent R/qtl: plot.map) (Broman and Sen 2009).

2) mq_g_genotype_plot, this figure shows marker alleles for all individuals across thegenome (equivalent to r/qtl: geno.image) (Broman and Sen 2009).

3) mq_g_missinggenotype_plot, this figure highlights missing marker scores for all individuals across the genome (equivalent to r/qtl: plot.missing) (Broman and Sen 2009).

4) mq_g_comparegenotypes_plot, this figure represents the pairwise number of alleles shared by each pair of individuals (equivalent to r/qtl: comparegeno) (Broman and Sen 2009).

6) mq_g_cf_plot, this figure represents the pairwise recombination fraction among each pair of markers (equivalent to r/qtl: plot.rf). (Broman and Sen 2009).

7) mq_g_identical_genotypes_plot, this figure is the histogram of the proportion of shared alleles among each pair of individuals.

Note

Performs marker quality daignostics for QTL and GWAS analyses

Author(s)

Lucia Gutierrez, Gaston Quero

References

Broman KW, Sen S (2009) A Guide to QTL Mapping with R/qtl. Springer, NewYork Hayes PM, Liu BH, Knapp SJ, Chen F, Jones B, Blake T, Franckowiak JD, Rasmusson DC, Sorrells M, Ullrich SE, Wesenberg DM, Kleinhofs A (1993) Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. Theor Appl Genet 87:392-401

See Also

gwas.cross

Examples

```
data (QA_geno)
data (QA_map)
data (QA_pheno)

P.data <- QA_pheno
G.data <- QA_geno
map.data <- QA_map

cross.data  <- gwas.cross (P.data, G.data, map.data,
cross='gwas', heterozygotes=FALSE)
summary (cross.data)

#Marker Quality

mq.g.diagnostics (crossobj=cross.data,I.threshold=0.1,
            p.val=0.01,na.cutoff=0.1)
```

`pca.analysis` *Principal Component Analysis.*

Description

Performs Principal Component Analysis of marker data from an object of cross class created by the gwas.cross function.

Usage

```
pca.analysis(crossobj, p.val)
```

Arguments

`crossobj`	An object of class = cross obtained from the gwas.cross function from this package, or the read.cross function from r/qtl package (Broman and Sen, 2009). This file contains phenotypic means, genotypic marker score, and genetic map.
`p.val`	Alpha level (a number) to identify the number of significant axis

Details

Performs two plots.

Value

A PCA plot with two principal components and a scree plot for all significant axes indicating the proportion of the variance explained by each marker.

Note

In gwas.memq function, the pca.anlysis function is already included.

Author(s)

Lucia Gutierrez

References

Comadran J, Thomas W, van Eeuwijk F, Ceccarelli S, Grando S, Stanca A, Pecchioni N, Akar T, Al-Yassin A, Benbelkacem A, Ouabbou H, Bort J, Romagosa I, Hackett C, Russell J (2009) Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association-mapping population for the Mediterranean basin. Theor Appl Genet 119:175-187

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) The New S Language. Wadsworth & Brooks/Cole.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979) Multivariate Analysis, London: Academic Press.

Venables, W. N. and B. D. Ripley (2002) Modern Applied Statistics with S, Springer-Verlag.

See Also

gwas.analysis

Examples

```
## Not run:
data (QA_geno)
data (QA_map)
data (QA_pheno)

P.data <- QA_pheno
G.data <- QA_geno
map.data <- QA_map

cross.data <- gwas.cross (P.data, G.data, map.data,
cross='gwas', heterozygotes=FALSE)
summary (cross.data)

pca <- pca.analysis(crossobj=cross.data, p.val=0.05)


## End(Not run)
```

QA_geno *Name of the file containing genotypic (marker scores) information.*

Description

A research program (MABDE) was set up to investigate patterns of adaptation in barley. In this project a large set of barley genotypes (~190 genotypes) were evaluated in Europe and in the Mediterranean region. More details about this population and research can be found in Comadran et al. (2009). In this example we use the information in one of those environments, for a set of 179 genotypes. The population was genotyped by DArTs.

Usage

```
QA_geno
```

Format

A data frame 179 genotypes and 811 markers.

Source

MABDE

QA_map *Name of the file containing marker map information*

Description

A research program (MABDE) was set up to investigate patterns of adaptation in barley. In this project a large set of barley genotypes (~190 genotypes) were evaluated in Europe and in the Mediterranean region. More details about this population and research can be found in Comadran et al. (2009). In this example we use the information in one of those environments, for a set of 179 genotypes. The population was genotyped by DArTs.

Usage

```
QA_map
```

Format

A data frame 811 row (markers) and 3 column. Column 1 gives the marker names, column 2 the chromosome on which the marker has been mapped, and column 3 indicates the position of the marker within the chromosome.

Source

MABDE

QA_pheno *Name of the file containing phenotypic information*

Description

A research program (MABDE) was set up to investigate patterns of adaptation in barley. In this project a large set of barley genotypes (~190 genotypes) were evaluated in Europe and in the Mediterranean region. More details about this population and research can be found in Comadran et al. (2009). In this example we use the information in one of those environments, for a set of 179 genotypes. The population was genotyped by DArTs.

Usage

```
QA_pheno
```

Format

A data frame 179 genotypes and 2 variables.

Source

MABDE

QA_sim *Name of the file containing phenotypic information*

Description

A research program (MABDE) was set up to investigate patterns of adaptation in barley. In this project a large set of barley genotypes (~190 genotypes) were evaluated in Europe and in the Mediterranean region. More details about this population and research can be found in Comadran et al. (2009). In this example we simulate the phenotipic data in three environments, for a set of 179 genotypes. The population was genotyped by DArTs.

Usage

```
QA_sim
```

Format

A data frame 179 genotypes and 3 variables.

Source

MABDE

Index