Package ‘PedGFLMM’

March 7, 2020

Type Package

Title Gene-Based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models for Family Data

Version 1.0.0

Description Implements family-based additive generalized linear mixed models (GLMM) and generalized functional linear mixed models (GFLMM) for gene-based association testing of dichotomous traits.

URL `https://github.com/DanielEWeeks/PedGFLMM`

BugReports `https://github.com/DanielEWeeks/PedGFLMM/issues`

License GPL-2

Depends R (>= 3.5)

biocViews

Imports fda, MASS, Matrix, nlme, pedigreemm, lme4, Mega2R

Suggests knitr,
rmarkdown,
formatR

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

VignetteBuilder knitr

R topics documented:

cov *cov*

Description

Example covariate data frame, available via data(exampleData).

Usage

```
cov
```

Format

An object of class `data.frame` with 456 rows and 4 columns.

Details

A data frame containing the covariate information. The first two columns are required to be named "ped" and "person", which are used to match subjects to their data in the pedigree data frame.

See Also

`Ped`, `geno`, `exampleData`, `snpPos`

DOPedGFLMM *PedGFLMM_beta_smooth_only call back function*

Description

First, ignore call backs that have less than two polymorphic markers. Second, convert the genotypesraw() patterns of 0x10001, 0x10002 (or 0x20001), 0x20002, 0 from the genotype matrix to the numbers 0, 1, 2, 0 for each marker. (Reverse, the order if allele "1" has the minor allele frequency.) Next, prepend the pedigree and person columns of the family data to this modified genotype matrix. Finally, invoke `PedGFLMM` with the family data and genotype matrix to compute the `PedGFLMM_beta_smooth_only` statistics. Save the p-values for each statistic in the *envir$PedGFLMM_results* data frame.

Usage

```
DOPedGFLMM(markers_arg, range_arg, envir = ENV)
```

Arguments

`markers_arg` a data.frame with the following 5 observations:

locus_link is the ordinal ranking of this marker among all loci

locus_link_fill is the position of corresponding genotype data in the *unified_genotype_table*

MarkerName is the text name of the marker

chromosome is the integer chromosome number

position is the integer base pair position of marker

`range_arg` one row of a ranges_arg. The latter is a data frame of at least three integer columns. The columns indicate a range: a chromosome number, a start base pair value, and an end base pair value.

`envir` 'environment' containing SQLite database and other globals

Value

None

Note

This function computes the PedGFLMM_beta_smooth_only statistics and appends the output to the data frame, *envir$PedGFLMM_results*. It will print out the lines as they are generated if *envir$verbose* is TRUE. The data frame *envir$PedGFLMM_results* is initialized by *init_PedGFLMM*, and is appended to each time *DOPedGFLMM* is run.

See Also

init_PedGFLMM

Examples

```
db = system.file("exdata", "seqsimmGFLMM.db", package="PedGFLMM")
ENV = init_PedGFLMM(db)
ENV$verbose = TRUE
Mega2R::applyFnToRanges(DOPedGFLMM, ENV$refRanges[50:60,], ENV$refIndices)


# Not run
Mega2R::applyFnToGenes(DOPedGFLMM, genes_arg = c("CEP104"))
```

dRule *dRule*

Description

This function applies the dynamic rule to determine the number of basis functions to use

Usage

```
dRule(geno.only)
```

Arguments

`geno.only`	The input matrix of SNP genotypes, coded 0, 1, 2.

exampleData *Example data for the PedGLMM package*

Description

Example data for the PedGLMM package

Usage

```
data(exampleData)
```

Format

Four data frames: Ped, geno, cov, snpPos

Value

None

See Also

`Ped`, `geno`, `cov`, `snpPos`

Examples

```
data(exampleData)
dim(Ped)
head(Ped)
dim(geno)
head(geno[,1:10])
dim(cov)
head(cov)
dim(snpPos)
head(snpPos)
```

geno *geno*

Description

Example genotype data frame, available via data(exampleData).

Usage

```
geno
```

Format

An object of class `data.frame` with 456 rows and 313 columns.

Details

A data frame containing the genotype information. This is a matrix with genotypes for subjects (rows) at each variant position (columns). The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data.frame. The genotypes are coded as 0, 1, 2 for autosomal markers (typically a count of the number of the minor alleles).

See Also

`Ped`, `exampleData`, `cov`, `snpPos`

`init_PedGFLMM` *load Mega2 SQLite database and perform initialization for PedGFLMM usage*

Description

This populates the **R** data frames from the specified **Mega2** SQLite database.

Usage

```
init_PedGFLMM(db = NULL, verbose = FALSE, traitname = "default")
```

Arguments

`db`	specifies the path of a **Mega2** SQLite database containing study data.
`verbose`	TRUE indicates that diagnostic printouts should be enabled. This value is saved in the returned environment.
`traitname`	Name of the affection status trait to use to set the case/control values; by default, "default"

Value

"environment" containing data frames from an SQLite database and some computed values.

Note

init_PedGFLMM sets up the schaidPed and pedPer data frames that are used later in the *DOPedGFLMM* calculation. In addition, it initializes a matrix to aid in translating a genotype allele matrix to a genotype count matrix.

It also initializes the results data frame *envir$PedGFLMM_results* to zero rows.

See Also

`DOPedGFLMM`, `Mega2PedGFLMM`

Examples

```
db = system.file("exdata", "seqsimmGFLMM.db", package="PedGFLMM")
ENV = init_PedGFLMM(db, traitname = "default")
ls(ENV)
```

Mega2PedGFLMM *Execute the PedGFLMM_beta_smooth_only function on a transcript ranges*

Description

This example function illustates how to use functions from the Mega2R R package to iterate over defined gene ranges, computing the PedGFLMM_beta_smooth_only statistics for each gene that contains more than two polymorphic markers.

Execute the PedGFLMM_beta_smooth_only function on the first *gs* default gene transcript ranges (gs = 1:100). Update the *envir$PedGFLMM_results* data frame with the results.

Usage

```
Mega2PedGFLMM(gs = 1:100, genes = NULL, envir = ENV)
```

Arguments

gs	a subrange of the default transcript ranges over which to calculate the *DOPedGFLMM* function.
genes	a list of genes over which to calculate the *DOPedGFLMM* function. The value, "*", means use all the transcripts in the selected Bioconductor database. If genes is NULL, the gs range of the internal *refRanges* will be used.
envir	'environment' containing SQLite database and other globals

Value

None the data frame with the PedGFLMM_beta_smooth_only results is stored in the environment and named *PedGFLMM_results*, viz. envir$PedGFLMM_results

See Also

init_PedGFLMM

Examples

```
db = system.file("exdata", "seqsimmGFLMM.db", package="PedGFLMM")
ENV = init_PedGFLMM(db)
ENV$verbose = TRUE
Mega2PedGFLMM(gs = 50:60)
```

M_GAO *M_GAO*

Description

Compute the effective number of independent SNPs in a region.

Usage

```
M_GAO(SNP_mx)
```

Arguments

`SNP_mx`	A matrix of polymorphic SNPs (coded 0, 1, 2) with SNPs in rows and individuals in columns.

Source

http://simplem.sourceforge.net/

References

Gao X, Starmer J and Martin ER (2008) A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. Genetic Epidemiology 32:361-369

Ped *Ped*

Description

Example pedigree data frame, available via data(exampleData)

Usage

```
Ped
```

Format

An object of class `data.frame` with 456 rows and 7 columns.

Details

A data frame containing the pedigree information with the following columns:

ID Person ID

ped pedigree ID,character or numeric allowed.

person person ID, a unique ID within each pedigree, numeric or character allowed.

father father ID, NA if no father.

mother mother ID, NA if no mother.

sex sex, coded as 1 for male, 2 for female.

trait trait phenotype, either case-control status coded as 1 for affected and 0 for unaffected. Subjects with missing (NA) will be removed from the analysis.

See Also

exampleData, geno, cov, snpPos

PedGFLMM *PedGFLMM package*

Description

This package implements family-based additive generalized linear mixed models (GLMM) and generalized functional linear mixed models (GFLMM) for gene-based association testing of dichotomous traits (Jiang et al, 2020).

Author(s)

Yingda Jiang, Chi-Yang Chiu, Daniel E. Weeks, Ruzong Fan

PedGFLMM_beta_smooth_only *PedGFLMM_beta_smooth_only*

Description

Computes the PedGFLMM statistics under the beta smooth only model.

Usage

```
PedGFLMM_beta_smooth_only(
  ped,
  geno,
  covariate = NULL,
  pos,
  order,
  beta_basis = NULL,
  base = "bspline",
  optimizer = "bobyqa",
  Wald = FALSE
)
```

Arguments

`ped` A data frame containing the pedigree information with the following columns:

- **ID** Person ID
- **ped** pedigree ID,character or numeric allowed.
- **person** person ID, a unique ID within each pedigree, numeric or character allowed.
- **father** father ID, NA if no father.
- **mother** mother ID, NA if no mother.
- **sex** sex, coded as 1 for male, 2 for female.
- **trait** trait phenotype, case-control status coded as 1 for affected and 0 for unaffected. Subjects with missing (NA) will be removed from the analysis.

`geno` A data frame containing the genotype information. This is a matrix with genotypes for subjects (rows) at each variant position (columns). The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data.frame. The genotypes are coded as 0, 1, 2 for autosomal markers (typically a count of the number of the minor alleles).

`covariate` A data frame containing the covariate information. The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data frame. This is optional and the default "covariate = NULL" is for the case when the covariate matrix is not provided.

`pos` Position of the markers in base pairs.

`order` The order used to generate the B-spline basis.

`beta_basis` The number of basis functions used to estimate the genetic effect function.

`base` Can be either ’bspline’ or ’fspline’.

`optimizer` Optimizer to use (default = "bobyqa").

`Wald` If Wald is set to true, return the Wald p-value in addition to the LRT p-value (Default: Wald = FALSE).

Value

A list containing the following components:

LRT The p-value based on a likelihood ratio test

Wald The p-value based on a Wald test, returned if ’Wald’ is TRUE

nbetabasis The number of basis functions used to estimate the genetic effect function

M_gao The effective number of variants in the region, as computed by M_GAO function

References

Chiu CY, Yuan F, Zhang BS, Yuan A, Li X, Fang HB, Lange K, Weeks DE, Wilson AF, Bailey-Wilson JE, Lakhal-Chaieb ML, Cook RJ, McMahon FJ, Amos CI, Xiong MM, and Fan RZ (2019) Pedigree-based linear mixed models for association analysis of quantitative traits with next-generation sequencing data. Genetic Epidemiology 43(2):189-206.

Fan RZ, Wang YF, Mills JL, Wilson AF, Bailey-Wilson JE, and Xiong MM (2013) Functional linear models for association analysis of quantitative traits. Genetic Epidemiology 37 (7):726- 742.

Fan RZ, Wang YF, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, and Xiong MM (2014) Generalized functional linear models for case-control association studies. Genetic Epidemiology 38 (7):622-637.

Jiang YD, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, Lakhal-Chaieb ML, Cook RJ, Amos CI, Wilson AF, Bailey-Wilson JE, McMahon FJ, Vazquez AI, Yuan A, Zhong XG, Xiong MM, Weeks DE, and Fan RZ (2020) Gene-based association testing of dichotomous traits with generalized linear mixed models for family data.

Schaid DJ, McDonnell SK, Sinnwell JP, and Thibodeau SN (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genetic Epidemiology 37:409-418.

See Also

PedGLMM_additive_effect_model, PedGFLMM_fixed_model, exampleData

Examples

```
data(exampleData)

order  =  4

bsmooth_bsp=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno),
   pos = snpPos$pos, order = order, covariate = as.matrix(cov),
   base = "bspline")
bsmooth_bsp

bsmooth_fsp=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno),
   pos = snpPos$pos, order = order, covariate = as.matrix(cov),
   base = "fspline")
bsmooth_fsp

bsmooth_bsp_no_cov=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno),
   pos = snpPos$pos, order = order, covariate = NULL,
   base = "bspline")
bsmooth_bsp_no_cov

bsmooth_fsp_no_cov=PedGFLMM_beta_smooth_only(ped = Ped, geno = as.matrix(geno),
   pos = snpPos$pos, order = order, covariate = NULL,
   base = "fspline")
bsmooth_fsp_no_cov
```

PedGFLMM_fixed_model *PedGFLMM_fixed_model*

Description

Computes the PedGFLMM statistics under a fixed model.

Usage

```
PedGFLMM_fixed_model(
  ped,
  geno,
  covariate = NULL,
  pos,
  order,
  beta_basis = NULL,
  geno_basis = NULL,
  base = "bspline",
  optimizer = "bobyqa",
  Wald = FALSE
)
```

Arguments

`ped` A data frame containing the pedigree information with the following columns:

- **ID** Person ID
- **ped** pedigree ID,character or numeric allowed.
- **person** person ID, a unique ID within each pedigree, numeric or character allowed.
- **father** father ID, NA if no father.
- **mother** mother ID, NA if no mother.
- **sex** sex, coded as 1 for male, 2 for female.
- **trait** trait phenotype, case-control status coded as 1 for affected and 0 for unaffected. Subjects with missing (NA) will be removed from the analysis.

`geno` A data frame containing the genotype information. This is a matrix with genotypes for subjects (rows) at each variant position (columns). The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data.frame. The genotypes are coded as 0, 1, 2 for autosomal markers (typically a count of the number of the minor alleles).

`covariate` A data frame containing the covariate information. The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data frame. This is optional and the default "covariate = NULL" is for the case when the covariate matrix is not provided.

`pos` Position of the markers in base pairs.

`order` The order used to generate the B-spline basis.

`beta_basis` The number of basis functions used to estimate the genetic effect function.

`geno_basis` The number of basis functions used to estimate the genetic variant functions.

`base` Can be either ’bspline’ or ’fspline’.

`optimizer` Optimizer to use (default = "bobyqa").

`Wald` If Wald is set to true, return the Wald p-value in addition to the LRT p-value (Default: Wald = FALSE).

Value

A list containing the following components:

LRT The p-value based on a likelihood ratio test

Wald The p-value based on a Wald test, returned if ’Wald’ is TRUE

nbetabasis The number of basis functions used to estimate the genetic effect function

ngenobasis The number of basis functions used to estimate the genetic variant functions

M_gao The effective number of variants in the region, as computed by M_GAO function

References

Chiu CY, Yuan F, Zhang BS, Yuan A, Li X, Fang HB, Lange K, Weeks DE, Wilson AF, Bailey-Wilson JE, Lakhal-Chaieb ML, Cook RJ, McMahon FJ, Amos CI, Xiong MM, and Fan RZ (2019) Pedigree-based linear mixed models for association analysis of quantitative traits with next-generation sequencing data. Genetic Epidemiology 43(2):189-206.

Fan RZ, Wang YF, Mills JL, Wilson AF, Bailey-Wilson JE, and Xiong MM (2013) Functional linear models for association analysis of quantitative traits. Genetic Epidemiology 37 (7):726- 742.

Fan RZ, Wang YF, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, and Xiong MM (2014) Generalized functional linear models for case-control association studies. Genetic Epidemiology 38 (7):622-637.

Jiang YD, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, Lakhal-Chaieb ML, Cook RJ, Amos CI, Wilson AF, Bailey-Wilson JE, McMahon FJ, Vazquez AI, Yuan A, Zhong XG, Xiong MM, Weeks DE, and Fan RZ (2020) Gene-based association testing of dichotomous traits with generalized linear mixed models for family data.

Schaid DJ, McDonnell SK, Sinnwell JP, and Thibodeau SN (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genetic Epidemiology 37:409-418.

See Also

PedGFLMM_beta_smooth_only, PedGLMM_additive_effect_model, exampleData

Examples

```
data(exampleData)

# betabasis_Bsp = 10
# genobasis_Bsp = 10

# betabasis_Fsp = 11
# genobasis_Fsp = 11
order  =   4

fixed_bsp=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), pos = snpPos$pos,
    order = order, covariate = as.matrix(cov), base = "bspline")
fixed_bsp

fixed_fsp=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), pos = snpPos$pos,
    order = order, covariate = as.matrix(cov), base = "fspline")
fixed_fsp

fixed_bsp_no_cov=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), pos = snpPos$pos,
    order = order, covariate = NULL, base = "bspline")
fixed_bsp_no_cov

fixed_fsp_no_cov=PedGFLMM_fixed_model(ped = Ped, geno = as.matrix(geno), pos = snpPos$pos,
    order = order, covariate = NULL, base = "fspline")
fixed_fsp_no_cov
```

PedGLMM_additive_effect_model *PedGLMM_additive_effect_model*

Description

Computes the PedGFLMM statistics under an additive effect model

Usage

```
PedGLMM_additive_effect_model(
  ped,
  geno,
  covariate = NULL,
  optimizer = "bobyqa",
  Wald = FALSE
)
```

Arguments

`ped` A data frame containing the pedigree information with the following columns:

- **ID** Person ID
- **ped** pedigree ID,character or numeric allowed.
- **person** person ID, a unique ID within each pedigree, numeric or character allowed.
- **father** father ID, NA if no father.
- **mother** mother ID, NA if no mother.
- **sex** sex, coded as 1 for male, 2 for female.
- **trait** trait phenotype, case-control status coded as 1 for affected and 0 for unaffected. Subjects with missing (NA) will be removed from the analysis.

`geno` A data frame containing the genotype information. This is a matrix with genotypes for subjects (rows) at each variant position (columns). The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data.frame. The genotypes are coded as 0, 1, 2 for autosomal markers (typically a count of the number of the minor alleles).

`covariate` A data frame containing the covariate information. The first two columns are required to be named “ped” and “person”, which are used to match subjects to their data in the pedigree data frame. This is optional and the default "covariate = NULL" is for the case when the covariate matrix is not provided.

`optimizer` Optimizer to use (default = "bobyqa").

`Wald` If Wald is set to true, return the Wald p-value in addition to the LRT p-value (Default: Wald = FALSE).

Value

A list containing the following components:

LRT The p-value based on a likelihood ratio test

Wald The p-value based on a Wald test, returned if ’Wald’ is TRUE

nbetabasis The number of basis functions used to estimate the genetic effect function

ngenobasis The number of basis functions used to estimate the genetic variant functions

M_gao The effective number of variants in the region, as computed by M_GAO function

References

Chiu CY, Yuan F, Zhang BS, Yuan A, Li X, Fang HB, Lange K, Weeks DE, Wilson AF, Bailey-Wilson JE, Lakhal-Chaieb ML, Cook RJ, McMahon FJ, Amos CI, Xiong MM, and Fan RZ (2019) Pedigree-based linear mixed models for association analysis of quantitative traits with next-generation sequencing data. Genetic Epidemiology 43(2):189-206.

Fan RZ, Wang YF, Mills JL, Wilson AF, Bailey-Wilson JE, and Xiong MM (2013) Functional linear models for association analysis of quantitative traits. Genetic Epidemiology 37 (7):726- 742.

Fan RZ, Wang YF, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, and Xiong MM (2014) Generalized functional linear models for case-control association studies. Genetic Epidemiology 38 (7):622-637.

Jiang YD, Chiu CY, Yan Q, Chen W, Gorin MB, Conley YP, Lakhal-Chaieb ML, Cook RJ, Amos CI, Wilson AF, Bailey-Wilson JE, McMahon FJ, Vazquez AI, Yuan A, Zhong XG, Xiong MM, Weeks DE, and Fan RZ (2020) Gene-based association testing of dichotomous traits with generalized linear mixed models for family data.

Schaid DJ, McDonnell SK, Sinnwell JP, and Thibodeau SN (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genetic Epidemiology 37:409-418.

See Also

PedGFLMM_beta_smooth_only, PedGFLMM_fixed_model, exampleData

Examples

```
data(exampleData)

add=PedGLMM_additive_effect_model(ped=Ped, geno = as.matrix(geno),
    covariate = as.matrix(cov))
add

add_no_cov=PedGLMM_additive_effect_model(ped=Ped, geno = as.matrix(geno), covariate = NULL)
add_no_cov
```

---

snpPos *snpPos*

Description

Example marker position data frame, available via data(exampleData).

Usage

```
snpPos
```

Format

An object of class `data.frame` with 311 rows and 3 columns.

Details

This data frame provides marker positions for each SNP. The first column, chr, contains the chromosome number, the second column, snp, contains the SNP name, and the third column, pos, contains the position of the SNP in base pairs.

See Also

`Ped`, `geno`, `cov`, `exampleData`

Index