Package 'FREGAT'

April 20, 2017

Title Family REGional Association Tests

Version 1.0.3

Author Nadezhda M. Belonogova <belon@bionet.nsc.ru> and
Gulnara R. Svishcheva <gulsvi@mail.ru>,
with contributions from:
Seunggeun Lee (kernel functions), Pierre Lafaye de Micheaux ('davies' method),
Thomas Lumley ('kuonen' method), James O. Ramsay (functional data analysis functions),
David Clayton ('read.plink' function) and Brian Ripley ('ginv' function)

Maintainer Nadezhda M. Belonogova <belon@bionet.nsc.ru>

Depends R (>= 3.0.0)

Imports methods, Matrix, splines, parallel

Suggests doParallel, foreach, GenABEL, seqminer

Repository CRAN

Description
Fast regional association analysis of quantitative traits for family-based and population studies.

License GPL-3

LazyLoad yes

R topics documented:

FREGAT-package *FREGAT: Family REGional Association Tests*

Description

The FREGAT package supplies the most common and efficient tests for the region-based association analysis aimed at identification of rare genetic variants for family-based, genetically related or population samples. Tests implemented are FFBSKAT (fast family-based sequence kernel association test), FFBSKAT optimal, famBT (family burden test), famFLM (family functional linear model, a functional data analysis approach), and MLR (standard multiple linear regression). The methods provide regional association testing of a set of SNPs with a continuous phenotype in the presence of additional covariates and within-family correlations.

Details

Package: FREGAT
Type: Package
License: GPLv3

Author(s)

Nadezhda Belonogova <belon@bionet.nsc.ru>
Gulnara Svishcheva <gulsvi@mail.ru>

The authors are very grateful to Dr. Anatoly Kirichenko for technical support
and Prof. Tatiana Axenovich for scientific guidance.

References

Belonogova N.M., Svishcheva G.R., Axenovich T.I. (2016) FREGAT: an R package for region-based association analysis. Bioinformatics, V32(15), P. 2392-2393.

example.data *A small example data set*

Description

genodata A matrix containing genotypes of 50 genetic variants (given in columns) in 66 individuals (given in rows). Three genotypes are coded as 0, 1 and 2.

phenodata A data frame containing "trait", "sex" and "age" columns: a quantitative trait to be analyzed and its covariates.

snpdata A data frame with descriptive information on 50 genetic variants in genodata. The important column is "gene": it assigns each variant to a certain gene region.

kin A kinship matrix for the 66 individuals.

Usage

```
data(example.data)
```

famBT *Family Burden Test*

Description

Burden test in related or population samples

Usage

famBT(formula, phenodata, genodata, kin = NULL, nullmod,
regions = NULL, sliding.window = c(20, 10), mode = "add",
ncores = 1, return.time = FALSE, beta.par = c(1, 25),
weights = NULL, flip.genotypes = FALSE, impute.method = 'mean',
write.file = FALSE, ...)

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait and/or covariates will be omitted.
genodata	an object with genotypes to analyze. Several formats are allowed: - a data frame or matrix (with individuals in the rows and genetic variants in the columns) containing genotypes coded as AA = 0, Aa = 1 and aa = 2, where a is a minor allele. - for PLINK binary data format, a character string indicating a *.bed file name (*.bim and *.fam files should have the same prefix). This will make use of read.plink() function. - for VCF format, a character string indicating a *vcf.gz file name. This will require seqminer R-package to be installed. Its readVCFToMatrixByGene() function will be used to read VCF file gene-wise. The function also requires a geneFile, a text file listing all genes in refFlat format (see Examples below). VCF file should be bgzipped and indexed by Tabix. - an object of gwaa.data or snp.data class (this will require GenABEL R-package to be installed).
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
nullmod	an object containing parameter estimates under the null model. Setting nullmod allows to avoid re-estimation of the null model that does not depend on genotypes and can be calculated once for a trait. If not set, the null model parameters will be estimated within the function. The nullmod object in proper format can be obtained by null.model() function or any analysis function in FREGAT.
regions	an object assigning regions to be analyzed. This can be: - a vector of length equal to the number of genetic variants assigning the region for each variant (see Examples). - a data frame / matrix with names of genetic variants in the first column and

	names of regions in the second column (this format allows overlapping regions). - for VCF format, a character vector with names of genes to analyze. If NULL, sliding.window parameters will be used.
sliding.window	the sliding window size and step. Has no effect if regions is defined.
mode	the mode of inheritance: "add", "dom" or "rec" for additive, dominant or recessive mode, respectively. For dominant (recessive) mode genotypes will be recoded as AA = 0, Aa = 1 and aa = 1 (AA = 0, Aa = 0 and aa = 1), where a is a minor allele. Default mode is additive.
ncores	number of CPUs for parallel calculations. Default = 1.
return.time	a logical value indicating whether the running time should be returned.
beta.par	two positive numeric shape parameters in the beta distribution to assign weights for each SNP. Default = c(1, 25) is recommended for analysis of rare variants. For unweighted burden test, use beta.par = c(1, 1). Has no effect if weights are defined.
weights	a numeric vector or a function of minor allele frequency (MAF) to assign weights for each SNP. If NULL, the weights will be calculated using beta distribution (see Details).
flip.genotypes	a logical value indicating whether the genotypes of some genetic variants should be flipped (relabeled) to ensure that all MAFs < 0.5. Default = FALSE, with warning of any MAF > 0.5.
impute.method	a method for imputation of missing genotypes. It can be either "mean" (default) or "blue". If "mean" the genotypes will be imputed by the simple mean values. If "blue" the best linear unbiased estimates (BLUEs) of mean genotypes will be calculated taking into account the relationships between individuals [McPeek, et al., 2004] and used for imputation.
write.file	output file name to write results as they come (sequential mode only).
...	other arguments that could be passed to null(), read.plink() and readVCFToMatrixByGene().

Details

Burden test (collapsing technique) suggests that the effects of causal genetic variants within a region have the same direction. If this is not the case, other regional tests (FFBSKAT, FLM) are shown to have higher power compared to famBT [Svishcheva, et al., 2015].
By default, famBT assigns weights calculated using the beta distribution. Given the shape parameters of the beta function, beta.par = c(a, b), the weights are defined using probability density function of the beta distribution:

$$W_i = (B(a,b))^{-1} MAF_i^{a-1}(1-MAF_i)^{b-1},$$

where MAF_i is a minor allelic frequency for the i^{th} genetic variant in region, which is estimated from genotypes, and $B(a,b)$ is the beta function.
beta.par = c(1, 1) corresponds to the unweighted burden test.

Value

A list with values:

results	a data frame containing P values, estimates of betas and their s.e., numbers of variants and polymorphic variants for each of analyzed regions.

nullmod	an object containing the estimates of the null model parameters: heritability (h2), total variance (total.var), estimates of fixed effects of covariates (alpha), the gradient (df), and the total log-likelihood (logLH).
sample.size	the sample size after omitting NAs.
time	If return.time = TRUE a list with running times for null model, regional analysis and total analysis is returned. See proc.time() for output format.

References

Svishcheva G.R., Belonogova N.M. and Axenovich T.I. (2015) Region-based association test for familial data under functional linear models. PLoS ONE 10(6): e0128999.
McPeek M.S., Wu X. and Ober C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics (60): 359-367.

Examples

```
data(example.data)

## Run famBT with sliding window (default):
out <- famBT(trait ~ age + sex, phenodata, genodata, kin)

## Run famBT with regions defined in snpdata$gene and with
## null model parameters obtained in the first run:
out <- famBT(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, regions = snpdata$gene)

## Run famBT parallelized on two cores (this will require
## 'foreach' and 'doParallel' R-packages installed and
## cores available):
out <- famBT(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, ncores = 2)

## Run famBT with genotypes in VCF format:
VCFfileName <- system.file(
"testfiles/1000g.phase1.20110521.CFH.var.anno.vcf.gz",
package = "FREGAT")
geneFile <- system.file("testfiles/refFlat_hg19_6col.txt.gz",
package = "FREGAT")
phe <- data.frame(trait = rnorm(85))
out <- famBT(trait, phe, VCFfileName, geneFile = geneFile,
reg = "CFH", annoType = "Nonsynonymous",
flip.genotypes = TRUE)

## Run famBT with genotypes in PLINK binary data format:
bedFile <- system.file("testfiles/sample.bed",
package = "FREGAT")
phe <- data.frame(trait = rnorm(120))
out <- famBT(trait, phe, bedFile)
```

famFLM *family Functional Linear Model*

Description

A region-based association test for familial or population data under functional linear models (functional data analysis approach)

Usage

famFLM(formula, phenodata, genodata, kin = NULL, nullmod,
regions = NULL, sliding.window = c(20, 10), mode = "add",
ncores = 1, return.time = FALSE, beta.par = c(1, 1),
weights = NULL, positions = NULL, GVF = FALSE,
BSF = "fourier", kg = 30, kb = 25, order = 4, stat = "F",
flip.genotypes = FALSE, impute.method = 'mean',
write.file = FALSE, ...)

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait or covariates will be omitted.
genodata	an object with genotypes to analyze. Several formats are allowed: - a data frame or matrix (with individuals in the rows and genetic variants in the columns) containing genotypes coded as AA = 0, Aa = 1 and aa = 2, where a is a minor allele. - for PLINK binary data format, a character string indicating a *.bed file name (*.bim and *.fam files should have the same prefix). This will make use of read.plink() function. - for VCF format, a character string indicating a *vcf.gz file name. This will require seqminer R-package to be installed. Its readVCFToMatrixByGene() function will be used to read VCF file gene-wise. The function also requires a geneFile, a text file listing all genes in refFlat format (see Examples below). VCF file should be bgzipped and indexed by Tabix. - an object of gwaa.data or snp.data class (this will require GenABEL R-package to be installed).
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
nullmod	an object containing parameter estimates under the null model. Setting nullmod allows to avoid re-estimation of the null model that does not depend on genotypes and can be calculated once for a trait. If not set, the null model parameters will be estimated within the function. The nullmod object in proper format can be obtained by null.model() function or any analysis function in FREGAT.
regions	an object assigning regions to be analyzed. This can be: - a vector of length equal to the number of genetic variants assigning the region

	for each variant (see Examples). - a data frame / matrix with names of genetic variants in the first column and names of regions in the second column (this format allows overlapping regions). - for VCF format, a character vector with names of genes to analyze. If NULL, sliding.window parameters will be used.
sliding.window	the sliding window size and step. Has no effect if regions is defined.
mode	the mode of inheritance: "add", "dom" or "rec" for additive, dominant or recessive mode, respectively. For dominant (recessive) mode genotypes will be recoded as AA = 0, Aa = 1 and aa = 1 (AA = 0, Aa = 0 and aa = 1), where a is a minor allele. Default mode is additive.
ncores	number of CPUs for parallel calculations. Default = 1.
return.time	a logical value indicating whether the running time should be returned.
beta.par	two positive numeric shape parameters in the beta distribution to assign weights for each genetic variant as a function of MAF (see Details). Default = c(1, 1) corresponds to standard unweighted FLM. Has no effect if weights are defined.
weights	a numeric vector or a function of minor allele frequency (MAF) to assign weights for each genetic variant in the weighted kernels. Has no effect if one of unweighted kernels was chosen. If NULL, the weights will be calculated using the beta distribution (see Details).
positions	a vector of physical positions for genetic variants in genodata. Not used when VCF file supplied.
GVF	a basis function type for Genetic Variant Functions. Can be set to "bspline" (B-spline basis) or "fourier" (Fourier basis). The default GVF = FALSE assumes beta-smooth only. If GVF = TRUE the B-spline basis will be used.
BSF	a basis function type for beta-smooth. Can be set to "bspline" (B-spline basis) or "fourier" (Fourier basis, default).
kg	the number of basis functions to be used for GVF (default = 30, has no effect under GVF = FALSE).
kb	the number of basis functions to be used for BSF (default = 25).
order	a polynomial order to be used in "bspline". Default = 4 corresponds to the cubic B-splines. as no effect if only Fourier bases are used.
stat	the statistic to be used to calculate the P values. One of "F" (default), "Chisq", "LRT".
flip.genotypes	a logical value indicating whether the genotypes of some genetic variants should be flipped (relabeled) for their better functional representation [Vsevolozhskaya, et al., 2014]. Default = FALSE.
impute.method	a method for imputation of missing genotypes. It can be either "mean" (default) or "blue". If "mean" the genotypes will be imputed by the simple mean values. If "blue" the best linear unbiased estimates (BLUEs) of mean genotypes will be calculated taking into account the relationships between individuals [McPeek, et al., 2004, DOI: 10.1111/j.0006-341X.2004.00180.x] and used for imputation.
write.file	output file name to write results as they come (sequential mode only).
...	other arguments that could be passed to null(), read.plink() and readVCFToMatrixByGene().

Details

The test assumes that the effects of multiple genetic variants (and also their genotypes if GVFs are used) can be described as a continuous function, which can be modelled through B-spline or Fourier basis functions. When the number of basis functions (set by Kg and Kb) is less than the number of variants within the region, the famFLM test may have an advantage of using less degrees of freedom [Svishcheva, et al., 2015].

Several restrictions exist in combining B-spline or Fourier bases for construction of GVFs and BSF [Svishcheva, et al., 2015], and the famFLM function takes them into account. Namely:

1) $m \geq Kg \geq Kb$, where m is the number of polymorphic genetic variants within a region.

2) Under $Kg = Kb$, B-B and B-F models are equivalent to 0-B model, and F-F and F-B models are equivalent to 0-F model. 0-B and 0-F models will be used for these cases, respectively.

3) Under $m = Kb$, 0-B and 0-F models are equivalent to a standard multiple linear regression, and it will be used for these cases.

4) When Fourier basis is used, the number of basis functions should be an odd integer. Even values will be changed accordingly.

Because of these restrictions, the model in effect may not always be the same as it has been set. The ultimate model name is returned in results in the "model" column (see below).

$\mathrm{beta.par} = \mathrm{c(a,\ b)}$ can be used to set weights for genetic variants. Given the shape parameters of the beta function, $\mathrm{beta.par} = \mathrm{c(a,\ b)}$, the weights are defined using probability density function of the beta distribution:

$$W_i = (B(a,b))^{-1} MAF_i^{a-1}(1-MAF_i)^{b-1},$$

where MAF_i is a minor allelic frequency for the i^{th} genetic variant in the region, which is estimated from genotypes, and $B(a,b)$ is the beta function. This way of defining weights is the same as in original SKAT (see [Wu, et al., 2011] for details).

Value

A list with values:

results	a data frame containing P values, numbers of variants and informative polymorphic variants for each of analyzed regions. It also contains the names of the functional models used for each region (it may not always coincide with what was set, because of restrictions described in Details section). The first part of the name relates to the functional basis of GVFs and the second one to that of BSF, e.g. "F30-B25" means that 30 Fourier basis functions were used for construction of GVFs and 25 B-spline basis functions were used for construction of BSF. "0-F25" means that genotypes were not smoothed and 25 Fourier basis functions were used for beta-smooth. "MLR" means that standard multiple linear regression was applied.
nullmod	an object containing the estimates of the null model parameters: heritability (h2), total variance (total.var), estimates of fixed effects of covariates (alpha), the gradient (df), and the total log-likelihood (logLH).

sample.size	the sample size after omitting NAs.
time	If return.time = TRUE a list with running times for null model, regional analysis and total analysis is returned. See proc.time() for output format.

References

Svishcheva G.R., Belonogova N.M. and Axenovich T.I. (2015) Region-based association test for familial data under functional linear models. PLoS ONE 10(6): e0128999.
Vsevolozhskaya O.A., et al. (2014) Functional Analysis of Variance for Association Studies. PLoS ONE 9(9): e105074.
Wu M.C., et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., Vol. 89, P. 82-93.

Examples

```
data(example.data)

## Run famFLM with sliding window (default):
out <- famFLM(trait ~ age + sex, phenodata, genodata, kin,
positions = snpdata$position)

## Run famFLM with regions defined in snpdata$gene and with
## null model parameters obtained in the first run:
out <- famFLM(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, positions = snpdata$position,
regions = snpdata$gene)

## Run famFLM parallelized on two cores (this will require
## 'foreach' and 'doParallel' R-packages installed and
## cores available):
out <- famFLM(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, positions = snpdata$position, ncores = 2)

## Run MLR with genotypes in VCF format:
VCFfileName <- system.file(
"testfiles/1000g.phase1.20110521.CFH.var.anno.vcf.gz",
package = "FREGAT")
geneFile <- system.file("testfiles/refFlat_hg19_6col.txt.gz",
package = "FREGAT")
phe <- data.frame(trait = rnorm(85))
out <- famFLM(trait, phe, VCFfileName, geneFile = geneFile,
reg = "CFH", annoType = "Nonsynonymous",
flip.genotypes = TRUE)

## Run famFLM with genotypes in PLINK binary data format:
bedFile <- system.file("testfiles/sample.bed",
package = "FREGAT")
data <- read.plink(bedFile)
phe <- data.frame(trait = rnorm(120))
out <- famFLM(trait, phe, bedFile, positions = data$map$position)
```

FFBSKAT *Fast Family-Based SKAT*

Description

A fast regional association analysis in related or population samples

Usage

```
FFBSKAT(formula, phenodata, genodata, kin = NULL, nullmod,
regions = NULL, sliding.window = c(20, 10), mode = "add",
ncores = 1, return.time = FALSE, kernel = "linear.weighted",
beta.par = c(1, 25), weights = NULL, method = "kuonen",
acc = 1e-8, lim = 1e+6, return.variance.explained = FALSE,
reml = TRUE, flip.genotypes = FALSE, impute.method = 'mean',
rho = FALSE, write.file = FALSE, ...)
```

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait or covariates will be omitted.
genodata	an object with genotypes to analyze. Several formats are allowed: - a data frame or matrix (with individuals in the rows and genetic variants in the columns) containing genotypes coded as AA = 0, Aa = 1 and aa = 2, where a is a minor allele. - for PLINK binary data format, a character string indicating a *.bed file name (*.bim and *.fam files should have the same prefix). This will make use of read.plink() function. - for VCF format, a character string indicating a *vcf.gz file name. This will require seqminer R-package to be installed. Its readVCFToMatrixByGene() function will be used to read VCF file gene-wise. The function also requires a geneFile, a text file listing all genes in refFlat format (see Examples below). VCF file should be bgzipped and indexed by Tabix. - an object of gwaa.data or snp.data class (this will require GenABEL R-package to be installed).
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
nullmod	an object containing parameter estimates under the null model. Setting nullmod allows to avoid re-estimation of the null model that does not depend on genotypes and can be calculated once for a trait. If not set, the null model parameters will be estimated within the function. The nullmod object in proper format can be obtained by null.model() function or any analysis function in FREGAT.
regions	an object assigning regions to be analyzed. This can be: - a vector of length equal to the number of genetic variants assigning the region for each variant (see Examples).

	- a data frame / matrix with names of genetic variants in the first column and names of regions in the second column (this format allows overlapping regions). - for VCF format, a character vector with names of genes to analyze. If NULL, sliding.window parameters will be used.
sliding.window	the sliding window size and step. Has no effect if regions is defined.
mode	the mode of inheritance: "add", "dom" or "rec" for additive, dominant or recessive mode, respectively. For dominant (recessive) mode genotypes will be recoded as AA = 0, Aa = 1 and aa = 1 (AA = 0, Aa = 0 and aa = 1), where a is a minor allele. Default mode is additive.
ncores	number of CPUs for parallel calculations. Default = 1.
return.time	a logical value indicating whether the running time should be returned.
kernel	one of "linear.weighted" (default), "quadratic", "IBS", "IBS.weighted", "2wayIX". See Details for "linear.weighted" kernel description and [Wu, et al., 2011] for other kernel types. "2wayIX" kernel considers SNP-SNP interaction terms along with main effects. For "linear.weighted" and "IBS.weighted" kernels, weights can be varied by defining weights or beta.par.
beta.par	two positive numeric shape parameters in the beta distribution to assign weights for each SNP in weighted kernels (see Details). Default = c(1, 25) is recommended for analysis of rare variants. Has no effect for unweighted kernels or if weights are defined.
weights	a numeric vector or a function of minor allele frequency (MAF) to assign weights for each genetic variant in the weighted kernels. Has no effect if one of unweighted kernels was chosen. If NULL, the weights will be calculated using the beta distribution (see Details).
method	either "kuonen" or "davies". Method for computing the P value (see Details). Default = "kuonen".
acc	accuracy parameter for "davies" method.
lim	limit parameter for "davies" method.
return.variance.explained	a logical value indicating whether the (marginal) variance explained by each region should be returned. Default = FALSE for faster performance.
reml	a logical value indicating whether the restricted maximum likelihood should be used to estimate the variance explained by each region. Default = TRUE for faster performance. Has no effect if return.variance.explained = FALSE.
flip.genotypes	a logical value indicating whether the genotypes of some genetic variants should be flipped (relabeled) to ensure that all minor allele frequencies (MAFs) < 0.5. Default = FALSE, with warning of any MAF > 0.5.
impute.method	a method for imputation of missing genotypes. It can be either "mean" (default) or "blue". If "mean" the genotypes will be imputed by the simple mean values. If "blue" the best linear unbiased estimates (BLUEs) of mean genotypes will be calculated taking into account the relationships between individuals [McPeek, et al., 2004, DOI: 10.1111/j.0006-341X.2004.00180.x] and used for imputation.
rho	If TRUE the optimal test is used [Lee, et al., 2012]. rho can be a vector of grid values from 0 to 1. The default grid is (0 : 10) / 10.
write.file	output file name to write results as they come (sequential mode only).
...	other arguments that could be passed to null(), read.plink() and readVCFToMatrixByGene().

Details

By default, FFBSKAT uses the linear weighted kernel function to set the inter-individual similarity matrix $K = GWWG^T$, where G is the $n \times p$ genotype matrix for n individuals and p genetic variants in the region, and W is the $p \times p$ diagonal weight matrix. Given the shape parameters of the beta function, $\mathrm{beta.par} = \mathrm{c(a,\ b)}$, the weights are defined using probability density function of the beta distribution:

$$W_i = (B(a,b))^{-1} MAF_i^{a-1}(1-MAF_i)^{b-1},$$

where MAF_i is a minor allelic frequency for the i^{th} genetic variant in the region, which is estimated from genotypes, and $B(a,b)$ is the beta function. This way of defining weights is the same as in original SKAT (see [Wu, et al., 2011] for details). $\mathrm{beta.par} = \mathrm{c}(1,\ 1)$ corresponds to the unweighted SKAT. The formula:

$$Q = 0.5\tilde{y}^T\Omega^{-1}K\Omega^{-1}\tilde{y}$$

is used to calculate score statistic, where $\tilde{y}$ and Ω are environmental residuals and covariance matrix obtained under the null hypothesis, respectively. Depending on the method option chosen, either Kuonen or Davies method is used to calculate P values from the score statistic Q. Both an Applied Statistics algorithm that inverts the characteristic function of the mixture chisq [Davies, 1980] and a saddlepoint approximation [Kuonen, 1999] are nearly exact, with the latter usually being a bit faster. For other kernel types, see [Wu, et al., 2011].

Value

A list with values:

results	a data frame containing P values, numbers of variants and polymorphic variants for each of analyzed regions. If $\mathrm{return.variance.explained} = \mathrm{TRUE}$ it contains also the column with marginal amounts of variance explained by each region. If $\mathrm{reml} = \mathrm{FALSE}$ the new estimates of heritability (h2) and total variance with corresponding total log-likelihood are also returned.
nullmod	an object containing the estimates of the null model parameters: heritability (h2), total variance (total.var), estimates of fixed effects of covariates (alpha), the gradient (df), and the total log-likelihood (logLH).
sample.size	the sample size after omitting NAs.
time	If $\mathrm{return.time} = \mathrm{TRUE}$ a list with running times for null model, regional analysis and total analysis is returned. See $\mathrm{proc.time()}$ for output format.

References

Svishcheva G.R., Belonogova N.M. and Axenovich T.I. (2014) FFBSKAT: Fast Family-Based Sequence Kernel Association Test. PLoS ONE 9(6): e99407. doi:10.1371/journal.pone.0099407
Davies R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 29, N 3, P. 323-333.
Kuonen D. (1999) Saddlepoint Approximations for Distributions of Quadratic Forms in Normal Variables. Biometrika, Vol. 86, No. 4, P. 929-935.
Wu M.C., et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., Vol. 89, P. 82-93.
Lee S., et al. (2012) Optimal unified approach for rare variant association testing with application to

small sample case-control whole-exome sequencing studies. American Journal of Human Genetics, 91, 224-237.

Examples

```
data(example.data)

## Run FFBSKAT with sliding window (default):
out <- FFBSKAT(trait ~ age + sex, phenodata, genodata, kin)

## Run FFBSKAT with regions defined in snpdata$gene and with
## null model obtained in first run:
out <- FFBSKAT(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, regions = snpdata$gene)

## Run FFBSKAT parallelized on two cores (this will require
## 'foreach' and 'doParallel' R-packages installed and
## cores available):
out <- FFBSKAT(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, ncores = 2)

## Run FFBSKAT with genotypes in VCF format:
VCFfileName <- system.file(
"testfiles/1000g.phase1.20110521.CFH.var.anno.vcf.gz",
package = "FREGAT")
geneFile <- system.file("testfiles/refFlat_hg19_6col.txt.gz",
package = "FREGAT")
phe <- data.frame(trait = rnorm(85))
out <- FFBSKAT(trait, phe, VCFfileName, geneFile = geneFile,
reg = "CFH", annoType = "Nonsynonymous",
flip.genotypes = TRUE)

## Run FFBSKAT with genotypes in PLINK binary data format:
bedFile <- system.file("testfiles/sample.bed",
package = "FREGAT")
phe <- data.frame(trait = rnorm(120))
out <- FFBSKAT(trait, phe, bedFile)
```

MLR *multiple linear regression*

Description

A multiple linear regression for familial or population data

Usage

```
MLR(formula, phenodata, genodata, kin = NULL, nullmod, regions = NULL,
sliding.window = c(20, 10), mode = "add", ncores = 1,
return.time = FALSE, stat = "F", impute.method = 'mean',
write.file = FALSE, ...)
```

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait or covariates will be omitted.
genodata	an object with genotypes to analyze. Several formats are allowed: - a data frame or matrix (with individuals in the rows and genetic variants in the columns) containing genotypes coded as AA = 0, Aa = 1 and aa = 2, where a is a minor allele. - for PLINK binary data format, a character string indicating a *.bed file name (*.bim and *.fam files should have the same prefix). This will make use of read.plink() function. - for VCF format, a character string indicating a *vcf.gz file name. This will require seqminer R-package to be installed. Its readVCFToMatrixByGene() function will be used to read VCF file gene-wise. The function also requires a geneFile, a text file listing all genes in refFlat format (see Examples below). VCF file should be bgzipped and indexed by Tabix. - an object of gwaa.data or snp.data class (this will require GenABEL R-package to be installed).
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
nullmod	an object containing parameter estimates under the null model. Setting nullmod allows to avoid re-estimation of the null model that does not depend on genotypes and can be calculated once for a trait. If not set, the null model parameters will be estimated within the function. The nullmod object in proper format can be obtained by null.model() function or any analysis function in FREGAT.
regions	an object assigning regions to be analyzed. This can be: - a vector of length equal to the number of genetic variants assigning the region for each variant (see Examples). - a data frame / matrix with names of genetic variants in the first column and names of regions in the second column (this format allows overlapping regions). - for VCF format, a character vector with names of genes to analyze. If NULL, sliding.window parameters will be used.
sliding.window	the sliding window size and step. Has no effect if regions is defined.
mode	the mode of inheritance: "add", "dom" or "rec" for additive, dominant or recessive mode, respectively. For dominant (recessive) mode genotypes will be recoded as AA = 0, Aa = 1 and aa = 1 (AA = 0, Aa = 0 and aa = 1), where a is a minor allele. Default mode is additive.
ncores	number of CPUs for parallel calculations. Default = 1.
return.time	a logical value indicating whether the running time should be returned.
stat	the statistic to be used to calculate the P values. One of "F" (default), "Chisq", "LRT".
impute.method	a method for imputation of missing genotypes. It can be either "mean" (default) or "blue". If "mean" the genotypes will be imputed by the simple mean values. If "blue" the best linear unbiased estimates (BLUEs) of mean genotypes will be calculated taking into account the relationships between individuals [McPeek, et al., 2004, DOI: 10.1111/j.0006-341X.2004.00180.x] and used for imputation.

write.file	output file name to write results as they come (sequential mode only).
...	other arguments that could be passed to null(), read.plink() and readVCFToMatrixByGene().

Value

A list with values:

results	a data frame containing P values, numbers of variants and polymorphic variants for each of analyzed regions.
nullmod	an object containing the estimates of the null model parameters: heritability (h2), total variance (total.var), estimates of fixed effects of covariates (alpha), the gradient (df), and the total log-likelihood (logLH).
sample.size	the sample size after omitting NAs.
time	If return.time = TRUE a list with running times for null model, regional analysis and total analysis is returned. See proc.time() for output format.

Examples

```
data(example.data)

## Run MLR with sliding window (default):
out <- MLR(trait ~ age + sex, phenodata, genodata, kin)

## Run MLR with regions defined in snpdata$gene and with
## null model parameters obtained in the first run:
out <- MLR(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, regions = snpdata$gene)

## Run MLR parallelized on two cores (this will require
## 'foreach' and 'doParallel' R-packages installed and
## cores available):
out <- MLR(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, ncores = 2)

## Run MLR with genotypes in VCF format:
VCFfileName <- system.file(
"testfiles/1000g.phase1.20110521.CFH.var.anno.vcf.gz",
package = "FREGAT")
geneFile <- system.file("testfiles/refFlat_hg19_6col.txt.gz",
package = "FREGAT")
phe <- data.frame(trait = rnorm(85))
out <- MLR(trait, phe, VCFfileName, geneFile = geneFile,
reg = "CFH", annoType = "Nonsynonymous")

## Run MLR with genotypes in PLINK binary data format:
bedFile <- system.file("testfiles/sample.bed",
package = "FREGAT")
phe <- data.frame(trait = rnorm(120))
out <- MLR(trait, phe, bedFile)
```

null *Fitting the null model*

Description

Gives estimation of model parameters under the null hypothesis

Usage

null(formula, phenodata, kin = NULL, opt.method = 'optimize',
ih2 = 0.3, eps = 1.e-04)

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait or covariates will be omitted.
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
opt.method	optimization method, one of "optimize" (default), "optim" and "nlminb". Corresponding R functions will be used in optimization.
ih2	initial value for h2. Default = 0.3.
eps	epsilon (precision) value for optimization. Default = 1e-04.

Details

The function performs one-dimensional optimization for h2 with analytical calculations for the other parameters.

Value

A list with values:

h2	estimate of the heritability
total.var	estimate of the total variance
alpha	estimates of fixed effects of covariates
df	the gradient
logLH	the total log-likelihood
p.normality	p value of Shapiro-Wilk normality test for the null model residuals.

Examples

```
data(example.data)

## Run the null model:
nullmod <- null(trait ~ age + sex, phenodata, kin)

## SKAT with the null model object obtained in the first run:
out <- FFBSKAT(trait ~ age + sex, phenodata, genodata, kin, nullmod)
```

---

read.plink *Read a PLINK binary data file*

Description

The package PLINK saves genome-wide association data in groups of three files, with the extensions .bed, .bim, and .fam. This function reads these files and creates a matrix with numeric genotypes and two data frames with information from the .bim, and .fam files.

Usage

```
read.plink(bed, bim, fam, na.strings = c("0", "-9"), sep = "." ,
select.subjects = NULL, select.snps = NULL)
```

Arguments

bed	The name of the file containing the packed binary SNP genotype data. It should have the extension .bed; if it doesn't, then this extension will be appended.
bim	The file containing the SNP descriptions.
fam	The file containing subject (and, possibly, family) identifiers. This is basically a tab-delimited "pedfile".
na.strings	Strings in .bam and .fam files to be recoded as NA.
sep	A separator character for constructing unique subject identifiers.
select.subjects	A numeric vector indicating a subset of subjects to be selected from the input file (see Details).
select.snps	Either a numeric or a character vector indicating a subset of SNPs to be selected from the input file (see Details).

Details

If the bed argument does not contain a file name with the file extension .bed, then this extension is appended to the argument. The remaining two arguments are optional; their default values are obtained by replacing the .bed file name extension by .bim and .fam respectively. See the PLINK documentation for the detailed specification of these files.

The select.subjects or select.snps argument can be used to read a subset of the data. Use of select.snps requires that the .bed file is in SNP-major order (the default in PLINK). Likewise, use of select.subjects requires that the .bed file is in individual-major order. Subjects are selected by their numeric order in the PLINK files, while SNPs are selected either by order or by name. Note

that the order of selected SNPs/subjects in the output objects will be the same as their order in the PLINK files.

Row names for the output object and for the accompanying subject description data frame are taken as the pedigree identifiers, when these provide the required unique identifiers. When these are duplicated, an attempt is made to use the pedigree-member identifiers instead but, when these too are duplicated, row names are obtained by concatenating, with a separator character, the pedigree and pedigree-member identifiers.

Value

A list with three elements:

genotypes	The output genotype data as a numeric matrix.
fam	A data frame corresponding to the .fam file, containing the first six fields in a standard pedfile. The row names will correspond with those of the genotype matrix.
map	A data frame corresponding to the .bim file. the row names correspond with the column names of the genotype matrix.

Author(s)

Originally written by David Clayton (snpStats package), modified by Nadezhda Belonogova

References

PLINK: Whole genome association analysis toolset. http://zzz.bwh.harvard.edu/plink/

Examples

```
bedFile <- system.file("testfiles/sample.bed", package = "FREGAT")
data <- read.plink(bedFile)
```

---

single.point *A single-point association test*

Description

A single-point association test for familial or population data

Usage

```
single.point(formula, phenodata, genodata, kin = NULL, nullmod,
regions = NULL, mode = "add", ncores = 1, return.time = FALSE,
impute.method = 'mean', ...)
```

Arguments

formula	referring to the column(s) in phenodata to be analyzed as outcome and, if needed, covariates.
phenodata	a data frame containing columns mentioned in formula: trait to analyze and, if needed, covariates. Individuals not measured for trait or covariates will be omitted.
genodata	an object with genotypes to analyze. Several formats are allowed: - a data frame or matrix (with individuals in the rows and genetic variants in the columns) containing genotypes coded as AA = 0, Aa = 1 and aa = 2, where a is a minor allele. - for PLINK binary data format, a character string indicating a *.bed file name (*.bim and *.fam files should have the same prefix). This will make use of read.plink() function. - an object of gwaa.data or snp.data class (this will require GenABEL R-package to be installed).
kin	a square symmetric matrix giving the pairwise kinship coefficients between analyzed individuals. Under default kin = NULL all individuals will be considered as unrelated.
nullmod	an object containing parameter estimates under the null model. Setting nullmod allows to avoid re-estimation of the null model that does not depend on genotypes and can be calculated once for a trait. If not set, the null model parameters will be estimated within the function. The nullmod object in proper format can be obtained by null.model() function or any analysis function in FREGAT.
regions	an object assigning regions for genetic variants in genodata can be either - a vector of length equal to the number of genetic variants assigning the region for each variant (see Examples). - a data frame / matrix with names of genetic variants in the first column and names of regions in the second column (this format allows overlapping regions). If NULL, all variants in genodata will be analysed.
mode	the mode of inheritance: "add", "dom" or "rec" for additive, dominant or recessive mode, respectively. For dominant (recessive) mode genotypes will be recoded as AA = 0, Aa = 1 and aa = 1 (AA = 0, Aa = 0 and aa = 1), where a is a minor allele. Default mode is additive.
ncores	number of CPUs for parallel calculations. Default = 1.
return.time	a logical value indicating whether the running time should be returned.
impute.method	a method for imputation of missing genotypes. It can be either "mean" (default) or "blue". If "mean" the genotypes will be imputed by the simple mean values. If "blue" the best linear unbiased estimates (BLUEs) of mean genotypes will be calculated taking into account the relationships between individuals [McPeek, et al., 2004, DOI: 10.1111/j.0006-341X.2004.00180.x] and used for imputation.
...	other arguments that could be passed to null() and read.plink().

Value

A list with values:

results	a data frame containing P values, estimates of betas and their s.e., MAFs and sample size for each of analyzed variants.

nullmod	an object containing the estimates of the null model parameters: heritability (h2), total variance (total.var), estimates of fixed effects of covariates (alpha), the gradient (df), and the total log-likelihood (logLH).
time	If return.time = TRUE a list with running times for null model, regional analysis and total analysis is returned. See proc.time() for output format.

Examples

```
data(example.data)

## Run single.point analysis for all variants in genodata (default):
out <- single.point(trait ~ age + sex, phenodata, genodata, kin)

## Run single.point analysis for regions defined in snpdata$gene and with
## null model parameters obtained in the first run:
out <- single.point(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, regions = snpdata$gene)

## Run single.point analysis parallelized on two cores (this will require
## 'foreach' and 'doParallel' R-packages installed and
## cores available):
out <- single.point(trait ~ age + sex, phenodata, genodata, kin,
out$nullmod, ncores = 2)

## Run single.point analysis with genotypes in PLINK binary data format:
bedFile <- system.file("testfiles/sample.bed",
package = "FREGAT")
phe <- data.frame(trait = rnorm(120))
out <- single.point(trait, phe, bedFile)
```

Index