Package 'MiRKAT'

April 14, 2020

Title Microbiome Regression-Based Analysis Tests

Version 1.1.0

Maintainer Anna Plantinga <amp9@williams.edu>

Description Test for overall association between microbiome composition data and phenotypes via phylogenetic kernels.
The phenotype can be univariate continuous or binary (Zhao et al. (2015) <doi:10.1016/j.ajhg.2015.04.003>),
survival outcomes (Plantinga et al. (2017) <doi:10.1186/s40168-017-0239-9>),
multivariate (Zhan et al. (2017) <doi:10.1002/gepi.22030>) and
structured phenotypes (Zhan et al. (2017) <doi:10.1111/biom.12684>).
The package can also use robust and quantile regression (Fu et al. (2020+), in preparation).
In each case, the microbiome community effect is modeled nonparametrically through a kernel function,
which can incorporate phylogenetic tree information.

Depends R (>= 3.0.2)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

NeedsCompilation no

Imports MASS, CompQuadForm, quantreg, GUniFrac, PearsonDS, lme4, Matrix, permute, mixtools, survival, ecodist, stats

Suggests knitr, propr, cluster, dirmult, vegan, rmarkdown, tidyverse, kableExtra

VignetteBuilder knitr, rmarkdown

Author Anna Plantinga [aut, cre],
Nehemiah Wilson [aut, ctb],
Haotian Zheng [aut, ctb],
Xiang Zhan [aut, ctb],
Michael Wu [aut],
Ni Zhao [aut, ctb],
Jun Chen [aut]

Repository CRAN

Date/Publication 2020-04-14 08:50:16 UTC

R topics documented:

`bindata` *Simulated DEPENDENT data with BINOMIAL traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)*

Description

Simulated DEPENDENT data with BINOMIAL traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)

Usage

```
data(bindata)
```

Format

A list containing three data objects for correlated microbiome data with binary response variable (described below).

bin.otu.tab Simulated OTU data for correlated regression-based analysis; 59 rows and 730 columns, rows being patients and columns being OTUs

bin.meta Simulated metadata for correlated regression-based analysis; 59 rows and 4 columns, rows being patients and columns being the outcome variable, subject identifier, and covariates to possibly account for in any regression modeling

bin.tree Simulated rooted phylogenetic tree with 730 tips and 729 nodes

---

CSKAT *Small-sample SKAT for correlated (continuous) data ('c' stands for 'correlated')*

Description

Compute the adjusted score statistic and p-value

Usage

```
CSKAT(formula.H0, data = NULL, Ks, nperm = 999)
```

Arguments

formula.H0	A two-sided linear formula object describing both the fixed-effects and random-effects part of the model under the null, use the same syntax as the "lmer" in "lme4" package
data	An optional data frame containing the variables named in formula. Default: NULL.
Ks	A kernel matrix or list of kernels, quantifying the similarities between samples.
nperm	Number of permutations for calculating the omnibus p-value. Ignored unless Ks is a list of candidate kernels.

Value

p.value Association p-values

Q.adj Adjusted score statistics

Author(s)

Nehemiah Wilson, Anna Plantinga, Xiang Zhan, Jun Chen.

References

Zhan X, et al. (2018) A small-sample kernel association test for correlated data with application to microbiome association studies. Genet Epidemiol.

Examples

```
Y <- rnorm(100)
Z <- matrix(rnorm(200), 100, 2)
ID <- gl(20, 5)
G <- matrix(rbinom(1000, 2, 0.05), 100, 10)
K <- G %*% t(G)
CSKAT(formula.H0 = Y ~ Z + (1 | ID), Ks = K)
```

D2K *D2K*

Description

Construct kernel matrix from distance matrix.

Usage

```
D2K(D)
```

Arguments

D	An n by n matrix giving pairwise distances or dissimilarites, where n is sample size.

Details

Converts a distance matrix (matrix of pairwise distances) into a kernel matrix for microbiome data. The kernel matrix is constructed as $K = -(I - 11'/n)D^2(I - 11'/n)/2$, where D is the pairwise distance matrix, I is the identity matrix, and 1 is a vector of ones.

D^2 represents element-wise square.

To ensure that K is positive semi-definite, a positive semi-definiteness correction is conducted

Value

An n by n kernel or similarity matrix corresponding to the distance matrix given.

Author(s)

Ni Zhao

References

Zhao, Ni, et al. "Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test

Examples

```
library(GUniFrac)

#Load in data and create a distance matrix
data(throat.tree)
data(throat.otu.tab)
unifracs <- GUniFrac(throat.otu.tab, throat.tree, alpha=c(1))$unifracs
D1 <- unifracs[,,"d_1"]

#Function call
K <- D2K(D1)
```

GLMMMiRKAT *The Microbiome Regression-based Kernel Association Test Based on the Generalized Linear Mixed Model*

Description

GLMMMiRKAT utilizes a generalized linear mixed model to allow dependence among samples.

Usage

```
GLMMMiRKAT(
  y,
  X = NULL,
  Ks,
  id = NULL,
  time.pt = NULL,
  model,
  method = "perm",
  formula.H0 = NULL,
  slope = FALSE,
  nperm = 5000
)
```

Arguments

y	A numeric vector of Gaussian (e.g., body mass index), Binomial (e.g., disease status, treatment/placebo) or Poisson (e.g., number of tumors/treatments) traits.
X	A vector or matrix of numeric covariates, if applicable (default = NULL).
Ks	A list of n-by-n OTU kernel matrices or one singular n-by-n OTU kernel matrix, where n is sample size.
id	A vector of cluster (e.g., family or subject including repeated measurements) IDs. Defaults to NULL since it is unnecessary for the CSKAT call.
time.pt	A vector of time points for the longitudinal studies. 'time.pt' is not required (i.e., 'time.pt = NULL') for the random intercept model. Default is time.pt = NULL.

model	A string declaring which model ("gaussian", "binmoial" or "poisson") is to be used; should align with whether a Gaussian, Binomial, or Poisson trait is being inputted for the y argument.
method	A string declaring which method ("permu" or "davies) will be used to calculate the p-value. Davies is only available for Gaussian traits. Defaults to "perm".
formula.H0	A two-sided linear formula object under the null, indicating the variables to adjust. Handles both the random and mixed effects. Needed only if model = "gaussian" and method = "davies. Defaults to NULL.
slope	An indicator to include random slopes in the model (slope = TRUE) or not (slope = FALSE). 'slope = FALSE' is for the random intercept model. 'slope = TRUE' is for the random slope model. For the random slope model (slope = TRUE), 'time.pt' is required.
nperm	The number of permutations used to calculate the p-values and omnibus p-value. Defaults to 5000.

Details

Missing data is not permitted. Please remove all individuals with missing y, X, and Ks prior to input for analysis.

y and X (if not NULL) should be numerical matrices or vectors with the same number of rows.

Ks should either be a list of n by n kernel matrices (where n is sample size) or a single kernel matrix. If you have distance matrices from metagenomic data, each kernel can be constructed through function D2K. Each kernel can also be constructed through other mathematical approaches.

If model="gaussian" and method="davies", CSKAT is called. CSKAT utilizes the same omnibus test as GLMMMiRKAT. See ?CSKAT for more details.

formula.H0 is required only if model="gaussian" and method="davies". In all other situations, it may be left NULL.

The "method" argument only determines kernel-specific p-values are generated. When Ks is a list of multiple kernels, an omnibus p-value is computed via permutation.

Value

Returns a p-value for each inputted kernel matrix, as well as an overall omnibus p-value if more than one kernel matrix is inputted

p_values	p-value for each individual kernel matrix
omnibus_p	overall omnibus p-value calculated by permutation for the adaptive GLMM-MiRKAT analysis

Author(s)

Hyunwook Koh

References

Koh H, Li Y, Zhan X, Chen J, Zhao N. (2019) A distance-based kernel association test based on the generalized linear mixed model for correlated microbiome studies. Front. Genet. 458(10), 1-14.

Examples

```
library(vegan)

## Example with Gaussian (e.g., body mass index) traits
## For non-Gaussian traits, see vignette.

# Import example microbiome data with Gaussian traits
data(nordata)
otu.tab <- nordata$nor.otu.tab
meta <- nordata$nor.meta

# Create kernel matrices
# could use phylogenetic kernels as below; computation time is slightly higher
# tree <- nordata$nor.tree
# unifracs <- GUniFrac::GUniFrac(otu.tab, tree, alpha=c(1))$unifracs
D_BC = as.matrix(vegdist(otu.tab, 'bray'))
K_BC = D2K(D_BC)

# Run GLMM-MiRKAT
GLMMMiRKAT(y = meta$y, X = cbind(meta$x1, meta$x2), id = meta$id,
          Ks = K_BC, model = "gaussian", nperm = 500)
```

`inner.CSKAT` *Inner Function for CSKAT, Correlated Sequence Kernel Association Test*

Description

Small-sample SKAT for correlated (continuous) data ('c' stands for 'correlated'). Computes the adjusted score statistic and p-value.

Usage

```
inner.CSKAT(formula.H0, data = NULL, K)
```

Arguments

`formula.H0`	A two-sided linear formula object under the null, indicating the variables to adjust. Only handles the mixed effects. Needed only if model = "gaussian" and method = "davies".
`data`	an optional data frame containing the variables named in formula.
`K`	the kernel matrix, which quantifies the similarities between samples

Value

p.value association p-value

Q.adj adjusted score statistic

References

Zhan X, et al. (2018) A small-sample kernel association test for correlated data with application to microbiome association studies. Genet Epidemiol., submitted.

`inner.KRV` *Kernel RV Coefficient Test; Inner Function*

Description

Function called when user calls function KRV. For each kernel matrix inputted into KRV, KRV runs inner.KRV on that kernel with the inputted kernel.y outcome matrix.

Usage

```
inner.KRV(
  y = NULL,
  X = NULL,
  kernel.otu,
  kernel.y,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

`y`	A numeric n by p matrix of p continuous phenotype variables and sample size n (default = NULL). If y is NULL, a phenotype kernel matrix must be entered for "kernel.y". Defaults to NULL.
`X`	A numeric n by q matrix, containing q additional covariates (default = NULL). If NULL, an intercept only model was fit. No covariate adjustment is possible if a matrix is provided for kernel.y.
`kernel.otu`	A numeric n by n kernel matrix, where n is sample size. It can be constructed from microbiome data, such as by transforming from a distance metric.
`kernel.y`	Either a numeric n by n kernel matrix of phenotype or a method to compute the kernel of phenotype. Gaussian kernel (kernel.y="Gaussian") can capture general relationship between microbiome and phenotypes; and linear kernel (kernel.y="linear") can be preferred if the underlying relationship is close to linear.
`returnKRV`	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
`returnR2`	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

y and X (if not NULL) should all be numerical matrices or vectors with the same number of rows.

Ks should be a list of n by n matrices or a single matrix. If you have distance metric from metagenomic data, each kernel can be constructed through function D2K. Each kernel can also be constructed through other mathematical approaches.

Missing data is not permitted. Please remove all individuals with missing y, X, Ks prior to analysis

Parameter "method" only concerns how kernel specific p-values are generated. When Ks is a list of multiple kernels, omnibus p-value is computed through permutation from each individual p-value, which are calculated through method of choice.

Value

Returns a p-value for the candidate kernel matrix

pv	p-value for the candidate kernel matrix
KRV	KRV statistic for the candidate kernel matrix. Only returned if returnKRV = TRUE.
R2	R-squared for the candidate kernel matrix. Only returned if returnR2 = TRUE.

Author(s)

Haotian Zheng, Xiang Zhan, Ni Zhao

References

Zhan, X., Plantinga, A., Zhao, N., and Wu, M.C. A Fast Small-Sample Kernel Independence Test for Microbiome Community-Level Association Analysis. Biometrics. 2017 Mar 10. doi: 10.1111/biom.12684.

KRV *Kernel RV Coefficient Test (KRV)*

Description

Kernel RV coefficient test to evaluate the overall association between microbiome composition and high-dimensional or structured phenotype.

Usage

```
KRV(
  y = NULL,
  X = NULL,
  kernels.otu,
  kernel.y,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

y	A numeric n by p matrix of p continuous phenotype variables and sample size n (default = NULL). If it is NULL, a phenotype kernel matrix must be entered for "kernel.y". Defaults to NULL.
X	A numeric n by q matrix, containing q additional covariates (default = NULL). If NULL, an intercept only model is used. No covariate adjustment is possible if a matrix is provided in kernel.y.
kernels.otu	A numeric OTU n by n kernel matrix or a list of matrices, where n is the sample size. It can be constructed from microbiome data, such as by transforming from a distance metric.
kernel.y	Either a numerical n by n kernel matrix for phenotypes or a method to compute the kernel of phenotype. Methods are "Gaussian" or "linear". A Gaussian kernel (kernel.y="Gaussian") can capture the general relationship between microbiome and phenotypes; a linear kernel (kernel.y="linear") may be preferred if the underlying relationship is close to linear.
returnKRV	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
returnR2	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

kernels.otu should be a list of numerical n by n kernel matrices, or a single n by n kernel matrix, where n is sample size.

When kernel.y is a method ("Gaussian" or "linear") to compute the kernel of phenotype, y should be a numerical phenotype matrix, and X (if not NULL) should be a numeric matrix of covariates. Both y and X should have n rows.

When kernel.y is a kernel matrix for the phenotype, there is no need to provide X and y, and they will be ignored if provided. In this case, kernel.y and kernel.otu should both be numeric matrices with the same number of rows and columns.

Missing data is not permitted. Please remove all individuals with missing kernel.otu, y (if not NULL), X (if not NULL), and kernel.y (if a matrix is entered) prior to analysis.

Value

If only one candidate kernel matrix is considered, returns a list containing the p-value for the candidate kernel matrix. If more than one candidate kernel matrix is considered, returns a list of two elements:

p_values	P-value for each candidate kernel matrix
omnibus_p	Omnibus p-value
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Nehemiah Wilson, Haotian Zheng, Xiang Zhan, Ni Zhao

References

Zheng, Haotian, Zhan, X., Plantinga, A., Zhao, N., and Wu, M.C. A Fast Small-Sample Kernel Independence Test for Microbiome Community-Level Association Analysis. Biometrics. 2017 Mar 10. doi: 10.1111/biom.12684.

Examples

```
library(GUniFrac)
library(MASS)

data(throat.tree)
data(throat.otu.tab)
data(throat.meta)


set.seed(123)
n = nrow(throat.otu.tab)
Sex <- throat.meta$Sex
Smoker <- throat.meta$SmokingStatus
anti <- throat.meta$AntibioticUsePast3Months_TimeFromAntibioticUsage
Male = (Sex == "Male")**2
Smoker =(Smoker == "Smoker") **2
anti =  (anti != "None")^2
cova = cbind(Male,  anti)

otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff
unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs

D.weighted = unifracs[,,"d_1"]
D.unweighted = unifracs[,,"d_UW"]
D.BC= as.matrix(vegdist(otu.tab.rff , method="bray"))


K.weighted = D2K(D.weighted)
K.unweighted = D2K(D.unweighted)
K.BC = D2K(D.BC)

rho = 0.2
Va = matrix(rep(rho, (2*n)^2), 2*n, 2*n)+diag(1-rho, 2*n)
G = mvrnorm(n, rep(0, 2*n), Va)

KRV(kernels.otu = K.weighted, kernel.y = G %*% t(G))
```

MiRKAT *Microbiome Regression-based Kernel Association Test*

Description

Test for association between microbiome composition and a continuous or dichotomous outcome by incorporating phylogenetic or nonphylogenetic distance between different microbiomes.

Usage

```
MiRKAT(
  y,
  X = NULL,
  Ks,
  out_type = "C",
  method = "davies",
  nperm = 999,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

y	A numeric vector of the a continuous or dichotomous outcome variable.
X	A numeric matrix or data frame, containing additional covariates that you want to adjust for. If NULL, a intercept only model is used. Defaults to NULL.
Ks	A list of n by n kernel matrices or a single n by n kernel matrix, where n is the sample size. It can be constructed from microbiome data through distance metric or other approaches, such as linear kernels or Gaussian kernels.
out_type	An indicator of the outcome type ("C" for continuous, "D" for dichotomous).
method	Method used to compute the kernel specific p-value. "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "moment" represents an approximation method that matches the first two moments. "permutation" represents a permutation approach for p-value calculation. Defaults to "davies".
nperm	The number of permutations if method = "permutation" or when multiple kernels are considered. If method = "davies" or "moment", nperm is ignored. Defaults to 999.
returnKRV	A logical indicating whether to return the KRV statistic (a measure of effect size). Defaults to FALSE.
returnR2	A logical indicating whether to return R-squared. Defaults to FALSE.

Details

y and X (if not NULL) should all be numeric matrices or vectors with the same number of rows.

Ks should be a list of n by n matrices or a single matrix. If you have distance metric(s) from metagenomic data, each kernel can be constructed through function D2K. Each kernel can also be constructed through other mathematical approaches.

Missing data is not permitted. Please remove all individuals with missing y, X, Ks prior to analysis

Parameter "method" only concerns with how kernel specific p-values are generated. When Ks is a list of multiple kernels, omnibus p-value is computed through permutation from each individual p-values, which are calculated through method of choice.

Value

Returns a list containing the following elements:

`p_values`	P-value for each candidate kernel matrix
`omnibus_p`	Omnibus p-value considering multiple candidate kernel matrices
`KRV`	Kernel RV statistic (a measure of effect size). Only returned if returnKRV = TRUE.
`R2`	R-squared. Only returned if returnR2 = TRUE.

Author(s)

Ni Zhao

References

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40: 5-19. doi: 10.1002/gepi.21934

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 , 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (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.

Zhou, J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.

Examples

```
library(GUniFrac)
library(vegan)

data(throat.tree)
data(throat.otu.tab)
data(throat.meta)

unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
Ds = list(w = unifracs[,,"d_1"], uw = unifracs[,,"d_UW"],
          BC= as.matrix(vegdist(throat.otu.tab, method="bray")))
Ks = lapply(Ds, FUN = function(d) D2K(d))

covar = cbind(throat.meta$Age, as.numeric(throat.meta$Sex == "Male"))

# Continuous phenotype
n = nrow(throat.meta)
y = rnorm(n)
MiRKAT(y, X = covar, Ks = Ks, out_type="C", method = "davies")

# Binary phenotype
y = as.numeric(runif(n) < 0.5)
MiRKAT(y, X = covar, Ks = Ks, out_type="D")
```

`MiRKAT.Q` *Robust MiRKAT (quantile regression)*

Description

A more robust version of MiRKAT utlizing a linear model that uses quantile regression.

Usage

```
MiRKAT.Q(y, X, Ks, returnKRV = FALSE, returnR2 = FALSE)
```

Arguments

`y`	A numeric vector of the a continuous or dichotomous outcome variable.
`X`	A numerical matrix or data frame, containing additional covariates that you want to adjust for. Mustn't be NULL.
`Ks`	A list of n by n kernel matrices (or a single n by n kernel matrix), where n is the sample size. If you have distance metric from metagenomic data, each kernel can be constructed through function D2K. Each kernel can also be constructed through other mathematical approaches, such as linear or Gaussian kernels.
`returnKRV`	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
`returnR2`	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

MiRKAT.Q creates a kernel matrix using the linear model created with the function rq, a quantile regression function, then does the KRV analysis on Ks and the newly formed kernel matrix representing the outome traits.

Missing data is not permitted. Please remove all individuals with missing y, X, Ks prior to analysis

Value

Returns p-values for each individual kernel matrix, an omnibus p-value if multiple kernels were provided, and measures of effect size KRV and R2.

p_values	labeled individual p-values for each kernel
omnibus_p	omnibus p_value, calculated as for the KRV test
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Weija Fu

Examples

```
library(quantreg)
library(vegan)

# Generate data
library(GUniFrac)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)

unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
Ds = list(w = unifracs[,,"d_1"], uw = unifracs[,,"d_UW"],
         BC= as.matrix(vegdist(throat.otu.tab, method="bray")))
Ks = lapply(Ds, FUN = function(d) D2K(d))

covar = scale(cbind(throat.meta$Age, as.numeric(throat.meta$Sex == "Male")))

# Continuous phenotype
n = nrow(throat.meta)
y = rchisq(n, 2) + apply(covar, 1, sum)
MiRKAT.Q(y, X = covar, Ks = Ks)
```

MiRKAT.R *Robust MiRKAT (robust regression)*

Description

A more robust version of MiRKAT utilizing a linear model by robust regression using an M estimator.

Usage

```
MiRKAT.R(y, X, Ks, returnKRV = FALSE, returnR2 = FALSE)
```

Arguments

y	A numeric vector of the a continuous or dichotomous outcome variable.
X	A numerical matrix or data frame, containing additional covariates that you want to adjust for Mustn't be NULL
Ks	list of n by n kernel matrices (or a single n by n kernel matrix), where n is the sample size. It can be constructed from microbiome data through distance metric or other approaches, such as linear kernels or Gaussian kernels.
returnKRV	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
returnR2	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

MiRKAT.R creates a kernel matrix using the linear model created with the function rlm, a robust regression function, then does the KRV analysis on Ks and the newly formed kernel matrix representing the outome traits.

y and X should all be numerical matrices or vectors with the same number of rows, and mustn't be NULL.

Ks should be a list of n by n matrices or a single matrix. If you have distance metric from metagenomic data, each kernel can be constructed through function D2K. Each kernel may also be constructed through other mathematical approaches.

Missing data is not permitted. Please remove all individuals with missing y, X, Ks prior to analysis

Value

Returns p-values for each individual kernel matrix, an omnibus p-value if multiple kernels were provided, and measures of effect size KRV and R2.

p_values	labeled individual p-values for each kernel
omnibus_p	omnibus p_value, calculated as for the KRV test
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Weijia Fu

Examples

```
# Generate data
library(GUniFrac)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)

unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
Ds = list(w = unifracs[,,"d_1"], uw = unifracs[,,"d_UW"],
         BC= as.matrix(vegdist(throat.otu.tab, method="bray")))
Ks = lapply(Ds, FUN = function(d) D2K(d))

covar = cbind(throat.meta$Age, as.numeric(throat.meta$Sex == "Male"))

# Continuous phenotype
n = nrow(throat.meta)
y = rchisq(n, 2)
MiRKAT.R(y, X = covar, Ks = Ks)
```

MiRKATS *Microiome Regression-based Kernel Association Test for Survival*

Description

Community level test for association between microbiome composition and survival outcomes (right-censored time-to-event data) using kernel matrices to compare similarity between microbiome profiles with similarity in survival times.

Usage

```
MiRKATS(
  obstime,
  delta,
  X = NULL,
  Ks,
  beta = NULL,
  perm = FALSE,
  nperm = 999,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

obstime	A numeric vector of follow-up (survival/censoring) times.
delta	Event indicator: a vector of 0/1, where 1 indicates that the event was observed for a subject (so "obstime" is survival time), and 0 indicates that the subject was censored.
X	A vector or matrix of numeric covariates, if applicable (default = NULL).
Ks	A list of or a single numeric n by n kernel matrices or matrix (where n is the sample size).
beta	A vector of coefficients associated with covariates. If beta is NULL and covariates are present, coxph is used to calculate coefficients (default = NULL).
perm	Logical, indicating whether permutation should be used instead of analytic p-value calculation (default=FALSE). Not recommended for sample sizes of 100 or more.
nperm	Integer, number of permutations used to calculate p-value if perm==TRUE (default=1000)
returnKRV	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
returnR2	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

obstime, delta, and X should all have n rows, and the kernel or distance matrix should be a single n by n matrix. If a distance matrix is entered (distance=TRUE), a kernel matrix will be constructed from the distance matrix.

Update in v1.1.0: MiRKATS also utilizes the OMiRKATS omnibus test if more than one kernel matrix is provided by the user. The OMiRKATS omnibus test calculates an overall p-value for the test via permutation.

Missing data is not permitted. Please remove individuals with missing data on y, X or in the kernel or distance matrix prior to using the function.

The Efron approximation is used for tied survival times.

Value

Return value depends on the number of kernel matrices inputted. If more than one kernel matrix is given, MiRKATS returns two items; a vector of the labeled individual p-values for each kernel matrix, as well as an omnibus p-value from the Optimal-MiRKATS omnibus test. If only one kernel matrix is given, then only its p-value will be given, as no omnibus test will be needed.

p_values	individual p-values for each inputted kernel matrix
omnibus_p	overall omnibus p-value
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Nehemiah Wilson, Anna Plantinga

References

Plantinga, A., Zhan, X., Zhao, N., Chen, J., Jenq, R., and Wu, M.C. MiRKAT-S: a distance-based test of association between microbiome composition and survival times. Microbiome, 2017:5-17. doi: 10.1186/s40168-017-0239-9

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40:5-19. doi: 10.1002/gepi.21934

Efron, B. (1977) "The efficiency of Cox's likelihood function for censored data." Journal of the American statistical Association 72(359):557-565.

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, 29:323-333

Examples

```
###################################
# Generate data
library(GUniFrac)

# Throat microbiome data
data(throat.tree)
data(throat.otu.tab)

unifracs = GUniFrac(throat.otu.tab, throat.tree, alpha = c(1))$unifracs
Ds = list(w = unifracs[,,"d_1"], uw = unifracs[,,"d_UW"],
         BC= as.matrix(vegdist(throat.otu.tab, method="bray")))
Ks = lapply(Ds, FUN = function(d) D2K(d))

# Covariates and outcomes
covar <- matrix(rnorm(120), nrow=60)
S <- rexp(60, 3)   # survival time
C <- rexp(60, 1)   # censoring time
D <- (S<=C)        # event indicator
U <- pmin(S, C)    # observed follow-up time

MiRKATS(obstime = U, delta = D, X = covar, Ks = Ks, beta = NULL)
```

`MiRKAT_binary` *Microbiome Regression-Based Kernel Association Test for binary outcomes*

Description

Called by MiRKAT if the outcome variable is dichotomous (out_type="D")

This function is called by the exported function MiRKAT if the argument "out_type" of MiRKAT is equal to "D" (for dichotomous).

Each argument of MiRKAT_continuous is given the value of the corresponding argument given by the user to MiRKAT.

Function not exported

Usage

```
MiRKAT_binary(
  y,
  X = NULL,
  Ks,
  method = "davies",
  family = "binomial",
  nperm = 999,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

`y`	A numeric vector of the dichotomous outcome variable
`X`	A numerical matrix or data frame, containing additional covariates that you want to adjust for (Default = NULL). If it is NULL, a intercept only model was fit.
`Ks`	A list of n by n kernel matrices (or a single n by n kernel matrix), where n is the sample size. It can be constructed from microbiome data through distance metric or other approaches, such as linear kernels or Gaussian kernels.
`method`	A string telling R which method to use to compute the kernel specific p-value (default = "davies"). "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "moment" represents an approximation method that matches the first two moments. "permutation" represents a permutation approach for p-value calculation.
`family`	A string describing the error distribution and link function to be used in the linear model.
`nperm`	the number of permutations if method = "permutation" or when multiple kernels are considered. if method = "davies" or "moment", nperm is ignored.

returnKRV	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
returnR2	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Value

If only one candidate kernel matrix is considered, returns a list containing the p-value for the candidate kernel matrix. If more than one candidate kernel matrix is considered, returns a list with two elements: the individual p-values for each candidate kernel matrix, and the omnibus p-value.

indivP	p-value for each candidate kernel matrix
omnibus_p	omnibus p-value if multiple kernel matrices are considered
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Ni Zhao

References

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40: 5-19. doi: 10.1002/gepi.21934

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 , 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (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.

Zhou, J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.

`MiRKAT_continuous` *Microbiome Regression-based Analysis Test for a continuous outcome variable*

Description

Inner function for MiRKAT; computes MiRKAT for continuous outcomes. Called by MiRKAT if out_type="C"

Usage

```
MiRKAT_continuous(
  y,
  X = NULL,
  Ks,
  method = "davies",
  nperm = 999,
  returnKRV = FALSE,
  returnR2 = FALSE
)
```

Arguments

`y`	A numeric vector of the continuous outcome variable
`X`	A numeric matrix or data frame containing additional covariates (default = NULL). If NULL, an intercept only model is used.
`Ks`	A list of n by n kernel matrices (or a single n by n kernel matrix), where n is the sample size. It can be constructed from microbiome data through distance metric or other approaches, such as linear kernels or Gaussian kernels.
`method`	A method to compute the kernel specific p-value (Default= "davies"). "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq. We adopt an exact variance component tests because most of the studies concerning microbiome compositions have modest sample size. "moment" represents an approximation method that matches the first two moments. "permutation" represents a permutation approach for p-value calculation.
`nperm`	the number of permutations if method = "permutation" or when multiple kernels are considered. If method = "davies" or "moment", nperm is ignored. Defaults to 999.
`returnKRV`	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
`returnR2`	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

This function is called by the exported function "MiRKAT" when the argument of MiRKAT, out_type, is set equal to "C".

Each argument of MiRKAT_continuous is given the value of the corresponding argument given by the user to MiRKAT.

Function not exported

Value

If only one candidate kernel matrix is considered, returns a list containing the p-value for the candidate kernel matrix. If more than one candidate kernel matrix is considered, returns a list of two elements: the individual p-values for each candidate kernel matrix, and the omnibus p-value

`indivP`	p-value for each candidate kernel matrix
`omnibus_p`	omnibus p-value considering multiple candidate kernel matrices
`KRV`	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
`R2`	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Ni Zhao

References

Zhao, N., Chen, J.,Carroll, I. M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J. J., Ringel, Y., Li, H. and Wu, M.C. (2015)). Microbiome Regression-based Kernel Association Test (MiRKAT). American Journal of Human Genetics, 96(5):797-807

Chen, J., Chen, W., Zhao, N., Wu, M~C.and Schaid, D~J. (2016) Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. 40: 5-19. doi: 10.1002/gepi.21934

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 , 29, 323-333.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. Biom. Bull. 2, 110-114.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (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.

Zhou, J. J. and Zhou, H.(2015) Powerful Exact Variance Component Tests for the Small Sample Next Generation Sequencing Studies (eVCTest), in submission.

MMiRKAT *Multivariate Microbiome Regression-based Kernel Association Test*

Description

Test for association between overall microbiome composition and multiple continuous outcomes.

Usage

```
MMiRKAT(Y, X = NULL, Ks, returnKRV = FALSE, returnR2 = FALSE)
```

Arguments

Y	A numerical n by p matrix of p continuous outcome variables, n being sample size.
X	A numerical n by q matrix or data frame, containing q additional covariates that you want to adjust for (Default = NULL). If it is NULL, an intercept only model is fit.
Ks	A list of numerical n by n kernel matrices, or a single n by n kernel matrix, where n is the sample size. Kernels can be constructed from distance matrices (such as Bray-Curtis or UniFrac distances) using the function D2K, or through other mathematical approaches.
returnKRV	A logical indicating whether to return the KRV statistic. Defaults to FALSE.
returnR2	A logical indicating whether to return the R-squared coefficient. Defaults to FALSE.

Details

Missing data is not permitted. Please remove all individuals with missing Y, X, K prior to analysis

The method of generating kernel specific p-values is "davies", which represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq.

Value

Returns a list of the MMiRKAT p-values for each inputted kernel matrix, labeled with the names of the kernels, if given.

p_values	list of the p-values for each individual kernel matrix inputted
KRV	A vector of kernel RV statistics (a measure of effect size), one for each candidate kernel matrix. Only returned if returnKRV = TRUE
R2	A vector of R-squared statistics, one for each candidate kernel matrix. Only returned if returnR2 = TRUE

Author(s)

Nehemiah Wilson, Haotian Zheng, Xiang Zhan, Ni Zhao

References

Zheng, H., Zhan, X., Tong, X., Zhao, N., Maity,A., Wu, M.C., and Chen,J. A small-sample multivariate kernel machine test for microbiome association studies. Genetic Epidemiology, 41(3), 210-220. DOI: 10.1002/gepi.22030

Examples

```
library(GUniFrac)
data(throat.tree)
data(throat.otu.tab)
data(throat.meta)

unifracs <- GUniFrac(throat.otu.tab, throat.tree, alpha=c(0, 0.5, 1))$unifracs

Ds = list(w = unifracs[,,"d_1"], u = unifracs[,,"d_UW"],
         BC= as.matrix(vegdist(throat.otu.tab , method="bray")))
Ks = lapply(Ds, FUN = function(d) D2K(d))

n = nrow(throat.otu.tab)
Y = matrix(rnorm(n*3, 0, 1), n, 3)

covar = cbind(as.numeric(throat.meta$Sex == "Male"), as.numeric(throat.meta$PackYears))
MMiRKAT(Y = Y, X = covar, Ks = Ks)
```

---

nordata *Simulated DEPENDENT data with GAUSSIAN traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)*

Description

Simulated DEPENDENT data with GAUSSIAN traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)

Usage

```
data(nordata)
```

Format

A list containing three data objects for correlated microbiome data with continuous response variable (described below).

nor.otu.tab Simulated OTU data for correlated regression-based analysis; 59 rows and 730 columns, rows being patients and columns being OTUs

nor.meta Simulated metadata for correlated regression-based analysis; 59 rows and 4 columns, rows being patients and columns being the outcome variable, subject identifier, and covariates to possibly account for in any regression modeling

nor.tree Simulated rooted phylogenetic tree with 730 tips and 729 nodes

---

`poisdata` *Simulated DEPENDENT data with POISSON (count) traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)*

Description

Simulated DEPENDENT data with POISSON (count) traits for correlated regression-based analysis (i.e. CSKAT, GLMMMiRKAT)

Usage

```
data(poisdata)
```

Format

A list containing three data objects for correlated microbiome data with binary response variable (described below).

pois.otu.tab Simulated OTU data for correlated regression-based analysis; 59 rows and 730 columns, rows being patients and columns being OTUs

pois.meta Simulated metadata for correlated regression-based analysis; 59 rows and 4 columns, rows being patients and columns being the outcome variable, subject identifier, and covariates to possibly account for in any regression modeling

pois.tree Simulated rooted phylogenetic tree with 730 tips and 729 nodes

---

`throat.meta` *Simulated metadata for microbiome regression-based analysis*

Description

Simulation code can be seen in ?KRV Corresponding OTU matrix is stored in "throat.otu.tab"

Usage

```
data(throat.meta)
```

Format

A data frame with 59 rows and 16 columns, rows being participants and columns being different covariates to possibly be accounted for in any utilized linear models.

throat.meta is part of a microbiome data set for studying the effect of smoking on the upper respiratory tract microbiome. This data set comes from the throat microbiome of left body side. It contains 60 subjects consisting of 32 nonsmokers and 28 smokers.

BarcodeSequence Sequence of DNA that allows for the identification of the specific species of bacteria. See GUniFrac for more details

LinkerPrimerSequence Sequence of DNA that aids in locating the Barcode Sequence. See GUniFrac for more details

SmokingStatus whether or not each patient is a "Smoker" or a "NonSmoker"

PatientID Identifying integer label given to each patient

SampleIndex Labels each patient as being from this particular sample, so as possibly be able to use multiple samples at once

AirwaySite Part of body where our samples were taken from from in each participant

SideOfBody Which side of the body the samples were taken from

SampleType What kind of sample each one is; should all be patientsamples

RespiratoryDiseaseStatus_severity_timeframe Whether or not the patient has had a respiratory disease, and if so which one_severity of said disease_whether or not that disease is still active. If there has been no such disease in the patient's medical history, the patient's value is "healthy" in this column

AntibioticUsePast3Months_TimeFromAntibioticUsage Whether or not the patient has used antibiotics in the past month_if so, how long ago it was. If not antibiotics have been used in the past month, the patient's value is "None" in this column

Age Age of the patient

Sex The sex of the patient

PackYears Unit of measurement measuring the intensity of smoking; average number of packs per day times the number of years the patient has been smoking. If patient has never smoked, their value is 0 for this column

TimeFromLastCig Minutes since the patient's last cigarette

TimeFromLastMeal Minutes since the patient's last meal

Description See Charleston paper and other sources

---

throat.otu.tab *Simulated OTU data for microbiome regression analysis*

Description

Simulated code can be seen in ?KRV Corresponding metadata is stored in "throat.meta"

Usage

```
data(throat.otu.tab)
```

Format

60 rows and 856 columns, where rows are patients and columns are OTUs

---

throat.tree *Simulated rooted phylogenetic tree*

Description

Simulation code can be seen in ?KRV

Usage

```
data(throat.tree)
```

Format

Phylogenetic tree with 856 tips and 855 internal nodes

Details

Corresponding OTU matrix stored in "throat.otu.tab" See the GUniFrac package for more details

Index