Package 'GUniFrac'

August 18, 2021

Type Package

Title Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis

Version 1.3

Date 2021-08-12

Author Jun Chen

Maintainer Jun Chen <chen.jun2@mayo.edu>

Description A suite of methods for powerful and robust microbiome data analysis including data normalization, data simulation, community-level association testing and differential abundance analysis. It implements generalized UniFrac distances, Geometric Mean of Pairwise Ratios (GMPR) normalization, semiparametric data simulator, distance-based statistical methods, and feature-based statistical methods. The distance-based statistical methods include three extensions of PERMANOVA: (1) PERMANOVA using the Freedman-Lane permutation scheme, (2) PERMANOVA omnibus test using multiple matrices, and (3) analytical approach to approximating PERMANOVA p-value. Feature-based statistical methods include linear model-based permutation tests for differential abundance analysis of zero-inflated compositional data.

Depends R (>= 3.5.0), vegan

Suggests ade4

Imports Rcpp (>= 0.12.13), matrixStats, Matrix, ape, parallel, stats, utils, statmod, rmutil, dirmult, MASS

LinkingTo Rcpp

NeedsCompilation yes

License GPL-3

Encoding UTF-8

Repository CRAN

Date/Publication 2021-08-18 21:50:22 UTC

R topics documented:

`adonis3` *Permutational Multivariate Analysis of Variance Using Distance Matrices (Freedman-Lane permutation)*

Description

Analysis of variance using distance matrices — for partitioning distance matrices among sources of variation and fitting linear models (e.g., factors, polynomial regression) to distance matrices; uses a permutation test (Freedman-Lane permutation) with pseudo-F ratios.

Usage

```
adonis3(formula, data, permutations = 999, method = "bray",
    strata = NULL, contr.unordered = "contr.sum",
    contr.ordered = "contr.poly", parallel = getOption("mc.cores"), ...)
```

Arguments

`formula`	model formula. The LHS must be either a community data matrix or a dissimilarity matrix, e.g., from `vegdist` or `dist`. If the LHS is a data matrix, function `vegdist` will be used to find the dissimilarities. The RHS defines the independent variables. These can be continuous variables or factors, they can be transformed within the formula, and they can have interactions as in a typical `formula`.
`data`	the data frame for the independent variables.
`permutations`	a list of control values for the permutations as returned by the function `how`, or the number of permutations required, or a permutation matrix where each row gives the permuted indices.

method	the name of any method used in vegdist to calculate pairwise distances if the left hand side of the formula was a data frame or a matrix.
strata	groups (strata) within which to constrain permutations.
contr.unordered, contr.ordered	contrasts used for the design matrix (default in R is dummy or treatment contrasts for unordered factors).
parallel	number of parallel processes or a predefined socket cluster. With parallel = 1 uses ordinary, non-parallel processing. The parallel processing is done with **parallel** package.
...	Other arguments passed to vegdist.

Details

adonis3 is the re-implementation of the famous adonis function in the vegan package based on the Freedman-Lane permutation scheme. (Freedman & Lane (1983), Hu & Satten (2020)). adonis is the function for the analysis and partitioning sums of squares using dissimilarities. The original implementation in the vegan package is directly based on the algorithm of Anderson (2001) and performs a sequential test of terms. Statistical significance is calculated based on permuting the distance matrix. As shown in Chen & Zhang (2020+), such permutation will lead to power loss in testing the effect of a covariate of interest while adjusting for other covariates (confounders). The power loss is more evident when the confounders' effects are strong, the correlation between the covariate of interest and the confounders is high, and the sample size is small. When the sample size is large than 100, the difference is usually small. The new implementation is revised on the adonis function with the same interface.

Value

Function adonis3 returns an object of class "adonis" with following components:

aov.tab	typical AOV table showing sources of variation, degrees of freedom, sequential sums of squares, mean squares, F statistics, partial R^2 and P values, based on N permutations.
coefficients	matrix of coefficients of the linear model, with rows representing sources of variation and columns representing species; each column represents a fit of a species abundance to the linear model. These are what you get when you fit one species to your predictors. These are NOT available if you supply the distance matrix in the formula, rather than the site x species matrix
coef.sites	matrix of coefficients of the linear model, with rows representing sources of variation and columns representing sites; each column represents a fit of a sites distances (from all other sites) to the linear model. These are what you get when you fit distances of one site to your predictors.
f.perms	an N by m matrix of the null F statistics for each source of variation based on N permutations of the data. The permutations can be inspected with permustats and its support functions.
model.matrix	the model.matrix for the right hand side of the formula.
terms	the terms component of the model.

Author(s)

Martin Henry H. Stevens (adonis) and Jun Chen (adonis3).

References

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**: 32–46.

Freedman D. & Lane D. 1983. A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, **1**292–298.

Hu, Y. J. & Satten, G. A. 2020. Testing hypotheses about the microbiome using the linear decomposition model (LDM). Bioinformatics.

Examples

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

groups <- throat.meta$SmokingStatus

# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff

# Calculate the UniFrac distance
unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs

# Test the smoking effect based on unweighted UniFrac distance, adjusting sex
adonis3(as.dist(unifracs[, , 'd_UW']) ~ Sex + SmokingStatus, data = throat.meta)
```

---

dmanova *Distance-based Multivariate Analysis of Variance (Analytical P-value Calculation)*

Description

Analysis of variance using distance matrices — for partitioning distance matrices among sources of variation and fitting linear models (e.g., factors, polynomial regression) to distance matrices; calculate the analytical p-value based on pseudo-F statistic without permutation.

Usage

```
dmanova(formula, data = NULL, positify = FALSE,
contr.unordered = "contr.sum", contr.ordered = "contr.poly",
returnG = FALSE)
```

Arguments

formula	model formula. The LHS must be a dissimilarity matrix (either class matrix or class dist, e.g., from vegdist or dist. The RHS defines the independent variables. These can be continuous variables or factors, they can be transformed within the formula, and they can have interactions as in a typical formula.
data	the data frame for the independent variables.
positify	a logical value indicating whether to make the Gower's matrix positive definite using the nearPD function in Matrix package. This is equivalent to modifying the distance matrix so that it has an Euclidean embedding.
contr.unordered, contr.ordered	contrasts used for the design matrix (default in R is dummy or treatment contrasts for unordered factors).
returnG	a logical value indicating whether the Gower's matrix should be returned.

Details

dmanova is a permutation-free method for approximating the p-value from distance-based permutational multivariate analysis of variance (PERMANOVA). PERMANOVA is slow when the sample size is large. In contrast, dmanova provides an analytical solution, which is several orders of magnitude faster for large sample sizes. The covariate of interest should be put as the last term in formula while the variables to be adjusted are put before the covariate of interest.

Value

Function dmanova returns a list with the following components:

aov.tab	typical AOV table showing sources of variation, degrees of freedom, sums of squares, mean squares, F statistics, partial R^2 and P values.
df	degree of freedom for the Chisquared distribution.
G	The Gower's matrix if returnG is true.
call	the call made

Author(s)

Jun Chen and Xianyang Zhang

References

Chen, J. & Zhang, X. 2021. D-MANOVA: fast distance-based multivariate analysis of variance for large-scale microbiome association studies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btab498

See Also

adonis3

Examples

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

groups <- throat.meta$SmokingStatus

# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff

# Calculate the UniFrac distance
unifracs <- GUniFrac(otu.tab.rff, throat.tree, alpha=c(0, 0.5, 1))$unifracs

# Test the smoking effect based on unweighted UniFrac distance, adjusting sex
# 'Sex' should be put before 'SmokingStatus'
dmanova(as.dist(unifracs[, , 'd_UW']) ~ Sex + SmokingStatus, data = throat.meta)
```

GMPR *Geometric Mean of Pairwise Ratios (GMPR) Normalization for Zero-inflated Count Data*

Description

A robust normalization method for zero-inflated count data such as microbiome sequencing data.

Usage

```
GMPR(OTUmatrix, min_ct = 2, intersect_no = 4)
```

Arguments

`OTUmatrix`	An OTU count table, where OTUs are arranged in rows and samples in columns.
`min_ct`	The minimal number of OTU counts. Only those OTU pairs with at least `min_ct` counts are considered in the ratio calculation. The default is 2.
`intersect_no`	The minimal number of shared OTUs between samples. Only those sample pairs sharing at least `intersect_no` OTUs are considered in geometric mean calculation. The default is 4.

Details

Normalization is a critical step in microbiome sequencing data analysis to account for variable library sizes. Microbiome data contains a vast number of zeros, which makes the traditional RNA-Seq normalization methods unstable. The proposed GMPR normalization remedies this problem by switching the two steps in DESeq2 normalization:

First, to calculate rij, the median count ratio of nonzero counts between samples: rij=median(cki/ckj) (k in 1:OTU_number and cki, ckj is the non-zero count of the kth OTU)

Second, to calculate the size factor si for a given sample i: si=geometric_mean(rij)

Value

A vector of GMPR size factor for each sample.

Author(s)

Jun Chen and Lujun Zhang

References

Li Chen, James Reeve, Lujun Zhang, Shenbing Huang, and Jun Chen. 2018. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ, 6, e4600.

Examples

```
data(throat.otu.tab)
size.factor <- GMPR(t(throat.otu.tab))
```

`GUniFrac` *Generalized UniFrac distances for comparing microbial communities.*

Description

A generalized version of commonly used UniFrac distances. It is defined as:

$$d^{(\alpha)} = \frac{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^\alpha \left| \frac{p_i^A - p_i^B}{p_i^A + p_i^B} \right|}{\sum_{i=1}^{m} b_i (p_i^A + p_i^B)^\alpha},$$

where m is the number of branches, b_i is the length of ith branch, p_i^A, p_i^B are the branch proportion for community A and B.

Generalized UniFrac distance contains an extra parameter α controlling the weight on abundant lineages so the distance is not dominated by highly abundant lineages. $\alpha = 0.5$ is overall very robust.

The unweighted and weighted UniFrac, and variance-adjusted weighted UniFrac distances are also implemented.

Usage

```
GUniFrac(otu.tab, tree, alpha = c(0, 0.5, 1))
```

Arguments

otu.tab	an OTU count table, row - n sample, column - q OTU
tree	a rooted phylogenetic tree of R class "phylo"
alpha	the parameter controlling weight on abundant lineages

Value

Return a list containing

unifracs	a three dimensional array containing all the UniFrac distance matrices

Note

The function only accepts rooted tree. To root a tree, you may consider using `midpoint` from the package `phangorn`.

Author(s)

Jun Chen <chen.jun2@mayo.edu>

References

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D. and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. 28(16): 2106–2113.

See Also

`Rarefy`, `PermanovaG`

Examples

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

groups <- throat.meta$SmokingStatus

# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff

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

dw <- unifracs[, , "d_1"] # Weighted UniFrac
du <- unifracs[, , "d_UW"] # Unweighted UniFrac
dv <- unifracs[, , "d_VAW"] # Variance adjusted weighted UniFrac
d0 <- unifracs[, , "d_0"]      # GUniFrac with alpha 0
d5 <- unifracs[, , "d_0.5"]    # GUniFrac with alpha 0.5
```

```
# Permanova - Distance based multivariate analysis of variance
adonis3(as.dist(d5) ~ groups)
```

PermanovaG *Permutational Multivariate Analysis of Variance Using Multiple Distance Matrices*

Description

In practice, we do not know a priori which type of change happens in the microbiome. Each distance measure is most powerful in detecting only a certain scenario. When multiple distance matrices are available, separate tests using each distance matrix will lead to loss of power due to multiple testing correction. Combing the distance matrices in a single test will improve power. PermanovaG combines multiple distance matrices by taking the minimum of the P values for individual distance matrices. Significance is assessed by permutation.

Usage

```
PermanovaG(formula, data = NULL, ...)
```

Arguments

formula	a formula, left side of the formula (Y ~ X) is a three dimensional ARRAY containing the supplied distance matrices as produced by GUniFrac function. Or it could be a list of distance matrices.
data	a data frame containing the covariates
...	parameter passing to adonis function

Value

Return a list containing:

p.tab	a data frame, columns: p-values for individual distance matrices and the omnibus test, rows: covariates. (Note: they are sequential p-values, put the variable of interest in the end)
aov.tab.list	a list of adonis AOV tables for individual distance matrices

Author(s)

Jun Chen <chen.jun2@mayo.edu>

References

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D. and Li, H.(2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. 28(16): 2106–2113.

See Also

Rarefy, GUniFrac

Examples

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

groups <- throat.meta$SmokingStatus

# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff

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

# Combine unweighted and weighted UniFrac for testing
PermanovaG(unifracs[, , c("d_1", "d_UW")] ~ groups)
# Combine d(0), d(0.5), d(1) for testing
```

PermanovaG2 *Permutational Multivariate Analysis of Variance Using Multiple Distance Matrices(Freedman-Lane Permutation)*

Description

In practice, we do not know a priori which type of change happens in the microbiome. Each distance measure is most powerful in detecting only a certain scenario. When multiple distance matrices are available, separate tests using each distance matrix will lead to loss of power due to multiple testing correction. Combing the distance matrices in a single test will improve power. PermanovaG combines multiple distance matrices by taking the minimum of the P values for individual distance matrices. Significance is assessed by permutation.

Usage

```
PermanovaG2(formula, data = NULL, ...)
```

Arguments

formula	a formula, left side of the formula (Y ~ X) is a three dimensional ARRAY containing the supplied distance matrices as produced by GUniFrac function. Or it could be a list of distance matrices.
data	a data frame containing the covariates
...	parameters passing to adonis function

Value

Return a list containing:

p.tab	a data frame, columns: p-values for individual distance matrices and the omnibus test, rows: covariates. (Note: they are sequential p-values, put the variable of interest in the end)
aov.tab.list	a list of adonis AOV tables for individual distance matrices

Author(s)

Jun Chen <chen.jun2@mayo.edu>

References

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D. and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. 28(16): 2106–2113.

See Also

Rarefy, GUniFrac, adonis3

Examples

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

groups <- throat.meta$SmokingStatus

# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab)$otu.tab.rff

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

# Combine unweighted and weighted UniFrac for testing
PermanovaG2(unifracs[, , c("d_1", "d_UW")] ~ groups)
```

Rarefy *Rarefy a Count Table to Equal Sequencing Depth*

Description

GUniFrac is also sensitive to different sequencing depth. To compare microbiomes on an equal basis, rarefaction might be used.

Usage

```
Rarefy(otu.tab, depth = min(rowSums(otu.tab)))
```

Arguments

otu.tab	OTU count table, row - n sample, column - q OTU
depth	required sequencing depth; If not specified, the lowest sequencing depth is used.

Value

Return a list containing:

otu.tab.rff	rarefied OTU table
discard	IDs of samples that does not reach the specified sequencing depth

Author(s)

Jun Chen <chen.jun2@mayo.edu>

References

Chen, J., Bittinger, K., Charlson, E.S., Hoffmann, C., Lewis, J., Wu, G.D., Collman, R.G., Bushman, F.D. and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. 28(16): 2106–2113.

Examples

```
data(throat.otu.tab)
# Rarefaction
otu.tab.rff <- Rarefy(throat.otu.tab, 1024)$otu.tab.rff
```

SimulateMSeq *A Semiparametric Model-based Microbiome Sequencing Data Simulator*

Description

The function generates synthetic microbiome sequencing data for studying the performance of differential abundance analysis methods. It uses a user-supplied (large) reference OTU table as a template to generate a synthetic OTU table of specified size. A subset of OTUs are affected by a simulated covariate of interest, either binary or continuous. Confounder effects can also be simulated. The function allows simulating different signal structures, i.e., the percentage of differential OTUs, their effect sizes, their direction of change, and whether these OTUs are relatively abundant or rare.

Usage

```
SimulateMSeq(
  ref.otu.tab,
  nSam = 100,
  nOTU = 500,
  diff.otu.pct = 0.1,
  diff.otu.direct = c("balanced", "unbalanced"),
  diff.otu.mode = c("abundant", "rare", "mix"),
  covariate.type = c("binary", "continuous"),
  grp.ratio = 1,
  covariate.eff.mean = 1,
  covariate.eff.sd = 0,
  confounder.type = c("none", "binary", "continuous", "both"),
  conf.cov.cor = 0.6,
  conf.diff.otu.pct = 0,
  conf.nondiff.otu.pct = 0.1,
  confounder.eff.mean = 0,
  confounder.eff.sd = 0,
  error.sd = 0,
  depth.mu = 10000,
  depth.theta = 5,
  depth.conf.factor = 0
)
```

Arguments

`ref.otu.tab`	a matrix, the reference OTU count table (row - OTUs, column - samples), serving as the template for synthetic sample generation.
`nSam`	the number of samples to be simulated.
`nOTU`	the number of OTUs to be simulated.
`diff.otu.pct`	a numeric value between 0 and 1, the percentage of differential OTUs to be simulated. If 0, global null setting is simulated. The default is 0.1.
`diff.otu.direct`	a character string of "balanced" or "unbalanced". "balanced" - the direction of change for these differential OTUs is random, "unbalanced" - direction of change is the same. The default is "balanced".
`diff.otu.mode`	a character string of "rare", "mix" or "abundant". "abundant" - differential OTU come from the top quartile of the abundance distribution, "rare" - differential OTU come from the bottom quartile of the abundance distribution, and "mix" - random set. The default is "abundant".
`covariate.type`	a character string of "binary" or "continuous", indicating the type of the covariate to be simulated. The default is "binary" (e.g., case v.s. control).
`grp.ratio`	a numeric value between 0 and 1. Group size ratio. The default is 1, i.e., equal group size. Only relevant when `covariate.type` is "binary".
`covariate.eff.mean`	a numeric value, the mean log fold change (effect size) in response to one unit change of the covariate. The default is 1.

`covariate.eff.sd`
: a positive numeric value, the standard deviation of the log fold change. The default is 0, i.e., the log fold change is the same across differential OTUs.

`confounder.type`
: a character string of "none", "binary", "continuous" or "both". The default is "none", no confounder will be simulated. If "both", both a binary and continuous confounder will be simulated. The default is "none".

`conf.cov.cor`
: a numeric value between 0 and 1. The correlation between the covariate of interest and the confounder. The default is 0.6.

`conf.diff.otu.pct`
: a numeric value between 0 and 1. The percentage of OTUs affected by the confounder and the covariate of interest. The default is 0.

`conf.nondiff.otu.pct`
: a numeric value between 0 and 1. The percentage of OTUs affected by the confounder but not the covariate of interest. The default is 0.1.

`confounder.eff.mean`
: a numeric value, the mean log fold change (effect size) in response to one unit change of the confounder. The default is 1.

`confounder.eff.sd`
: a positive numeric value, the standard deviation of the log fold change for the confounder. The default is 0, i.e., the log fold change is the same across OTUs affected by the confounder.

`error.sd`
: the sd of the log fold change unexplained by the covariate and the confounder (i.e., the error term under the log linear model). The default is 0.

`depth.mu`
: the mean sequencing depth to be simulated. The default is 10,000.

`depth.theta`
: the theta value of the negative binomial distribution controlling the variance (`mu + mu^2/theta`). The default is 5.

`depth.conf.factor`
: a numeric value controlling the dependence of the sequencing depth on the covariate of interest (`depth.mu * exp(scale(X) * depth.conf.factor)`). The default is 0, i.e., the depth is not associated with the covariate of interest. This parameter can be used to simulate depth confounding.

Details

This function implements a semiparametric approach for realistic microbiome sequencing data generation. The method draws random samples from a large reference dataset (non-parametric part) and uses these reference samples as templates to generate new samples (parametric part). Specifically, for each drawn reference sample, it infers the underlying composition based on a Bayesian model and then adds covariate/confounder effects to the composition vector, based on which a new sequencing sample is generated. The method circumvents the difficulty in modeling the inter-subject variation of the microbiome composition.

Value

Return a list with the elements:

`otu.tab.sim`
: simulated OTU table

covariate	simulated covariate of interest
confounder	simulated confounder(s)
diff.otu.ind	indices of the differential OTUs, i.e., affected by the covariate of interest
otu.names	the names of the simulated OTUs
conf.otu.ind	indices of OTUs affected by the confounder(s)

Author(s)

Jun Chen and Lu Yang

References

Yang, L. & Chen, J. 2021+. A comprehensive evaluation of differential abundance analysis methods: current status and potential solutions. Submitted.

Examples

```
# Use throat microbiome for illustration
data(throat.otu.tab)
comm <- t(throat.otu.tab)
comm <- comm[rowMeans(comm != 0) > 0.2, ]

# Simulate binary covariate, 10% signal density, abundant differential OTUs, unbalanced change
# This setting simulates strong compositional effects
sim.obj <- SimulateMSeq(
ref.otu.tab = comm, nSam = 50, nOTU = 50,
# True signal setting
diff.otu.pct = 0.1, diff.otu.direct = c("unbalanced"),
diff.otu.mode = c("abundant"),
covariate.type = c("binary"), grp.ratio = 1,
covariate.eff.mean = 1.0, covariate.eff.sd = 0,
# Confounder signal setting
confounder.type = c("both"), conf.cov.cor = 0.6,
conf.diff.otu.pct = 0.1, conf.nondiff.otu.pct = 0.1,
confounder.eff.mean = 1.0, confounder.eff.sd = 0,
# Depth setting
depth.mu = 10000, depth.theta = 5, depth.conf.factor = 0
)

meta.dat <- data.frame(X = sim.obj$covariate, Z1 = sim.obj$confounder[, 1],
                       Z2 = sim.obj$confounder[, 2])
otu.tab.sim <- sim.obj$otu.tab.sim

# Run ZicoSeq for differential abundance analysis
zico.obj <- ZicoSeq(meta.dat = meta.dat, feature.dat = otu.tab.sim,
grp.name = 'X', adj.name = c('Z1', 'Z2'), feature.dat.type = "count",
# Filter to remove rare taxa
prev.filter = 0.2, mean.abund.filter = 0,  max.abund.filter = 0.002, min.prop = 0,
# Winsorization to replace outliers
is.winsor = TRUE, outlier.pct = 0.03, winsor.end = 'top',
```

```
# Posterior sampling to impute zeros
is.post.sample = TRUE, post.sample.no = 25,
# Multiple link functions to capture diverse taxon-covariate relation
link.func = list(function (x) x^0.25, function (x) x^0.5, function (x) x^0.75),
stats.combine.func = max,
# Permutation-based multiple testing correction
perm.no = 99,  strata = NULL,
# Reference-based multiple stage normalization
ref.pct = 0.5, stage.no = 6, excl.pct = 0.2,
# Family-wise error rate control
is.fwer = FALSE,
verbose = TRUE, return.feature.dat = FALSE)

# Detected differential OTUs
which(zico.obj$p.adj.fdr <= 0.05)

# True differential OTUs
sim.obj$otu.names[sim.obj$diff.otu.ind]
```

`stool.otu.tab` *Stool Microbiome OTU Count Table*

Description

OTU count table from 16S V3-V5 targeted sequencing of the stool microbiome samples from the HMP project. A total of 2,094 OTUs from 295 samples.

Usage

```
data(stool.otu.tab)
```

Format

The format is: chr "stool.otu.tab"

Details

The OTU table was taken from R bioconductor "HMP16SData" package. OTUs with prevalence less than 10% and maximum proportion less than 0.2% were removed. This OTU table can be used for simulating stool microbiome sequencing data.

Source

Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, Dowd JB, Segata N, Waldron L (2019). "HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor." American Journal of Epidemiology. doi: 10.1093/aje/kwz006.

Examples

```
data(stool.otu.tab)
```

throat.meta *Throat Microbiome Meta Data*

Description

It is part of a microbiome data set for studying the effect of smoking on the upper respiratory tract microbiome. The original data set contains samples from both throat and nose microbiomes, and from both body sides. This data set comes from the throat microbiome of left body side. It contains 60 subjects consisting of 32 nonsmokers and 28 smokers.

Usage

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

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

Examples

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

throat.otu.tab *Throat Microbiome OTU Count Table*

Description

It is part of a microbiome data set (16S V12-targeted 454 pyrosequencing) for studying the effect of smoking on the upper respiratory tract microbiome. The original data set contains samples from both throat and nose microbiomes, and from both body sides. This data set comes from the throat microbiome of left body side. It contains 60 subjects consisting of 32 nonsmokers and 28 smokers.

Usage

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

Details

The OTU table is produced by the QIIME software. Singleton OTUs have been discarded.

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

Examples

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

`throat.tree` *UPGMA Tree of Throat Microbiome OTUs*

Description

The OTU tree is constructed using UPGMA on the K80 distance matrice of the OTUs. It is a rooted tree of class "phylo".

Usage

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

Details

The OTUs are produced by the QIIME software. Singleton OTUs have been discarded.

Source

Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, et al. (2010) Disordered Microbial Communities in the Upper Respiratory Tract of Cigarette Smokers. PLoS ONE 5(12): e15216.

Examples

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

`vaginal.otu.tab` *Vaginal Microbiome OTU Count Table*

Description

OTU count table from 16S V3-V5 targeted sequencing of the vaginal microbiome samples from the HMP project. A total of 780 OTUs from 381 samples.

Usage

```
data(vaginal.otu.tab)
```

Details

The OTU table was taken from R bioconductor "HMP16SData" package. OTUs with prevalence less than 10% and maximum proportion less than 0.2% were removed. This OTU table can be used for simulating vaginal microbiome sequencing data.

Source

Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, Dowd JB, Segata N, Waldron L (2019). "HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor." American Journal of Epidemiology. doi: 10.1093/aje/kwz006.

Examples

```
data(stool.otu.tab)
```

`ZicoSeq` *A linear Model-based Permutation Test for Differential Abundance Analysis of Microbiome Data and Other Omics Data*

Description

ZicoSeq is a permutation test (Smith permutation) for differential abundance analysis of microbiome sequencing data. The input can be a count or a proportion matrix. When a count matrix is provided, it provides an option to draw posterior samples of the underlying proportions to account for the sampling variability during the sequencing process. The test results are aggregated over these posterior samples. For both count and proportion data, a reference-based ratio approach is used to account for compositional effects. As a general methodology, ZicoSeq can also be applied to differential analysis of other omics data. In this case, they are not treated as compositional data.

Usage

```
ZicoSeq(
  meta.dat,
  feature.dat,
  grp.name,
  adj.name = NULL,
  feature.dat.type = c('count', 'proportion', 'other'),
  prev.filter = 0,
  mean.abund.filter = 0,
  max.abund.filter = 0,
  min.prop = 0,
  is.winsor = TRUE,
  outlier.pct = 0.03,
  winsor.end = c('top', 'bottom', 'both'),
  is.post.sample = TRUE,
  post.sample.no = 25,
  link.func = list(function(x) x^0.5),
  stats.combine.func = max,
  perm.no = 99,
  strata = NULL,
  ref.pct = 0.5,
  stage.no = 6,
  excl.pct = 0.2,
```

```
  is.fwer = FALSE,
  verbose = TRUE,
  return.feature.dat = FALSE
)
```

Arguments

meta.dat	a data frame containing the sample meta data.
feature.dat	a matrix of counts, row - features (OTUs, genes, etc) , column - samples.
grp.name	the name for the variable of interest. It could be numeric or categorical; should be in meta.dat.
adj.name	the name(s) for the variable(s) to be adjusted. Multiple variables are allowed. They could be numeric or categorical; should be in meta.dat.
feature.dat.type	the type of the feature data. It could be "count", "proportion" or "other". For "proportion" data type, posterior sampling will not be performed, but the reference-based ratio approach will still be used to address compositional effects. For "other" data type, neither posterior sampling or reference-base ratio approach will be used.
prev.filter	the prevalence (percentage of nonzeros) cutoff, under which the features will be filtered. The default is 0.
mean.abund.filter	the mean relative abundance cutoff, under which the features will be filtered. The default is 0.
max.abund.filter	the max relative abundance cutoff, under which the features will be filtered. The default is 0.
min.prop	proportions less than this value will be replaced with this value. Only relevant when log transformation is used. Default is 0.
is.winsor	a logical value indicating whether winsorization should be performed to replace outliers. The default is TRUE.
outlier.pct	the expected percentage of outliers. These outliers will be winsorized. The default is 0.03.
winsor.end	a character indicating whether the outliers at the "top", "bottom" or "both" will be winsorized. The default is "top". If the feature.dat.type is "other", "both" may be considered.
is.post.sample	a logical value indicating whether to perform posterior sampling of the underlying proportions. Only relevant when the feature data are counts.
post.sample.no	the number of posterior samples if posterior sampling is used. The default is 25.
link.func	a list of transformation functions for the feature data or the ratios. Based on our experience, square-root transformation is a robust choice for many datasets.
perm.no	the number of permutations. If the raw p values are of the major interest, set perm.no to at least 999.

`strata` a factor such as subject IDs indicating the permutation strata. Permutation will be confined to each stratum. This can be used for paired or some longitudinal designs.

`stats.combine.func` function to combine the F-statistic for the omnibus test. The default is `max`.

`ref.pct` percentage of reference taxa. The default is 0.5.

`stage.no` the number of stages if multiple-stage normalization is used. The default is 6.

`excl.pct` the maximum percentage of significant features (nominal p-value < 0.05) in the reference set that should be removed. Only relevant when multiple-stage normalization is used.

`is.fwer` a logical value indicating whether the family-wise error rate control (West-Young) should be performed.

`verbose` a logical value indicating whether the trace information should be printed out.

`return.feature.dat` a logical value indicating whether the winsorized, filtered "feature.dat" matrix should be returned.

Details

`ZicoSeq` is a linear model-based permutation test developed for differential abundance analysis of zero-inflated compositional data. Although its development is motivated by zero-inflated microbiome sequence count data, it can be applied to proportion (composition) data and more generally to other types of omics data. Currently, it has the following components: 1. Winsorization to decrease the influence of outliers; 2. Posterior sampling based on a beta mixture prior to address sampling variability and zero inflation; 3. Reference-based multiple-stage normalization to address compositional effects; 4. An omnibus test to address diverse feature-covariate relationships; 5. Permutation-based false discovery rate control / family-wise error rate control for multiple testing correction, which takes into account the correlation structure in the feature data.

Value

A list with the elements

`call` the call

`feature.dat` the winsorized, filtered `feature.dat` matrix.

`filter.ind` a vector of logical values indicating which features are tested.

`R2` a matrix of percent explained variance (number of features by number of transformation functions).

`F0` a matrix of F-statistics (number of features by number of transformation functions).

`RSS` a matrix of residual sum squares (number of features by number of transformation functions).

`df.model, df.residual` degrees of freedom for the model and residual space.

`p.raw` the raw p-values based on permutations (not accurate if `perm.no` is small).

`p.adj.fdr` permutation-based FDR-adjusted p-values.

`p.adj.fwer` permutation-based FWER-adjusted (West-Young) p-values.

Author(s)

Jun Chen

References

Yang, L. & Chen, J. 2021+. A comprehensive evaluation of differential abundance analysis methods: current status and potential solutions. To be submitted.

Examples

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

comm <- t(throat.otu.tab)
meta.dat <- throat.meta

set.seed(123)
# For count data
zico.obj <- ZicoSeq(meta.dat = meta.dat, feature.dat = comm,
grp.name = 'SmokingStatus', adj.name = 'Sex', feature.dat.type = "count",
# Filter to remove rare taxa
prev.filter = 0.2, mean.abund.filter = 0,  max.abund.filter = 0.002, min.prop = 0,
# Winsorization to replace outliers
is.winsor = TRUE, outlier.pct = 0.03, winsor.end = 'top',
# Posterior sampling to impute zeros
is.post.sample = TRUE, post.sample.no = 25,
# Multiple link functions to capture diverse taxon-covariate relation
link.func = list(function (x) x^0.25, function (x) x^0.5, function (x) x^0.75),
stats.combine.func = max,
# Permutation-based multiple testing correction
perm.no = 99,  strata = NULL,
# Reference-based multiple stage normalization
ref.pct = 0.5, stage.no = 6, excl.pct = 0.2,
# Family-wise error rate control
is.fwer = FALSE,
verbose = TRUE, return.feature.dat = FALSE)

which(zico.obj$p.adj.fdr <= 0.05)

# For proportion data
comm.p <- t(t(comm) / colSums(comm))
zico.obj <- ZicoSeq(meta.dat = meta.dat, feature.dat = comm.p,
grp.name = 'SmokingStatus', adj.name = 'Sex', feature.dat.type = "proportion",
# Filter to remove rare taxa
prev.filter = 0.2, mean.abund.filter = 0,  max.abund.filter = 0.002, min.prop = 0,
# Winsorization to replace outliers
is.winsor = TRUE, outlier.pct = 0.03, winsor.end = 'top',
# Posterior sampling will be automatically disabled
is.post.sample = FALSE, post.sample.no = 25,
# Use the square-root transformation
```

```
link.func = list(function (x) x^0.5), stats.combine.func = max,
# Permutation-based multiple testing correction
perm.no = 99,  strata = NULL,
# Reference-based multiple stage normalization
ref.pct = 0.5, stage.no = 6, excl.pct = 0.2,
# Family-wise error rate control
is.fwer = FALSE,
verbose = TRUE, return.feature.dat = FALSE)

which(zico.obj$p.adj.fdr <= 0.05)


# For other type of data. The user should be responsible for the filtering.
comm.o <- comm[rowMeans(comm != 0) >= 0.2, ] + 1
comm.o <- log(t(t(comm.o) / colSums(comm.o)))
zico.obj <- ZicoSeq(meta.dat = meta.dat, feature.dat = comm.o,
grp.name = 'SmokingStatus', adj.name = 'Sex', feature.dat.type = "other",
# Filter will not be applied
prev.filter = 0, mean.abund.filter = 0,  max.abund.filter = 0, min.prop = 0,
# Winsorization to both ends of the distribution
is.winsor = TRUE, outlier.pct = 0.03, winsor.end = 'both',
# Posterior sampling will be automatically disabled
is.post.sample = FALSE, post.sample.no = 25,
# Identity function is used
link.func = list(function (x) x), stats.combine.func = max,
# Permutation-based multiple testing correction
perm.no = 99,  strata = NULL,
# Reference-based multiple-stage normalization will not be performed
ref.pct = 0.5, stage.no = 6, excl.pct = 0.2,
# Family-wise error rate control
is.fwer = TRUE,
verbose = TRUE, return.feature.dat = FALSE)

which(zico.obj$p.adj.fdr <= 0.05)
```

Index