Package 'ANCOMBC'

July 3, 2022

Type Package

Title Analysis of compositions of microbiomes with bias correction

Version 1.6.2

Description ANCOMBC is a package containing differential abundance (DA) and correlation analyses for microbiome data. Specifically, the package includes Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) and Analysis of Composition of Microbiomes (ANCOM) for DA analysis, and Sparse Estimation of Correlations among Microbiomes (SECOM) for correlation analysis. Microbiome data are typically subject to two sources of biases: unequal sampling fractions (sample-specific biases) and differential sequencing efficiencies (taxon-specific biases). Methodologies included in the ANCOMBC package were designed to correct these biases and construct statistically consistent estimators.

Date 2022-06-24

License Artistic-2.0

Imports magrittr, Rdpack, rlang, microbiome, phyloseq, stats, DescTools, Hmisc, MASS, doParallel, doRNG, dplyr, energy, foreach, nlme, nloptr, parallel, tibble, tidyr

Suggests knitr, rmarkdown, testthat, DT, corrplot, ggforce, limma, qwraps2 (>= 0.5.0), tidyverse

biocViews DifferentialExpression, Microbiome, Normalization, Sequencing, Software

BugReports https://github.com/FrederickHuangLin/ANCOMBC/issues

URL https://github.com/FrederickHuangLin/ANCOMBC

VignetteBuilder knitr

RdMacros Rdpack

Encoding UTF-8

RoxygenNote 7.2.0

git_url https://git.bioconductor.org/packages/ANCOMBC

git_branch RELEASE_3_15

git_last_commit 543c77a

git_last_commit_date 2022-06-24

Date/Publication 2022-07-03

Author Huang Lin [cre, aut] (<https://orcid.org/0000-0002-4892-7871>)

Maintainer Huang Lin <huanglinfrederick@gmail.com>

R topics documented:

---

ancom *Analysis of Composition of Microbiomes (ANCOM)*

Description

Determine taxa whose absolute abundances, per unit volume, of the ecosystem (e.g. gut) are significantly different with changes in the covariate of interest (e.g. group). The current version of `ancom` function implements ANCOM in cross-sectional and longitudinal datasets while allowing for covariate adjustment.

Usage

```
ancom(
  phyloseq,
  p_adj_method = "holm",
  prv_cut = 0.1,
  lib_cut = 0,
  main_var,
  adj_formula = NULL,
  rand_formula = NULL,
  lme_control = NULL,
  struc_zero = FALSE,
  neg_lb = FALSE,
  alpha = 0.05,
  n_cl = 1
)
```

Arguments

phyloseq	a phyloseq-class object, which consists of a feature table (microbial observed abundance table), a sample metadata, a taxonomy table (optional), and a phylogenetic tree (optional). The row names of the metadata must match the sample names of the feature table, and the row names of the taxonomy table must match the taxon (feature) names of the feature table. See ?phyloseq::phyloseq for more details.
p_adj_method	character. method to adjust p-values. Default is "holm". Options include "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none". See ?stats::p.adjust for more details.
prv_cut	a numerical fraction between 0 and 1. Taxa with prevalences less than prv_cut will be excluded in the analysis. Default is 0.10.
lib_cut	a numerical threshold for filtering samples based on library sizes. Samples with library sizes less than lib_cut will be excluded in the analysis. Default is 0, i.e. do not discard any sample.
main_var	character. The name of the main variable of interest.
adj_formula	character string representing the formula for covariate adjustment. Default is NULL.
rand_formula	character string representing the formula for random effects. For details, see ?nlme::lme. Default is NULL.
lme_control	a list specifying control values for lme fit. For details, see ?nlme::lmeControl. Default is NULL.
struc_zero	logical. whether to detect structural zeros based on main_var. main_var should be discrete. Default is FALSE.
neg_lb	logical. whether to classify a taxon as a structural zero using its asymptotic lower bound. Default is FALSE.
alpha	numeric. level of significance. Default is 0.05.
n_cl	numeric. The number of nodes to be forked. For details, see ?parallel::makeCluster. Default is 1 (no parallel computing).

Details

The definition of structural zero can be found at ANCOM-II. Setting neg_lb = TRUE indicates that you are using both criteria stated in section 3.2 of ANCOM-II to detect structural zeros; otherwise, the algorithm will only use the equation 1 in section 3.2 for declaring structural zeros. Generally, it is recommended to set neg_lb = TRUE when the sample size per group is relatively large (e.g. > 30).

Value

a list with components:

- res, a data.frame containing ANCOM result for the variable specified in main_var, each column is:
  - W, test statistics.

  - detected_0.9, detected_0.8, detected_0.7, detected_0.6, logical vectors representing whether a taxon is differentially abundant under a series of cutoffs. For example, TRUE in detected_0.7 means the number of ALR transformed models where the taxon is differentially abundant with regard to the main variable outnumbers 0.7 * (n_taxa - 1). detected_0.7 is commonly used. Choose detected_0.8 or detected_0.9 for more conservative results, or choose detected_0.6 for more liberal results.

- zero_ind, a logical matrix with TRUE indicating the taxon is identified as a structural zero for the specified main variable.
- beta_data, a numeric matrix containing pairwise coefficients for the main variable of interest in ALR transformed regression models.
- p_data, a numeric matrix containing pairwise p-values for the main variable of interest in ALR transformed regression models.
- q_data, a numeric matrix containing adjusted p-values by applying the p_adj_method to the p_data matrix.

Author(s)

Huang Lin

References

Mandal S, Van Treuren W, White RA, Eggesbo M, Knight R, Peddada SD (2015). "Analysis of composition of microbiomes: a novel method for studying microbial composition." *Microbial ecology in health and disease*, **26**(1), 27663.

See Also

ancombc

Examples

```
library(microbiome)
library(tidyverse)
data(dietswap)

# Subset to baseline
pseq = subset_samples(dietswap, timepoint == 1)
# Aggregate to family level
family_data = aggregate_taxa(pseq, "Family")

# Run ancombc function
set.seed(123)
out = ancom(phyloseq = family_data,  p_adj_method = "holm",
            prv_cut = 0.10, lib_cut = 0, main_var = "nationality",
            adj_formula = "bmi_group",
            rand_formula = NULL, lme_control = NULL,
            struc_zero = TRUE, neg_lb = TRUE, alpha = 0.05, n_cl = 2)

res = out$res
```

ancombc *Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC)*

Description

Determine taxa whose absolute abundances, per unit volume, of the ecosystem (e.g. gut) are significantly different with changes in the covariate of interest (e.g. group). The current version of ancombc function implements Analysis of Compositions of Microbiomes with Bias Correction (ANCOM-BC) in cross-sectional data while allowing for covariate adjustment.

Usage

```
ancombc(
  phyloseq,
  formula,
  p_adj_method = "holm",
  prv_cut = 0.1,
  lib_cut = 0,
  group = NULL,
  struc_zero = FALSE,
  neg_lb = FALSE,
  tol = 1e-05,
  max_iter = 100,
  conserve = FALSE,
  alpha = 0.05,
  global = FALSE
)
```

Arguments

phyloseq	a phyloseq-class object, which consists of a feature table (microbial observed abundance table), a sample metadata, a taxonomy table (optional), and a phylogenetic tree (optional). The row names of the metadata must match the sample names of the feature table, and the row names of the taxonomy table must match the taxon (feature) names of the feature table. See ?phyloseq::phyloseq for more details.
formula	the character string expresses how the microbial absolute abundances for each taxon depend on the variables in metadata.
p_adj_method	character. method to adjust p-values. Default is "holm". Options include "holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none". See ?stats::p.adjust for more details.
prv_cut	a numerical fraction between 0 and 1. Taxa with prevalences less than prv_cut will be excluded in the analysis. Default is 0.10.

lib_cut	a numerical threshold for filtering samples based on library sizes. Samples with library sizes less than lib_cut will be excluded in the analysis. Default is 0, i.e. do not discard any sample.
group	character. the name of the group variable in metadata. group should be discrete. Specifying group is required for detecting structural zeros and performing global test.
struc_zero	logical. whether to detect structural zeros based on group. Default is FALSE.
neg_lb	logical. whether to classify a taxon as a structural zero using its asymptotic lower bound. Default is FALSE.
tol	numeric. the iteration convergence tolerance for the E-M algorithm. Default is 1e-05.
max_iter	numeric. the maximum number of iterations for the E-M algorithm. Default is 100.
conserve	logical. whether to use a conservative variance estimator for the test statistic. It is recommended if the sample size is small and/or the number of differentially abundant taxa is believed to be large. Default is FALSE.
alpha	numeric. level of significance. Default is 0.05.
global	logical. whether to perform global test. Default is FALSE.

Details

The definition of structural zero can be found at ANCOM-II. Setting neg_lb = TRUE indicates that you are using both criteria stated in section 3.2 of ANCOM-II to detect structural zeros; otherwise, the algorithm will only use the equation 1 in section 3.2 for declaring structural zeros. Generally, it is recommended to set neg_lb = TRUE when the sample size per group is relatively large (e.g. > 30).

Value

a list with components:

- feature_table, a data.frame of pre-processed (based on prv_cut and lib_cut) microbial observed abundance table.
- zero_ind, a logical matrix with TRUE indicating the taxon is identified as a structural zero for the specified group variable.
- samp_frac, a numeric vector of estimated sampling fractions in log scale (natural log). Note that for each sample, if it contains missing values for any variable specified in the formula, the corresponding sampling fraction estimate for this sample will be NA since the sampling fraction is not estimable with the presence of missing values.
- resid, a matrix of residuals from the ANCOM-BC log-linear (natural log) model. Rows are taxa and columns are samples.
- delta_em, estimated sample-specific biases through E-M algorithm.
- delta_wls, estimated sample-specific biases through weighted least squares (WLS) algorithm.
- res, a list containing ANCOM-BC primary result, which consists of:

  - lfc, a data.frame of log fold changes obtained from the ANCOM-BC log-linear (natural log) model.
  - se, a data.frame of standard errors (SEs) of lfc.
  - W, a data.frame of test statistics. W = lfc/se.
  - p_val, a data.frame of p-values. P-values are obtained from two-sided Z-test using the test statistic W.
  - q_val, a data.frame of adjusted p-values. Adjusted p-values are obtained by applying p_adj_method to p_val.
  - diff_abn, a logical data.frame. TRUE if the taxon has q_val less than alpha.
- res_global, a data.frame containing ANCOM-BC global test result for the variable specified in group, each column is:
  - W, test statistics.
  - p_val, p-values, which are obtained from two-sided Chi-square test using W.
  - q_val, adjusted p-values. Adjusted p-values are obtained by applying p_adj_method to p_val.
  - diff_abn, A logical vector. TRUE if the taxon has q_val less than alpha.

Author(s)

Huang Lin

References

Kaul A, Mandal S, Davidov O, Peddada SD (2017). "Analysis of microbiome data in the presence of excess zeros." *Frontiers in microbiology*, **8**, 2114.

Lin H, Peddada SD (2020). "Analysis of compositions of microbiomes with bias correction." *Nature communications*, **11**(1), 1–11.

See Also

ancom

Examples

```
#================Build a Phyloseq-Class Object from Scratch==================
library(phyloseq)

otu_mat = matrix(sample(1:100, 100, replace = TRUE), nrow = 10, ncol = 10)
rownames(otu_mat) = paste0("taxon", 1:nrow(otu_mat))
colnames(otu_mat) = paste0("sample", 1:ncol(otu_mat))


meta = data.frame(group = sample(LETTERS[1:4], size = 10, replace = TRUE),
                  row.names = paste0("sample", 1:ncol(otu_mat)),
                  stringsAsFactors = FALSE)

tax_mat = matrix(sample(letters, 70, replace = TRUE),
                 nrow = nrow(otu_mat), ncol = 7)
```

```
rownames(tax_mat) = rownames(otu_mat)
colnames(tax_mat) = c("Kingdom", "Phylum", "Class", "Order",
                      "Family", "Genus", "Species")

OTU = otu_table(otu_mat, taxa_are_rows = TRUE)
META = sample_data(meta)
TAX = tax_table(tax_mat)
physeq = phyloseq(OTU, META, TAX)

#======================Run ANCOMBC Using a Real Data======================

library(microbiome)
library(tidyverse)
data(dietswap)

# Subset to baseline
pseq = subset_samples(dietswap, timepoint == 1)
# Aggregate to family level
family_data = aggregate_taxa(pseq, "Family")

# Run ancombc function
out = ancombc(phyloseq = family_data, formula = "bmi_group + nationality",
              p_adj_method = "holm", prv_cut = 0.10, lib_cut = 1000,
              group = "bmi_group", struc_zero = TRUE, neg_lb = FALSE,
              tol = 1e-5, max_iter = 100, conserve = TRUE,
              alpha = 0.05, global = TRUE)

res = out$res
res_global = out$res_global
```

`secom_dist` *Sparse estimation of distance correlations among microbiomes*

Description

Obtain the sparse correlation matrix for distance correlations between taxa.

Usage

```
secom_dist(
  pseqs,
  pseudo = 0,
  prv_cut = 0.5,
  lib_cut = 1000,
  corr_cut = 0.5,
  wins_quant = c(0.05, 0.95),
  R = 1000,
  thresh_hard = 0,
```

```
  max_p = 0.005,
  n_cl = 1
)
```

Arguments

pseqs	a list of phyloseq-class objects. For one single ecosystem, specify it as pseqs = list(c(phyloseq1, phyloseq2)), where phyloseq1 (typically in low taxonomic levels, such as OTU or species level) is used to estimate biases, while phyloseq2 (can be in any taxonomic level) is used to compute the correlation matrix. For multiple ecosystems, simply stack the phyloseq objects. For example, for two ecosystems (such as gut and tongue), specify it as pseqs = list(gut = c(phyloseq1, phyloseq2),tongue = c(phyloseq3, phyloseq4)).
pseudo	numeric. Add pseudo-counts to the data. Default is 0 (no pseudo-counts).
prv_cut	a numerical fraction between 0 and 1. Taxa with prevalences less than prv_cut will be excluded in the analysis. Default is 0.5.
lib_cut	a numerical threshold for filtering samples based on library sizes. Samples with library sizes less than lib_cut will be excluded in the analysis. Default is 1000.
corr_cut	numeric. To prevent false positives due to taxa with small variances, taxa with Pearson correlation coefficients greater than corr_cut with the estimated sample-specific bias will be flagged. The pairwise correlation coefficient between flagged taxa will be set to 0s. Default is 0.5.
wins_quant	a numeric vector of probabilities with values between 0 and 1. Replace extreme values in the abundance data with less extreme values. Default is c(0.05, 0.95). For details, see ?DescTools::Winsorize.
R	numeric. The number of replicates in calculating the p-value for distance correlation. For details, see ?energy::dcor.test. Default is 1000.
thresh_hard	Numeric. Set a hard threshold for the correlation matrix. Pairwise distance correlation less than or equal to thresh_hard will be set to 0. Default is 0 (No ad-hoc hard thresholding).
max_p	numeric. Obtain the sparse correlation matrix by p-value filtering. Pairwise correlation coefficient with p-value greater than max_p will be set to 0. Default is 0.005.
n_cl	numeric. The number of nodes to be forked. For details, see ?parallel::makeCluster. Default is 1 (no parallel computing).

Details

The distance correlation, which is a measure of dependence between two random variables, can be used to quantify any dependence, whether linear, monotonic, non-monotonic or nonlinear relationships.

Value

a list with components:

- s_diff_hat, a numeric vector of estimated sample-specific biases.

- y_hat, a matrix of bias-corrected abundances
- mat_cooccur, a matrix of taxon-taxon co-occurrence pattern. The number in each cell represents the number of complete (nonzero) samples for the corresponding pair of taxa.
- dcorr, the sample distance correlation matrix computed using the bias-corrected abundances y_hat.
- dcorr_p, the p-value matrix corresponding to the sample distance correlation matrix dcorr.
- dcorr_fl, the sparse correlation matrix obtained by p-value filtering based on the cutoff specified in max_p.

Author(s)

Huang Lin

See Also

secom_linear

Examples

```
library(microbiome)
library(tidyverse)
data(dietswap)

# Subset to baseline
pseq = subset_samples(dietswap, timepoint == 1)
# Genus level data
phyloseq1 = pseq
# Phylum level data
phyloseq2 = aggregate_taxa(pseq, level = "Phylum")

# print(phyloseq1)
# print(phyloseq2)

set.seed(123)
res_dist = secom_dist(pseqs = list(c(phyloseq1, phyloseq2)), pseudo = 0,
                      prv_cut = 0.5, lib_cut = 1000, corr_cut = 0.5,
                      wins_quant = c(0.05, 0.95), R = 1000,
                      thresh_hard = 0.3, max_p = 0.005, n_cl = 1)

dcorr_fl = res_dist$dcorr_fl
```

---

secom_linear *Sparse estimation of linear correlations among microbiomes*

Description

Obtain the sparse correlation matrix for linear correlations between taxa. The current version of `secom_linear` function supports either of the three correlation coefficients: Pearson, Spearman, and Kendall's τ.

Usage

```
secom_linear(
  pseqs,
  pseudo = 0,
  prv_cut = 0.5,
  lib_cut = 1000,
  corr_cut = 0.5,
  wins_quant = c(0.05, 0.95),
  method = c("pearson", "kendall", "spearman"),
  soft = FALSE,
  thresh_len = 100,
  n_cv = 10,
  thresh_hard = 0,
  max_p = 0.005,
  n_cl = 1
)
```

Arguments

`pseqs`	a list of phyloseq-class objects. For one single ecosystem, specify it as `pseqs = list(c(phyloseq1, phyloseq2))`, where `phyloseq1` (typically in low taxonomic levels, such as OTU or species level) is used to estimate biases, while `phyloseq2` (can be in any taxonomic level) is used to compute the correlation matrix. For multiple ecosystems, simply stack the phyloseq objects. For example, for two ecosystems (such as gut and tongue), specify it as `pseqs = list(gut = c(phyloseq1, phyloseq2),tongue = c(phyloseq3, phyloseq4))`.
`pseudo`	numeric. Add pseudo-counts to the data. Default is 0 (no pseudo-counts).
`prv_cut`	a numerical fraction between 0 and 1. Taxa with prevalences less than `prv_cut` will be excluded in the analysis. Default is 0.5.
`lib_cut`	a numerical threshold for filtering samples based on library sizes. Samples with library sizes less than `lib_cut` will be excluded in the analysis. Default is 1000.
`corr_cut`	numeric. To prevent false positives due to taxa with small variances, taxa with Pearson correlation coefficients greater than `corr_cut` with the estimated sample-specific bias will be flagged. The pairwise correlation coefficient between flagged taxa will be set to 0s. Default is 0.5.
`wins_quant`	a numeric vector of probabilities with values between 0 and 1. Replace extreme values in the abundance data with less extreme values. Default is `c(0.05, 0.95)`. For details, see `?DescTools::Winsorize`.
`method`	character. It indicates which correlation coefficient is to be computed. One of "pearson", "kendall", or "spearman": can be abbreviated.

soft	logical. TRUE indicates that soft thresholding is applied to achieve the sparsity of the correlation matrix. FALSE indicates that hard thresholding is applied to achieve the sparsity of the correlation matrix. Default is FALSE.
thresh_len	numeric. Grid-search is implemented to find the optimal values over thresh_len thresholds for the thresholding operator. Default is 100.
n_cv	numeric. The fold number in cross validation. Default is 10 (10-fold cross validation).
thresh_hard	Numeric. Set a hard threshold for the correlation matrix. Pairwise correlation coefficient (in its absolute value) less than or equal to thresh_hard will be set to 0. Default is 0 (No ad-hoc hard thresholding).
max_p	numeric. Obtain the sparse correlation matrix by p-value filtering. Pairwise correlation coefficient with p-value greater than max_p will be set to 0. Default is 0.005.
n_cl	numeric. The number of nodes to be forked. For details, see ?parallel::makeCluster. Default is 1 (no parallel computing).

Value

a list with components:

- s_diff_hat, a numeric vector of estimated sample-specific biases.
- y_hat, a matrix of bias-corrected abundances
- cv_error, a numeric vector of cross-validation error estimates, which are the Frobenius norm differences between correlation matrices using training set and validation set, respectively.
- thresh_grid, a numeric vector of thresholds in the cross-validation.
- thresh_opt, numeric. The optimal threshold through cross-validation.
- mat_cooccur, a matrix of taxon-taxon co-occurrence pattern. The number in each cell represents the number of complete (nonzero) samples for the corresponding pair of taxa.
- corr, the sample correlation matrix (using the measure specified in method) computed using the bias-corrected abundances y_hat.
- corr_p, the p-value matrix corresponding to the sample correlation matrix corr.
- corr_th, the sparse correlation matrix obtained by thresholding based on the method specified in soft.
- corr_fl, the sparse correlation matrix obtained by p-value filtering based on the cutoff specified in max_p.

Author(s)

Huang Lin

See Also

secom_dist

Examples

```
library(microbiome)
library(tidyverse)
data("dietswap")

# Subset to baseline
pseq = subset_samples(dietswap, timepoint == 1)
# Genus level data
phyloseq1 = pseq
# Phylum level data
phyloseq2 = aggregate_taxa(pseq, level = "Phylum")

# print(phyloseq1)
# print(phyloseq2)

set.seed(123)
res_linear = secom_linear(pseqs = list(c(phyloseq1, phyloseq2)), pseudo = 0,
                          prv_cut = 0.5, lib_cut = 1000, corr_cut = 0.5,
                          wins_quant = c(0.05, 0.95), method = "pearson",
                          soft = FALSE, thresh_len = 20, n_cv = 10,
                          thresh_hard = 0.3, max_p = 0.005, n_cl = 1)

corr_th = res_linear$corr_th
corr_fl = res_linear$corr_fl
```

Index