Package 'CBEA'

October 18, 2022

Title Competitive Balances for Taxonomic Enrichment Analysis in R

Version 1.0.0

Date 2022-03-03

Description This package implements CBEA, a method to perform set-based analysis for microbiome relative abundance data. This approach constructs a competitive balance between taxa within the set and remainder taxa per sample. More details can be found in the Nguyen et al. 2021+ manuscript. Additionally, this package adds support functions to help users perform taxa-set enrichment analyses using existing gene set analysis methods. In the future we hope to also provide curated knowledge driven taxa sets.

License MIT + file LICENSE

URL https://github.com/qpmnguyen/CBEA, https://qpmnguyen.github.io/CBEA/

BugReports https://github.com/qpmnguyen/CBEA//issues

Depends R (>= 4.2.0)

Imports BiocParallel, BiocSet, dplyr, lmom, fitdistrplus, magrittr, methods, mixtools, Rcpp (>= 1.0.7), stats, SummarizedExperiment, tibble, TreeSummarizedExperiment, tidyr, glue, generics, rlang, goftest

Suggests phyloseq, BiocStyle, covr, knitr, RefManageR, rmarkdown, sessioninfo, testthat (>= 3.0.0), tidyverse, roxygen2, mia, purrr

LinkingTo Rcpp

VignetteBuilder knitr

biocViews Software, Microbiome, Metagenomics, GeneSetEnrichment, DataImport

Config/testthat/edition 3

Encoding UTF-8

LazyData false

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

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

git_branch RELEASE_3_15

git_last_commit 20d5b52

git_last_commit_date 2022-04-26

Date/Publication 2022-10-18

Author Quang Nguyen [aut, cre] (<https://orcid.org/0000-0002-2072-3279>)

Maintainer Quang Nguyen <quangpmnguyen@gmail.com>

R topics documented:

cbea 2
get_raw_score 5
glance.CBEAout 6
gmean 7
gmeanRow 7
hmp_gingival 8
new_CBEAout 9
pmnorm 9
print.CBEAout 10
tidy.CBEAout 11

Index **12**

`cbea` *Enrichment analysis using competitive compositional balances (CBEA)*

Description

`cbea` is used compute enrichment scores per sample for pre-defined sets using the CBEA (Competitive Balances for Enrichment Analysis).

Usage

```
cbea(
  obj,
  set,
  output,
  distr = NULL,
  adj = FALSE,
  n_perm = 100,
  parametric = TRUE,
  thresh = 0.05,
```

```
  init = NULL,
  control = NULL,
  parallel_backend = NULL,
  ...
)

## S4 method for signature 'TreeSummarizedExperiment'
cbea(
  obj,
  set,
  output,
  distr = NULL,
  abund_values,
  adj = FALSE,
  n_perm = 100,
  parametric = TRUE,
  thresh = 0.05,
  init = NULL,
  control = NULL,
  parallel_backend = NULL,
  ...
)

## S4 method for signature 'data.frame'
cbea(
  obj,
  set,
  taxa_are_rows = FALSE,
  id_col = NULL,
  output,
  distr = NULL,
  adj = FALSE,
  n_perm = 100,
  parametric = TRUE,
  thresh = 0.05,
  init = NULL,
  control = NULL,
  parallel_backend = NULL,
  ...
)

## S4 method for signature 'matrix'
cbea(
  obj,
  set,
  taxa_are_rows = FALSE,
  output,
  distr = NULL,
```

```
  adj = FALSE,
  n_perm = 100,
  parametric = TRUE,
  thresh = 0.05,
  init = NULL,
  control = NULL,
  parallel_backend = NULL,
  ...
)
```

Arguments

obj	The element of class TreeSummarizedExperiment, data.frame, or matrix. phyloseq is not supported due to conflicting dependencies and TreeSummarizedExperiment is much more compact.
set	BiocSet. Sets to be tested for enrichment in the BiocSet format. Taxa names must be in the same format as elements in the set.
output	(String). The form of the output of the model. Has to be either zscore, cdf, raw, pval, or sig
distr	(String). The choice of distribution for the null. Can be either mnorm (2 component mixture normal), norm (Normal distribution), or NULL if parametric is TRUE.
adj	(Logical). Whether correlation adjustment procedure is utilized. Defaults to FALSE.
n_perm	(Numeric). Add bootstrap resamples to both the permuted and unpermuted data set. This might help with stabilizing the distribution fitting procedure, especially if the sample size is low. Defaults to 1.
parametric	(Logical). Indicate whether a parametric distribution will be fitted to estimate z-scores, CDF values, and p-values. Defaults to TRUE
thresh	(Numeric). Threshold for significant p-values if sig is the output. Defaults to 0.05
init	(Named List). Initialization parameters for estimating the null distribution. Default is NULL.
control	(Named List). Additional arguments to be passed to fitdistr and normmixEM. Defaults to NULL.
parallel_backend	See documentation cbea
...	Additional arguments not used at the moment.
abund_values	(Character). Character value for selecting the assay to be the input to cbea
taxa_are_rows	(Logical). Indicate whether the data frame or matrix has taxa as rows
id_col	(Character Vector). Vector of character to indicate metadata columns to keep (for example, sample_id)

Details

This function support different formats of the OTU table, however for best results please use TreeSummarizedExperiment. phyloseq is supported, however CBEA will not explicitly import phyloseq package and will require users to install them separately. If use data.frame or matrix, users should specify whether taxa are rows using the taxa_are_rows option. Additionally, for data.frame, users can specify metadata columns to be kept via the id_col argument.
The output argument specifies what type of values will be returned in the final matrix. The options pval or sig returns either unadjusted p-values or dummy variables indicating whether a set is significantly enriched in that sample (based on unadjusted p-values thresholded at thresh). The option raw returns raw scores computed for each set without any distribution fitting or inference procedure. Users can use this option to examine the distribution of CBEA scores under the null.

Value

R An n by m matrix of enrichment scores at the sample level

Examples

```
data(hmp_gingival)
seq <- hmp_gingival$data
set <- hmp_gingival$set
# n_perm = 10 to reduce runtime
mod <- cbea(obj = seq, set = set, output = "zscore",
    abund_values = "16SrRNA",
    distr = "norm", parametric = TRUE,
    adj = TRUE, thresh = 0.05, n_perm = 10)
```

---

get_raw_score *Get CBEA scores for a given matrix and a vector of column indices*

Description

Get CBEA scores for a given matrix and a vector of column indices

Usage

```
get_raw_score(X, idx)
```

Arguments

X	(Matrix). OTU table of matrix format where taxa are columns and samples are rows
idx	(Integer vector). Vector of integers indicating the column ids of taxa in a set

Value

A matrix of size n by 1 where n is the total number of samples

Examples

```
data(hmp_gingival)
seq <- hmp_gingival$data
seq_matrix <- SummarizedExperiment::assays(seq)[[1]]
seq_matrix <- t(seq_matrix) + 1
rand_set <- sample(seq_len(ncol(seq_matrix)), size = 10)
scores <- get_raw_score(X = seq_matrix, idx = rand_set)
```

`glance.CBEAout` *Glance at* `CBEAout` *object*

Description

This function cleans up all diagnostics of the cbea method (from the `CBEAout` object) into a nice `tibble::tibble()`

Usage

```
## S3 method for class 'CBEAout'
glance(x, statistic, ...)
```

Arguments

`x`	An object of type `CBEAout`
`statistic`	What type of diagnostic to return. Users can choose to return `fit_diagnostic` which returns goodness of fit statistics for the different fitted distributions (e.g. log likelihoods) while `fit_comparison` returns comparisons across different distributions and raw values (and data) across the 4 l-moments.
`...`	Unused, kept for consistency with generics

Value

A `tibble::tibble()` summarizing diagnostic fits per set (as row)

Examples

```
# load the data
data(hmp_gingival)
mod <- cbea(hmp_gingival$data, hmp_gingival$set, abund_values = "16SrRNA",
    output = "sig", distr = "norm", adj = FALSE, n_perm = 5, parametric = TRUE)
glance(mod, "fit_diagnostic")
```

gmean *Geometric mean of a vector*

Description

Compute geometric mean of a vector using `exp(mean(log(.x)))` format

Usage

```
gmean(vec)
```

Arguments

`vec`	A vector of values with length `n`

Value

A numeric value of the geometric mean of the vector `vec`

Examples

```
ex <- abs(rnorm(10))
gmean(ex)
```

gmeanRow *Geometric mean of rows of a matrix*

Description

This function computes the geometric mean by row of a numeric matrix

Usage

```
gmeanRow(X)
```

Arguments

`X`	A numeric matrix with `n` rows and `p` columns

Value

A numeric vector of the geometric mean of the matrix `X` with length `n`

Examples

```
ex <- matrix(rnorm(100), nrow = 10, ncol = 10)
ex <- abs(ex)
gmeanRow(ex)
```

`hmp_gingival` *Gingival data set from the Human Microbiome Project*

Description

Gingival data set from the Human Microbiome Project

Usage

```
data(hmp_gingival)
```

Format

A list with two elements

data The microbiome relative abundance data with relevant metadata obtained from the Human Microbiome Project via the `HMP16SData` package (snapshot: 11-15-2021). The data set is hosted the container of type `phyloseq`. Using the `mia` package users can convert it to the `TreeSummarizedExperiment` type.

set Sets of microbes based on their metabolism annotation at the Genera level. Annotations obtained via Calagaro et al.'s repository on Zenodo (https://doi.org/10.5281/zenodo.3942108)

References

Data can be downloaded directly from https://hmpdacc.org/hmp/

R interface of the data from https://doi.org/doi:10.18129/B9.bioc.HMP16SData

Beghini F, Renson A, Zolnik CP, Geistlinger L, Usyk M, Moody TU, et al. Tobacco Exposure Associated with Oral Microbiota Oxygen Utilization in the New York City Health and Nutrition Examination Study. Annals of Epidemiology. 2019;34:18–25.e3. doi:10.1016/j.annepidem.2019.03.005

Consortium THMP, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, et al. Structure, Function and Diversity of the Healthy Human Microbiome. Nature. 2012;486(7402):207–214. doi:10.1038/nature11234.

Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of Statistical Methods from Single Cell, Bulk RNA-Seq, and Metagenomics Applied to Microbiome Data. Genome Biology. 2020;21(1):191. doi:10.1186/s13059-020-02104-1

Schiffer L, Azhar R, Shepherd L, Ramos M, Geistlinger L, Huttenhower C, et al. HMP16SData: Efficient Access to the Human Microbiome Project through Bioconductor. American Journal of Epidemiology. 2019;doi:10.1093/aje/kwz006.

new_CBEAout *Creating an output object of type CBEAout*

Description

This function takes a list of lists from each object and turns it into a CBEAout type object

Usage

```
new_CBEAout(out, call)
```

Arguments

out	A list containing scores for each set
call	A list containing all important arguments for printing

Value

A new CBEAout object (which is a cleaner list of lists)

pmnorm *The Two Component Mixture Normal Distribution*

Description

The Two Component Mixture Normal Distribution

Usage

```
pmnorm(q, mu, sigma, lambda, log = FALSE, verbose = FALSE)

dmnorm(x, mu, sigma, lambda, log = FALSE, verbose = FALSE)
```

Arguments

q, x	(Vector). Values to calculate distributional values of.
mu	(Vector). A two value vector of mean values.
sigma	(Vector). A two value vector of component-wise variances
lambda	(Vector). A two value vector of component mixing coefficients
log	(Boolean). Whether returning probabilities are in log format
verbose	(Boolean). Whether to return component values.

Value

A numeric value representing the probability density value of a two-component mixture distribution

Functions

- pmnorm: Cumulative Distribution Function
- dmnorm: Probability Density Function

Examples

```
library(mixtools)
lambda <- c(0.7,0.3)
mu <- c(1,2)
sigma <- c(1,1)
v <- rnormmix(100, lambda=lambda, mu=mu, sigma=sigma)
pmnorm(v, lambda=lambda,mu=mu,sigma=sigma)
dmnorm(v, lambda=lambda,mu=mu,sigma=sigma)
```

---

print.CBEAout *Print dispatch for CBEAout objects*

Description

Print dispatch for CBEAout objects

Usage

```
## S3 method for class 'CBEAout'
print(x, ...)
```

Arguments

x	The CBEAout object
...	Undefined arguments, keeping consistency for generics

Value

Text for printing

`tidy.CBEAout` *Tidy a CBEAout object*

Description

This function takes in a CBEA type object and collects all values across all sets and samples that were evaluated.

Usage

```
## S3 method for class 'CBEAout'
tidy(x, ...)
```

Arguments

`x`	A `CBEAout` object.
`...`	Unused, included for generic consistency only.

Value

A tidy `tibble::tibble()` summarizing scores per sample per set.

Examples

```
# load the data
data(hmp_gingival)
mod <- cbea(hmp_gingival$data, hmp_gingival$set, abund_values = "16SrRNA",
    output = "sig", distr = "norm", adj = FALSE, n_perm = 5, parametric = TRUE)
tidy(mod)
```

Index

∗ **datasets**
hmp_gingival, 8

cbea, 2, *4*
cbea,data.frame-method *(*cbea*)*, 2
cbea,matrix-method *(*cbea*)*, 2
cbea,TreeSummarizedExperiment-method
*(*cbea*)*, 2

dmnorm *(*pmnorm*)*, 9

get_raw_score, 5
glance.CBEAout, 6
gmean, 7
gmeanRow, 7

hmp_gingival, 8

new_CBEAout, 9

pmnorm, 9
print.CBEAout, 10

tibble::tibble(), *6*, *11*
tidy.CBEAout, 11
TreeSummarizedExperiment, *5*