Package 'DiversitySeq'

September 2, 2018

Type Package

Title DiversitySeq: measuring diversity from count data sets

Version 1.0

Date 2016-10-18

Author Francesca Finotello, Eleonora Mastrorilli, Barbara Di Camillo

Maintainer Francesca Finotello <francesca.finotello@dei.unipd.it>

Description DiversitySeq is a package for the analysis of diversity from count data and for the simulation of 16S ribosomal RNA (16S rRNA) gene sequencing data sets.

License GPL-3

Depends vegan

R topics documented:

`aindex` *Compute alpha diversity from count data*

Description

Computes alpha diversity from a matrix of species counts.

Usage

```
aindex(countdata, index = c("Hill", "Renyi", "BergerParker",
    "Richness", "iSimpson", "cSimpson", "Shannon",
    "Chao1", "ACE", "Jackknife1", "Jackknife2",
    "Pielou","Tail", "EF", "IF", "RLE", "RLI"),
     q = NULL, keep0 = FALSE, scalemin = FALSE,
     group = NULL)
```

Arguments

countdata	a matrix of species counts, with species on the rows and samples on the columns (more generally, it can be a matrix of counts computed for a set of non-overlapping classes)
index	the index to be used for the computation of alpha diversity (see the package vignette for further details)
q	a number indicating the order of diversity, mandatory for Hill, Renyi, EF, IF, RLE and RLI indices.
keep0	a logical value TRUE/FALSE indicating whether species with null counts should be considered in the computation of alpha diversity
scalemin	a logical value indicating whether count data shold be scaled so to have the minimum equal to 1 (useful in case of normalized data)
group	a vector of strings indicating to which group the samples in 'countdata' belong. The length of 'group' vector must equal the number of columns of 'countdata'. When the parameter 'group' is not specified, all samples are assigned to the same group, called 'group1'

Value

List whose length equals the number of different groups. For each group, the list contains the alpha diversity values computed for all samples in 'countdata' which belong to the group.

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and data set
library(DiversitySeq)
data(salivaSimData)

# Assign samples to 2 groups
group <- c(rep("group1", ncol(simCounts)/2),
rep("group2", ncol(simCounts)/2))

# Compute alpha diversity with Hill numbers of order 2
adiv <- aindex(simCounts, index = "Hill", q=2, group = group)

# Plot diversity
divplot(adiv, col="default")
```

bindex *Compute beta diversity from count data*

Description

Computes beta diversity from a matrix of species counts.

Usage

```
bindex(countdata, index = c("w", "c", "r", "I",
    "e", "m", "mn", "-2","co", "cc", "-3",
    "-3n", "rs", "sim", "z"), group = NULL)
```

Arguments

countdata	a matrix of species counts, with species on the rows and samples on the columns (more generally, it can be a matrix of counts computed for a set of non-overlapping classes)
index	the index to be used for the computation of beta diversity (see the package vignette for further details)
group	vector of strings indicating to which group the samples in 'countdata' belong. The length of 'group' vector must equal the number of columns of 'countdata'. When the parameter 'group' is not specified, all samples are assigned to the same group, called 'group1'

Value

List whose length equals the number of different groups. For each group, the list contains the beta diversity values computed for all pairs of samples in 'countdata' which belong to the group.

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and data set
library(DiversitySeq)
data(salivaSimData)

# Assign samples to 2 groups
group <- c(rep("group1", ncol(simCounts)/2),
rep("group2", ncol(simCounts)/2))

# Compute beta diversity with Whittaker index
bdiv <- bindex(simCounts, index = "w", group = group)

# Plot diversity
divplot(bdiv, col="default")
```

`DivesrsitySeq` *DiversitySeq: measuring diversity from count data sets*

Description

Next-generation sequencing, and particularly 16S ribosomal RNA (16S rRNA) gene sequencing, is a powerful technique for the identification and quantification of human-resident microbes, collectively known as the human microbiota.

Once bacterial abundances are profiled via 16S rRNA gene sequencing and summarized in a count data set, diversity indices provide valuable mathematical tools to investigate the composition of the human microbiota. In brief, alpha diversity can be used to describe the compositional complexity of a single sample, whereas beta diversity can be used to identify taxonomical differences between samples.

The DiversitySeq package implements in a unified framework the whole panel of diversity indices reviewed in Finotello et al. (2016), enabling the assessment of diversity from count data sets. DiversitySeq also implements a simulator for the generation of synthetic count data sets from 16S rRNA gene sequencing.

Besides 16S rRNA gene sequencing data, this package can be employed with other data sets with similar characteristics, such as 5S rRNA gene sequencing, environmental metagenomics or, more generally, any kind of matrix were counts are computed for different types non-overlapping classes.

Details

Package:	DiversitySeq
Type:	Package
Version:	1.0
Date:	2016-01-09
License:	GPL-3

Author(s)

Francesca Finotello, Eleonora Mastrorilli, Barbara Di Camillo

Maintainer: Francesca Finotello <francesca.finotello@dei.unipd.it>

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and data
library(DiversitySeq)
data(salivaSimData)

# Compute alpha diversity
```

```
alphadiv <- aindex(simCounts, index = "Richness")
divplot(alphadiv)

# Compute beta diversity
betadiv <- bindex(simCounts, index = "r")

# Plot beta diversity
divplot(betadiv)

# Simulate new count data
newsimdata <- simulatecounts(avgAbund = avgAbundances, phi, sdepth)
```

`divplot` *Plot diversity computed with the DiversitySeq package*

Description

Plots the boxplots of alpha and beta diversity computed with the DiversitySeq package.

Usage

```
divplot(diversity, main="", ylab = "Diversity", las = 1,
    col = NULL, points = FALSE, pointcol = "black", cexpoints = 1)
```

Arguments

`diversity`	a list of alpha or beta diversity values computed with 'adiv' or 'bdiv' functions, respectively
`main`	the title of the plot
`ylab`	the label of the y-axis
`las`	numeric in [0, 1, 2, 3] indicating the style of axis labels; for further details, check 'las' help(par)
`col`	colors to be used for the boxplots. When set to NULL, the color is set to 'white'. When set to "default", a default palette is used (up to 4 groups). Alternatively, a vector of valid color can be provided (see the examples below).
`points`	a logical value indicating whether single diversity values should be shown as scatter plot over each boxplot
`pointcol`	the color of the points in the scatter plot, when 'points = TRUE'
`cexpoints`	a numerical value giving the amount by which the points should be magnified relative to the default of 1, when 'points = TRUE'; for further details, check 'cex' in help(par)

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and data set
library(DiversitySeq)
data(salivaSimData)

# Assign samples to 2 groups
group <- c(rep("group1", ncol(simCounts)/2), rep("group2", ncol(simCounts)/2))

# Compute alpha diversity with Hill numbers of order 2
adiv <- aindex(simCounts, index = "Hill", q=2, group = group)

# Plot diversity
divplot(adiv)
divplot(adiv, points = TRUE, cexpoints = 0.8)
divplot(adiv, col = "default", las = 2)
divplot(adiv, col = c("#4682B4","#FFD700"),
        main = "Hill diversity (q=2)",
        ylab = "Number equivalents")
```

`mergedatasets` *Merge count data sets*

Description

Merge count data sets.

Usage

```
mergedatasets(datasets, groups)
```

Arguments

`datasets`	a list containing the count matrices to be merged. Each count matrices must be a matrix of species counts, with species on the rows and samples on the columns (or, more generally, a matrix of counts conputed for a set of non-overlapping classes)
`groups`	a list containing the group annotation for each data set (in the same order as in the list above). Each 'group' vector is a vector of strings indicating to which group the samples in countdata belong. The length of 'group' vector must equal the number of columns of the corresponding count matrix

Value

List containing the merged matrix and the the corresponding vector of groups. The number of rows in the output matrix is equal to the union of all species assayed in the input data sets, and the number of columns is the sum of the samples of all input data sets.

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and data sets
library(DiversitySeq)
data(salivaSimData)
saliva.counts <- simCounts
data(stoolSimData)
stool.counts <- simCounts

# Generate 'group' vectors for the two data sets
saliva.group <- rep("Saliva", ncol(saliva.counts))
stool.group <- rep("Stool", ncol(stool.counts))

# Merge the data
mrgData <- mergedatasets(list(stool.counts, saliva.counts),
                         list(stool.group, saliva.group))
mrg.counts <- mrgData$data
mrg.group <- mrgData$group
```

`salivaSimData` *Simulated 16S rRNA gene sequencing data from saliva samples*

Description

Simulated 16S rRNA gene sequencing data from saliva samples generated with a negative binomial (NB) model (see References for further details on the simulation).

Usage

```
salivaSimData
```

Format

The data set contains the following objects:

simAbund: a matrix of simulated species abundances over 15,094 species (rows) and 20 samples (columns)

simCounts: a matrix of simulated counts over 15,094 species (rows) and 20 samples (columns), resulting from the (simulated) sequencing of 'simAbund'

avgAbundances: a numeric vector of average species abundances used for the simulation

phi: the coefficient of dispersion of the count data used for the simulation

sdepth: a vector of sequencing depths used for the simulation

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
library(DiversitySeq)
data(salivaSimData)
```

simulatecounts *Simulate 16S rRNA gene sequencing data*

Description

Simulate a count matrix from 16S rRNA gene sequencing.

Usage

```
simulatecounts(avgAbund, phi, sdepth)
```

Arguments

avgAbund	a numeric vector of average abundances over N species
phi	the coefficient of dispersion of the count data to be simulated
sdepth	a vector of sequencing depths for the M samples to be simulated

Value

List containing the N x M the matrix of species abundances and the N x M matrix of counts.

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
# Load package and simulation parameters
library(DiversitySeq)
data(salivaSimData)

# Simulate a new data set
newsimdata <- simulatecounts(avgAbund = avgAbundances, phi, sdepth)
newcounts<-newsimdata$counts
newabundances<-newsimdata$abundances
```

stoolSimData *Simulated 16S rRNA gene sequencing data from stool samples*

Description

Simulated 16S rRNA gene sequencing data from stool samples generated with a negative binomial (NB) model (see References for further details on the simulation).

Usage

```
stoolSimData
```

Format

The data set contains the following objects:

simAbund: a matrix of simulated species abundances over 8,048 species (rows) and 20 samples (columns)

simCounts: a matrix of simulated counts over 8,048 species (rows) and 20 samples (columns), resulting from the (simulated) sequencing of 'simAbund'

avgAbundances: a numeric vector of average species abundances used for the simulation

phi: the coefficient of dispersion of the count data used for the simulation

sdepth: a vector of sequencing depths used for the simulation

References

Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in Bioinformatics 19 (4), 679-692, 2018.

Examples

```
library(DiversitySeq)
data(salivaSimData)
```

Index