Package ‘MIDASim’

July 5, 2025

Type Package

Title Simulating Realistic Microbiome Data using 'MIDASim'

Version 2.0

Description The 'MIDASim' package is a microbiome data simulator for generating realistic microbiome datasets by adapting a user-provided template. It supports the controlled introduction of experimental signals-such as shifts in taxon relative abundances, prevalence, and sample library sizes-to create distinct synthetic populations under diverse simulation scenarios. For more details, see He et al. (2024) <doi:10.1186/s40168-024-01822-z>.

Imports psych, MASS, pracma, scam, stats

Suggests vegan, phyloseq, rmarkdown, knitr, roxygen2, testthat

URL `https://github.com/mengyu-he/MIDASim`

BugReports `https://github.com/mengyu-he/MIDASim/issues`

LazyData true

License GPL-2

Encoding UTF-8

Depends R (>= 3.5.0)

VignetteBuilder knitr

RoxygenNote 7.3.1

Config/testthat/edition 3

NeedsCompilation no

Author Mengyu He [aut, cre]

Maintainer Mengyu He <`mhe44@emory.edu`>

Repository CRAN

Date/Publication 2025-07-05 19:00:03 UTC

Contents

`count.ibd` *IBD microbiome dataset*

Description

A filtered microbiome dataset of patients with IBD(Inflammatory Bowel Disease) in Human Microbiome Project 2 (HMP2).

Usage

```
data(count.ibd)
```

Format

An object of class `matrix` (inherits from `array`) with 146 rows and 614 columns.

References

Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662 (2019). https://doi.org/10.1038/s41586-019-1237-9.

Examples

```
data(count.ibd)

MIDASim.setup(otu.tab = count.ibd, mode = "nonparametric")
```

`count.vaginal` *MOMS-PI microbiome dataset*

Description

A filtered microbiome dataset of Multi-Omic Microbiome Study-Pregnancy Initiative (MOMS-PI) in Human Microbiome Project 2 (HMP2).

Usage

```
data(count.vaginal)
```

Format

An object of class `matrix` (inherits from `array`) with 517 rows and 1146 columns.

References

Fettweis, J.M., Serrano, M.G., Brooks, J.P. et al. The vaginal microbiome and preterm birth. Nat Med 25, 1012–1021 (2019). https://doi.org/10.1038/s41591-019-0450-2

Examples

```
data(count.vaginal)

MIDASim.setup(otu.tab = count.vaginal, mode = "nonparametric")
```

`MIDASim` *Simulating Realistic Microbiome Data using MIDASim*

Description

Generate microbiome datasets using parameters from MIDASim.modify.

Usage

```
MIDASim(fitted.modified, only.rel = FALSE)
```

Arguments

`fitted.modified`	Output from MIDASim.modify.
`only.rel`	A logical indicating whether to only simulate relative- abundance data. If `TRUE`, then the count data will not be generated. Defaults to `FALSE`.

Value

Returns a list that has components:

`sim_01`	Matrix of simulated presence-absence data
`sim_rel`	Matrix of simulated relative-abundance data
`sim_count`	Matrix of simulated count data

Author(s)

Mengyu He

Examples

```
data("throat.otu.tab")
otu.tab = throat.otu.tab[,colSums(throat.otu.tab>0)>1]

fitted = MIDASim.setup(otu.tab)
fitted.modified = MIDASim.modify(fitted)
sim = MIDASim(fitted.modified, only.rel = FALSE)
```

`MIDASim.modify` *Modifying MIDASim model*

Description

MIDASim.modify() modifies the fitted MIDASim.setup model according to user specification that one or multiple of the following characteristics, such as the library sizes, taxa relative abundances, location parameters of the parametric model can be changed. This is useful if the users wants to introduce an 'effect' in simulation studies.

Usage

```
MIDASim.modify(
  fitted,
  lib.size = NULL,
  mean.rel.abund = NULL,
  gengamma.mu = NULL,
  sample.1.prop = NULL,
  taxa.1.prop = NULL,
  individual.rel.abund = NULL,
  ...
)
```

Arguments

`fitted`	Output from MIDASim.setup.
`lib.size`	Numeric vector of pre-specified library sizes (length should be equal to `n.sample` if specified). In nonparametric mode, if `lib.size` is specified, both `taxa.1.prop` and `sample.1.prop` should be specified.
`mean.rel.abund`	Numeric vector of specified mean relative abundances for taxa. Length should be equal to `n.taxa` in `fitted`.
`gengamma.mu`	Numeric vector of specified location parameters for the parametric model (generalized gamma model). Specify either `mean.rel.abund` or `gengamma.mu`, not both. Length should be equal to `n.taxa` in `fitted`. See Details. This argument is only applicable in parametric mode.
`sample.1.prop`	Numeric vector of specified proportion of non-zeros for subjects (the length should be equal to `n.sample` in `fitted`). This argument is only applicable in nonparametric mode.
`taxa.1.prop`	Numeric vector of specified proportion of non-zeros for taxa (the length should be equal to `n.taxa` in `fitted`). This argument is only applicable in nonparametric mode.
`individual.rel.abund`	Numeric matrix of expected relative abundances with `n.sample` rows and `n.taxa` columns (rows should sum to 1). Provides subject-specific mean compositions and therefore overrides `mean.rel.abund` and `gengamma.mu`. Only applicable in parametric mode.
`...`	Additional arguments. If SCAM model is chosen for parameter changes under the non-parametric mode, specify `SCAM = T`.

Details

The parametric model in MIDASim is a location-scale model, specifically, a generalized gamma model for relative abundances π of a taxon. Denote $t = 1/\pi$. The generalized gamma distribution for t is chosen so that

$$ln(t) = \mu + \sigma \cdot w$$

where w follows a log gamma distribution with a shape parameter $1/Q$. MIDASim fits the model to the template data and estimates parameters μ, σ and Q by matching the first two moments of π and maximizing the likelihood.

Value

Returns an updated list with different elements depending on the value of `fitted$mode`:

`n.sample`	Target sample size in the simulation.
`lib.size`	Target library sizes in the simulation.
`taxa.1.prop`	Updated proportions of non-zero values for each taxon.
`sample.1.prop`	Updated proportion of non-zero cells for each subject.

`theta`	Mean values of the multivariate normal distribution in generating presence-absence data.
`eta`	Adjustment to be applied to samples in generating presence- absence data.

Author(s)

Mengyu He

Examples

```
data("throat.otu.tab")
otu.tab = throat.otu.tab[,colSums(throat.otu.tab>0)>1]

fitted = MIDASim.setup(otu.tab, mode = 'parametric')

# modify library sizes
fitted.modified <- MIDASim.modify(fitted,
                                  lib.size = sample(fitted$lib.size, 2*nrow(otu.tab),
                                                    replace = TRUE) )

# modify mean relative abundances
fitted.modified <- MIDASim.modify(fitted,
                        mean.rel.abund = fitted$mean.rel.abund * runif(fitted$n.taxa))
```

`MIDASim.setup` *Fitting MIDAS model to microbiome data*

Description

Midas.setup estimates parameters from a template microbiome count dataset for downstream data simulation.

Usage

```
MIDASim.setup(otu.tab, n.break.ties = 100, mode = "nonparametric")
```

Arguments

`otu.tab`	Numeric matrix of template microbiome count dataset. Rows are samples, columns are taxa.
`n.break.ties`	Number of replicates to break ties when ranking relative abundances. Defaults to `100`.
`mode`	A character indicating the modeling approach for relative abundances. If `'parametric'`, a parametric model involving fitting a generalized gamma distribution is used. If `'nonparametric'`, the nonparametric approach involving quantile matching is applied. Note that a parametric model is required if library sizes or characteristics of taxa will be modified. Defaults to `'nonparametric'`.

Value

Returns a list that has components:

`mat01`	Presence-absence matrix of the template data.
`lib.size`	Observed library sizes of the template data.
`n.taxa`	Number of taxa in the template data.
`n.sample`	Sample size in the template data.
`ids`	Taxa ids present in all samples in the template.
`tetra.corr`	Estimated tetrachoric correlation of the presence-absence matrix of the template.
`corr.rel.corrected`	Estimated Pearson correlation of relative abundances, transformed from Spearman's rank correlation.
`sample.1.prop`	Proportion of non-zero cells for each subject.
`taxa.1.prop`	Proportion of non-zeros for each taxon.
`mean.rel.abund`	Observed mean relative abundances of each taxon.
`rel.abund.1`	Observed non-zero relative abundances of each taxon.
`taxa.names`	Names of taxa in the template.

Author(s)

Mengyu He

Examples

```
data("throat.otu.tab")
  otu.tab = throat.otu.tab[,colSums(throat.otu.tab>0)>1]

  # use nonparametric model
  fitted = MIDASim.setup(otu.tab)

  # use parametric model
  fitted = MIDASim.setup(otu.tab, mode = 'parametric')
```

`throat.otu.tab` *throat microbiome dataset*

Description

A microbiome dataset of 60 subjects with 856 OTUs. The data were collected from right and left nasopharynx and oropharynx region.

Usage

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

Format

An object of class `data.frame` with 60 rows and 856 columns.

References

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D., & Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. PloS one, 5(12), e15216. https://doi.org/10.1371/journal.pone.0015216

Examples

```
data(throat.otu.tab)

MIDASim.setup(otu.tab = throat.otu.tab, mode = "nonparametric")
```

Index