Package 'microbiomeDataSets'

May 26, 2026

Title Experiment Hub based microbiome datasets

Version 1.20.0

Description microbiomeDataSets is a collection of microbiome datasets loaded from Bioconductor'S ExperimentHub infrastructure. The datasets serve as reference for workflows and vignettes published adjacent to the microbiome analysis tools on Bioconductor. Additional datasets can be added overtime and additions from authors are welcome.

License CC0

Encoding UTF-8

biocViews ExperimentHub, ExperimentData, MicrobiomeData, SequencingData

Depends R (>= 4.1), SummarizedExperiment, TreeSummarizedExperiment, MultiAssayExperiment

Imports methods, utils, BiocGenerics, ExperimentHub, Biostrings, ape

Suggests knitr, rmarkdown, BiocStyle, SingleCellExperiment, testthat

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

VignetteBuilder knitr

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

git_branch RELEASE_3_23

git_last_commit f7f9739

git_last_commit_date 2026-04-28

Repository Bioconductor 3.23

Date/Publication 2026-05-26

Author Leo Lahti [cre, aut] (ORCID: <https://orcid.org/0000-0001-5537-637X>), Felix G.M. Ernst [aut] (ORCID: <https://orcid.org/0000-0001-5064-0928>), Sudarshan Shetty [aut] (ORCID: <https://orcid.org/0000-0001-7280-9915>), Chouaib Benchraka [ctb] (ORCID: <https://orcid.org/0000-0002-1167-7848>), Yagmur Simsek [ctb]

Maintainer Leo Lahti <leo.lahti@iki.fi>

Contents

`microbiomeDataSets-package` *Microbiome data sets*

Description

`microbiomeDataSets` is a collation of data from microbiome and associated studies, which are publicaly available.

Details

The data is made available through the ExperimentHub resources of the Bioconductor project. It is loaded as `TreeSummarizedExperiment` object or a `MultiAssayExperiment` objects.

`availableDataSets` *Load available microbiome data sets*

Description

To list the available datasets in `microbiomeDataSets`, run `availableDataSets`.

Usage

```
availableDataSets()
```

Details

For information visit the individual man pages.

Value

A `data.frame` containing the following columns:

- Dataset: the name of the function to load a dataset

Examples

```
availableDataSets()
```

`GrieneisenTSData` *Retrieve GrieneisenTS data*

Description

Obtain longitudinal gut microbiome data in wild baboons from Grieneisen et al. (2021).

Usage

```
GrieneisenTSData()

baboongut()
```

Details

The GrieneisenTS dataset contains 16,234 16S rRNA gene sequencing-based microbiome profiles from 585 baboon samples collected over 14 years to determine the heritability of the gut microbiome on various environmental factors such as diet, age, season. Each baboon had an average of 28 samples collected over 4.5 years. The data set can be used to investigate significance of longitudinal sampling at large sample sizes.

This data set contains the 613 most prevalent taxa with a phylogenetic tree.

Column metadata includes the following fields:

- sample: Sample ID (character)
- baboon_id: Baboon ID (factor)
- collection_date: Sample collection date (date ; YYYY-MM-DD)
- sex: Sex (factor; F/M)
- age: Age (numeric)
- social_group: Social group ID (factor)
- group_size: Social group size (integer)
- rain_month_mm: Rain per month(mm) (numeric)
- season: Season (factor; dry/wet)
- hydro_year: Hydro year (integer)
- month: Month (integer)
- readcount: Read count (numeric)
- plate: Plate (factor)
- post_pcr_dna_ng: Post PCR DNA(ng) (numeric)
- diet_PC1: Diet Principal coordinate 1 (numeric)
- diet_PC2: Diet Principal coordinate 2 (numeric)

- diet_PC3: Diet Principal coordinate 3 (numeric)
- diet_PC4: Diet Principal coordinate 4 (numeric)
- diet_PC5: Diet Principal coordinate 5 (numeric)
- diet_PC6: Diet Principal coordinate 6 (numeric)
- diet_PC7: Diet Principal coordinate 7 (numeric)
- diet_PC8: Diet Principal coordinate 8 (numeric)
- diet_PC9: Diet Principal coordinate 9 (numeric)
- diet_PC10: Diet Principal coordinate 10 (numeric)
- diet_PC11: Diet Principal coordinate 11 (numeric)
- diet_PC12: Diet Principal coordinate 12 (numeric)
- diet_PC13: Diet Principal coordinate 13 (numeric)
- diet_shannon_h: Dietary Shannon's H index (numeric)
- asv_richness: Amplicon sequence variant (ASV) richness (integer)
- asv_shannon_h: ASV Shannon's H index (numeric)
- pc1_bc: Principal coordinate 1 Bray-Curtis dissimilarity (numeric)
- pc2_bc: Principal coordinate 2 Bray-Curtis dissimilarity (numeric)
- pc3_bc: Principal coordinate 3 Bray-Curtis dissimilarity (numeric)
- pc4_bc: Principal coordinate 4 Bray-Curtis dissimilarity (numeric)
- pc5_bc: Principal coordinate 5 Bray-Curtis dissimilarity (numeric)

Row metadata of the microbiome data contains taxonomic information on the Domain, Phylum, Class, Order, Family, Genus, and ASV levels.

The row tree consists of a phylogenetic tree build using sequence information of 613 taxa.

As reference sequences the ASV are provided.

Value

A TreeSummarizedExperiment object.

Author(s)

Yagmur Simsek and Leo Lahti

References

Grieneisen et al. (2021): Gut microbiome heritability is nearly universal but environmentally contingent *Science* 373:6551 https://science.sciencemag.org/content/373/6551/181.full

Examples

```
tse <- GrieneisenTSData()
```

LahtiMLData *Obtain the Lahti Microbiome and Lipid data*

Description

Obtain the microbiome and lipid data from Lahti et al. (2013).

Usage

```
LahtiMLData()

LahtiMData()

peerj32()
```

Details

The LahtiML dataset contains high-throughput profiling data from 389 human blood serum lipids and 130 intestinal genus-level bacteria from 44 samples (22 subjects from 2 time points; before and after probiotic/placebo intervention). The data set can be used to investigate associations between intestinal bacteria and host lipid metabolism

This data set is based on the Human Intestinal Tract (HIT)Chip phylogenetic 16S microarray (Rajilic-Stojanovic *et al.* 2009. This profiling technology differs from the more widely used 16S rRNA amplicon sequencing.

Column metadata includes the time points, sex, subject identifier, sample identifier and treatment group.

Row metadata of the microbiome data contains taxonomic information on the Phylum, Family and Genus level.

All data are downloaded from ExperimentHub and cached for local re-use.

Value

for `LahtiMLData` a MultiAssayExperiment object with TreeSummarizedExperiment and a SummarizedExperiment, for `LahtiMData` a TreeSummarizedExperiment

Author(s)

Felix G.M. Ernst and Leo Lahti

References

Lahti L et al. (2013): Associations between the human intestinal microbiota, Lactobacillus rhamnosus GG and serum lipids indicated by integrated analysis of high-throughput profiling data. *PeerJ* 1:e32 https://doi.org/10.7717/peerj.32

Rajilic-Stojanovic M, Heilig HG, Molenaar D, Kajander K, Surakka A, Smidt H, de Vos WM (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol.* 11(7):1736-51 https://doi.org/10.1111/j.1462-2920.2009.01900.x

Examples

```
mae <- LahtiMLData()
tse <- LahtiMData()
```

`OKeefeDSData` *Obtain the O'Keefe diet swap microbiome data*

Description

Obtain the microbiome data from O'Keefe et al. (2015).

Usage

```
OKeefeDSData()

dietswap()
```

Details

The OKeefeDS dataset contains microbiome data from a study with African and African American groups undergoing a two-week diet swap.

This data set is based on the Human Intestinal Tract (HIT)Chip phylogenetic 16S microarray (Rajilic-Stojanovic *et al.* 2009. This profiling technology differs from the more widely used 16S rRNA amplicon sequencing.

Column metadata includes the subject identifier, sex, nationality, group information, sample identifier, time point information, time point information within group and BMI group.

Row metadata contains taxonomic information on the Phylum, Family and Genus level.

All data are downloaded from ExperimentHub and cached for local re-use.

Value

a TreeSummarizedExperiment

Author(s)

Felix G.M. Ernst and Leo Lahti

References

O'Keefe S et al. (2015): Fat, fibre and cancer risk in African Americans and rural Africans. *Nature Communications* 6:6342 (2015) https://dx.doi.org/10.1038/ncomms7342

Rajilic-Stojanovic M, Heilig HG, Molenaar D, Kajander K, Surakka A, Smidt H, de Vos WM (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ Microbiol.* 11(7):1736-51 https://doi.org/10.1111/j.1462-2920.2009.01900.x

Examples

```
tse <- OKeefeDSData()
```

`SilvermanAGutData` *Obtain the Silverman Artificial Gut data*

Description

Obtain the microbiome data from Silverman et al. (2018).

Usage

```
SilvermanAGutData()

artificialgut()
```

Details

The SilvermanAGutData dataset contains 16S rRNA gene based high-throughput profiling of 4 in vitro artificial gut models. The sampling was done hourly and daily to capture sub-daily dynamics of microbial community originating from human feces. The data consists of 413 taxa from 639 samples. The data set can be used to investigate longitudinal dynamics of microbial community in a controlled environment.

Column metadata includes the days of sampling, vessel identifier, sampling frequency pre-post challenge with Bacteroides ovatus.

The wow metadata of the microbiome data contains taxonomic information on the Kingdom, Phylum, Class, Order, Family and Genus and Species level.

The row tree consists of a phylogenetic tree build using sequence information of 413 taxa.

As reference sequences the ASV are provided.

All data are downloaded from ExperimentHub and cached for local re-use.

Value

a TreeSummarizedExperiment

Author(s)

Sudarshan A. Shetty and Felix G.M. Ernst

References

Silveman J.D et al. (2018): Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* 6:202 https://doi.org/10.1186/s40168-018-0584-3

Examples

```
tse <- SilvermanAGutData()
```

`SongQAData` *Obtain the Song quality assurance dataset*

Description

Obtain the microbiome data from Song et al. (2016).

Usage

```
SongQAData()

qa10934()
```

Details

The SongQA dataset contains microbiome data from a study comparing different sample preparation methods focusing on temperature and DNA extraction methods mainly.

The dataset was generated by sequencing the variable region 4 of 16S rRNA genes via Illumina HiSeq or MiSeq.

Column metadata was reconstructed from sample names and includes information on the source, timepoint, extraction method, temperature conditions and other optional data.

Row metadata contains taxonomic information on the Phylum, Family and Genus level.

All data are downloaded from ExperimentHub and cached for local re-use.

Value

a TreeSummarizedExperiment

Author(s)

Felix G.M. Ernst

Source

https://qiita.ucsd.edu/study/description/10394

References

Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. Se Jin Song, Amnon Amir, Jessica L. Metcalf, Katherine R. Amato, Zhenjiang Zech Xu, Greg Humphrey, Rob Knight. mSystems May 2016, 1 (3) e00021-16; DOI: 10.1128/mSystems.00021-16

Examples

```
tse <- SongQAData()
```

`SprockettTHData` *Obtain the Sprockett Tsimane Horticulturalists data*

Description

Obtain the microbiome data from Sprockett et al. (2020).

Usage

```
SprockettTHData()
```

Details

The SprockettTHData dataset contains 16S rRNA gene based high-throughput profiling of 1966 Feces, 120 Saliva samples from 319 participants from Bolivia, Finland and Bangladesh. These include samples from adults, children, and infants. Several participants have longitudinal samples. The data consists of 2319 taxa from 2086 samples. The data set can be used to investigate assembly, structure, and dynamics as well as associations between several host related parameters with microbiota.

Column metadata includes the sex, age, feeding status, delivery mode, country, and other information.

The row metadata of the microbiome data contains taxonomic information on the Kingdom, Phylum, Class, Order, Family and Genus, Species and lowest taxonomic rank.

The row tree consists of a phylogenetic tree build using sequence information of 2319 taxa.

As reference sequences the ASV are provided.

All data are downloaded from ExperimentHub and cached for local re-use.

Value

a TreeSummarizedExperiment

Author(s)

Sudarshan A. Shetty and Felix G.M. Ernst

References

Sprockett, D.D., Martin, M., Costello, E.K. et al. (2020) Microbiota assembly, structure, and dynamics among Tsimane horticulturalists of the Bolivian Amazon. *Nat Commun* 11, 3772 https://doi.org/10.1038/s41467-020-17541-6

Subramanian, S., Huq, S., Yatsunenko, T., et al. (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* 510, 417-421 https://doi.org/10.1038/nature13421

Vatanen, T., Kostic A.D., d'Hennezel E., et al. (2016) Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* 165, 842-853 https://doi.org/10.1016/j.cell.2016.04.007

Examples

```
tse <- SprockettTHData()
```

Index