Package 'HMP16SData'

October 30, 2025

Type Package

Title 16S rRNA Sequencing Data from the Human Microbiome Project

Version 1.31.0

Description HMP16SData is a Bioconductor ExperimentData package of the Human Microbiome Project (HMP) 16S rRNA sequencing data for variable regions 1–3 and 3–5. Raw data files are provided in the package as downloaded from the HMP Data Analysis and Coordination Center. Processed data is provided as SummarizedExperiment class objects via ExperimentHub.

License Artistic-2.0

Encoding UTF-8

LazyData true

Depends R (>= 4.1.0), SummarizedExperiment

Imports AnnotationHub, ExperimentHub, S4Vectors, assertthat, dplyr, kableExtra, knitr, magrittr, methods, readr, stringr, tibble, utils

Suggests Biobase, BiocCheck, BiocManager, BiocStyle, circlize, cowplot, dendextend, devtools, ggplot2, gridExtra, haven, phyloseq, rmarkdown, roxygen2, stats, testthat, tidyr

biocViews ExperimentData, ExperimentHub, Homo_sapiens_Data, MicrobiomeData, ReproducibleResearch, SequencingData

VignetteBuilder knitr

URL https://github.com/waldronlab/HMP16SData

BugReports https://github.com/waldronlab/HMP16SData/issues

RoxygenNote 7.1.2

git_url https://git.bioconductor.org/packages/HMP16SData

git_branch devel

git_last_commit dbabb16

git_last_commit_date 2025-10-29

Repository Bioconductor 3.23

Date/Publication 2025-10-30

Author Lucas schiffer [aut, cre] (ORCID:
<https://orcid.org/0000-0003-3628-0326>),
Rimsha Azhar [aut],
Marcel Ramos [ctb],
Ludwig Geistlinger [ctb],
Levi Waldron [aut]

Maintainer Lucas schiffer <schiffer.lucas@gmail.com>

Contents

`as_phyloseq` *Coerce a SummarizedExperiment object to a phyloseq object*

Description

The `phyloseq-package` provides a suite of methods for working with 16S rRNA sequencing and other microbiome data that may be of use to the users of `HMP16SData`. The `as_phyloseq` method provides a means to easily coerce a `SummarizedExperiment-class` object to a `phyloseq-class` object.

Usage

```
as_phyloseq(x)
```

Arguments

`x`	A `SummarizedExperiment-class` object from the `HMP16SData` package

Value

A `phyloseq-class` object

Examples

```
V13() %>%
    as_phyloseq()
```

`attach_dbGaP` *Attach dbGaP metadata to a SummarizedExperiment object*

Description

The National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP) has extensive metadata pretaining to the HMP study. Access to this metadata is controlled and its use requires authorization by the appropriate Data Access Committee – further information is available here. Those with authorization will be able to download a dbGaP repository key and can use it to attach the controlled-access metadata to a `SummarizedExperiment-class` object with a single command, provided the NCBI SRA Toolkit is installed and on the user's `PATH`. See note below for further information.

Usage

```
attach_dbGaP(x, dbGaP_repository_key = "")
```

Arguments

`x`	A `SummarizedExperiment-class` object from the `HMP16SData` package
`dbGaP_repository_key`	A repository key downloaded from dbGaP; only required for the initial download and not needed thereafter

Value

A `SummarizedExperiment-class` object with protected metadata from dbGaP attached

Note

The NCBI SRA Toolkit is called internally and must be on the user's `PATH` or an error message will be displayed. The NCBI SRA Toolkit can readily be installed on multiple platforms following these instructions.

See Also

`dictionary`

Examples

```
## Not run:
V13() %>%
    attach_dbGaP(dbGaP_repository_key = "~/prj_12146.ngc")

## End(Not run)
```

`dictionary` *A data dictionary describing dbGaP metadata variables*

Description

A data dictionary containing descriptions of the metadata variables attached using the `attach_dbGaP` method, as translated from dbGaP `XLM` data dictionary files.

Usage

```
dictionary
```

Format

A `data.frame` with 789 rows and 8 variables:

VARIABLE The name of the variable as it appears in `SummarizedExperiment` `colData`

DESCRIPTION A description of the variable

TYPE The type of variable, as specified by dbGaP

UNITS The units of the variable (e.g. days)

MINIMUM The smallest possible value of the variable

MAXIMUM The largest possible value of the variable

CODED_VALUES The encoding of coded categorical variables, if needed

COMMENT Any comments from dbGaP, most pertain to gated fields

Source

The following dbGaP `XLM` data dictionary files were used to construct the data dictionary.

Data dictionary for data table pht001185.v4.p1: https://tinyurl.com/y7rgelao. No description provided by dbGaP.

Data dictionary for data table pht001193.v3.p1: https://tinyurl.com/y7at85tv. The data table contains subject data with sample information for the Global Trace Screen (GTV). Variables include the specimen type (based on 18 specimens from 6 body sites), nucleic extraction method, and sample administrative information. The Global Trace dataset will be updated when additional visits or subjects have been added.

Data dictionary for data table pht002157.v1.p1: https://tinyurl.com/y7af3arf. The Targeted Physical Forms (DTP), Medical History Screening Forms (DHX), and the Visit Documentation

Forms (DVD) have been combined into a single data table. The DTP variables include measurements collected for blood pressure, pulse, body weight, height, and BMI. Variables also include physical assessment of areas/systems and the specific details of the abnormalities. Areas/Systems consist of general appearance, HEENT, cardiovascular, pulmonary, abdomen, neurological, musculoskeletal, and extremities. The DHX variables include data that indicate whether the subject noted medical problems in various areas/systems and the indications and specifications of medical problems. Areas/Systems consist of HEENT, cardiovascular, pulmonary, GI, hepatobiliary, renal, neurologic, blood lymphatic, endocrine/metabolic, musculoskeletal, genital/reproductive, dermatologic, allergies, cancer, immunodeficiency, drug or alcohol dependence, and autoimmune disease. Data were also collected for relevant medical or surgical history, medical abnormality with ongoing treatment and presence of acute disease. The DVD variables include education level, insurance status, occupation status, tobacco usage, and visit-specific information. This visit-specific information pertains to subject's oral temperature, pregnancy test results, education level, insurance status, occupation, and if blood and GI specimens were collected. DTP, DHX and DVD data will be collected at each visit though not every variable was collected.

Data dictionary for data table pht001187.v3.p1: https://tinyurl.com/ych4wywl. The data table contains data that were collected from the Concomitant Medications Form (DCM). Variables include medication number and code, indication, if the medication is ongoing, and the duration of the medication.

Data dictionary for data table pht001184.v3.p1: https://tinyurl.com/y9hwfn9e. The data table contains the subject and consent group information.

Data dictionary for data table pht002158.v1.p1: https://tinyurl.com/y7jtghxv. The Demographics Form (DEM) and the Eligibility Checklist (ENR) have been combined into a single data table. The DEM variables include sociodemographic measurements such as gender, ethnicity, race, and place of birth, while the ENR variables include the enrollment/first sampling time of subjects (n=1 variable) and administrative variables (n=4 variables). Both the Medical History and Targeted Physical exams must be completed prior to the completion of the Exclusion Criteria found in the ENR form. Data derived from the Medical History (DHX) and Targeted Physical (DTP) exams can be found in EMMES_HMP_DTP_DHX_DVD dataset.

Data dictionary for data table pht002156.v1.p1: https://tinyurl.com/y7znjmak. The data table contains data that were collected from the Suppplemental Questions Form (DSU) upon the completion of HMP study enrollment. Variables include additional health history such as dietary habits, breastfeeding, birthing history, delivery method, and if the subject failed any initial screenings prior to enrollment.

See Also

attach_dbGaP

Examples

```
head(dictionary)
```

HMP16SData *16S rRNA sequencing data from the Human Microbiome Project*

Description

HMP16SData is a Bioconductor ExperimentData package of the Human Microbiome Project (HMP) 16S rRNA sequencing data for variable regions 1–3 and 3–5. Raw data files are provided in the package as downloaded from the HMP Data Analysis and Coordination Center. Processed data is provided as SummarizedExperiment-class objects via ExperimentHub.

kable_one *Produce a summary HTML table of key demographic variables*

Description

Sometimes it is desirable to produce a summary of key demographic variables for presentation. The table_one and kable_one methods are a quick way to do so – they not only summarize key demographic variables from SummarizedExperiment-class object(s) in the HMP16SData package but remove abbreviations and underscores in column and variable names that might otherwise be ambigious. The table_one method returns a *tidy* (i.e. one sample observation per line) data.frame object or a named list of *tidy* data.frame objects. The kable_one method can then be used to produce a publication-ready HTML table that could, for example, be pasted into a word processor.

Usage

```
kable_one(x, significant_figures = 2)
```

Arguments

x	A data.frame object or a named list of data.frame objects returned from the table_one method
significant_figures	The number of significant figures to be used for decimals in the HTML table; if no value is specified, the default is 2

Value

A summary HTML table of key demographic variables

See Also

table_one

Examples

```
V13() %>%
    table_one() %>%
    kable_one()
```

`reexports` *Objects exported from other packages*

Description

These objects are imported from other packages. Follow the links below to see their documentation.

magrittr `%>%`

Examples

```
V13() %>%
    colData()
```

`table_one` *Produce a summary data.frame of key demographic variables*

Description

Sometimes it is desirable to produce a summary of key demographic variables for presentation. The `table_one` and `kable_one` methods are a quick way to do so – they not only summarize key demographic variables from `SummarizedExperiment-class` object(s) in the `HMP16SData` package but remove abbreviations and underscores in column and variable names that might otherwise be ambigious. The `table_one` method returns a *tidy* (i.e. one sample observation per line) `data.frame` object or a named `list` of *tidy* `data.frame` objects. The `kable_one` method can then be used to produce a publication-ready `HTML` table that could, for example, be pasted into a word processor.

Usage

```
table_one(
  x,
  VISITNO = TRUE,
  SEX = TRUE,
  RUN_CENTER = TRUE,
  HMP_BODY_SITE = TRUE,
  HMP_BODY_SUBSITE = TRUE
)
```

Arguments

x	A SummarizedExperiment-class object or a named list of SummarizedExperiment-class objects from the HMP16SData package
VISITNO	logical; if FALSE, the VISITNO column(s) of SummarizedExperiment colData will not be summarized
SEX	logical; if FALSE, the SEX column(s) of SummarizedExperiment colData will not be summarized
RUN_CENTER	logical; if FALSE, the RUN_CENTER column(s) of SummarizedExperiment colData will not be summarized
HMP_BODY_SITE	logical; if FALSE, the HMP_BODY_SITE column(s) of SummarizedExperiment colData will not be summarized
HMP_BODY_SUBSITE	logical; if FALSE, the HMP_BODY_SUBSITE column(s) of SummarizedExperiment colData will not be summarized

Value

A data.frame object or a named list of data.frame objects

See Also

kable_one

Examples

```
V13() %>%
    table_one()
```

V13 *HMP 16S rRNA sequencing data for variable regions 1–3*

Description

The NIH Human Microbiome Project (HMP) was a longitudinal study conducted from 2007 to 2012 across four institutions (Baylor College of Medicine, the Broad Institute, the J. Craig Venter Institute, and Washington University) of healthy adults aged 18 to 40 that produced a comprehensive reference for the composition, diversity, and variation of the healthy human microbiome. This SummarizedExperiment-class object represents 16S rRNA sequencing data for variable regions 1–3 that was performed on samples collected at five major body sites – available participant metadata as well as phylogenetic trees are included.

Usage

```
V13(metadata = FALSE)
```

Arguments

metadata	logical; if TRUE only the metadata is downloaded, rather than the entire resource

Format

A SummarizedExperiment-class object with 43,140 features and 2,898 samples:

colData:

RSID a random subject identifier

VISITNO visit number, between 1 and 3

SEX sex, female or male

RUN_CENTER center where sample sequencing took place: Baylor College of Medicine (BCM), the Broad Institute (BI), the J. Craig Venter Institute (JCVI), or the Genome Sequencing Center at Washington University (WUGC)

HMP_BODY_SITE body site where the sample was collected

HMP_BODY_SUBSITE body subsite where the sample was collected

SRS_SAMPLE_ID a sample identifier to be used when comparing 16S rRNA samples to whole metagenome shotgun (WMS) samples

rowData:

CONSENSUS_LINEAGE the most detailed lineage description shared by the sequences within an OTU

SUPERKINGDOM superkingdom taxonomy, assumed to be Bacteria

PHYLUM phylum taxonomy parsed from CONSENSUS_LINEAGE

CLASS calss taxonomy parsed from CONSENSUS_LINEAGE

ORDER order taxonomy parsed from CONSENSUS_LINEAGE

FAMILY family taxonomy parsed from CONSENSUS_LINEAGE

GENUS genus taxonomy parsed from CONSENSUS_LINEAGE

Value

A SummarizedExperiment-class object

Note

The "PSN" identifiers were used as the colnames of the SummarizedExperiment-class object, see source for additional information.

Source

The following source information is derived from the HMP Data Analysis and Coordination Center:

Following a July 2010 16S data freeze, data was downloaded from NCBI SRA projects SRP002395: Human Microbiome Project 16S rRNA Clinical Production Phase I, and SRP002012: Human Microbiome Project 454 Clinical Production Pilot. This dataset corresponds to over 5,700 samples and over 10,000 sequence preps. 16S variable region 3–5 (V35) was sequenced for the entire set of samples, and variable region 1–3 (V13) for a subset of samples.

The QIIME (Quantitative Insights Into Microbial Ecology) software package was used to process HMP 16S data using an OTU-binning strategy to which taxonomic classification is added.

Raw 16S sequence and metadata, available through https://tinyurl.com/y7ev836z, were demultiplexed using QIIME. OTU picking was performed for the V1–3 and V3–5 region sequences using OTUPipe, which includes error correction, chimera checking through UCHIME, and clustering via UCLUST, and postprocessing by picking the optimal representative sequence centroid. Taxonomy was assigned using the RDP classifier version 2.2.

The resulting OTU tables were checked for mislabeling and contamination, as described in the SOP available through https://tinyurl.com/y7ev836z. Alpha and beta diversity for each sample and Procrustes analysis were established using QIIME with default parameters.

All QIIME output files are available through https://tinyurl.com/y7ev836z, for both the V1–3 and V3–5 variable regions, as well as Procrustes summary data. SOPs and custom scripts are also available through https://tinyurl.com/y7ev836z.

If you're interested in joint analysis of 16S and shotgun metagenomic datasets from the HMP, pairing up data from the same microbiome samples can initially seem tricky. The HMP Sample Flow Schematic indicates how these sample IDs are related experimentally, and provides tables joining 16S dataset "SN" and "PSN" identifiers with metagenomic dataset "SRS" identifiers.

Four files were used to construct this SummarizedExperiment-class object.

OTU table file with PSN identifiers: https://tinyurl.com/y74gqpho

Subject metadata files with PSN identifiers: https://tinyurl.com/y8adlfso

Subject metadata files with SRS identifiers: https://tinyurl.com/ybmn7q8m

Representative sequence phylogenetic trees: https://tinyurl.com/ybp8mzgj

See Also

V35

Examples

```
V13()
```

V35 *HMP 16S rRNA sequencing data for variable regions 3–5*

Description

The NIH Human Microbiome Project (HMP) was a longitudinal study conducted from 2007 to 2012 across four institutions (Baylor College of Medicine, the Broad Institute, the J. Craig Venter Institute, and Washington University) of healthy adults aged 18 to 40 that produced a comprehensive reference for the composition, diversity, and variation of the healthy human microbiome. This SummarizedExperiment-class object represents 16S rRNA sequencing data for variable regions 3–5 that was performed on samples collected at five major body sites – available participant metadata as well as phylogenetic trees are included.

Usage

```
V35(metadata = FALSE)
```

Arguments

metadata	logical; if TRUE only the metadata is downloaded, rather than the entire resource

Format

A SummarizedExperiment with 45,383 features and 4,743 samples:

colData:

RSID a random subject identifier

VISITNO visit number, between 1 and 3

SEX sex, female or male

RUN_CENTER center where sample sequencing took place: Baylor College of Medicine (BCM), the Broad Institute (BI), the J. Craig Venter Institute (JCVI), or the Genome Sequencing Center at Washington University (WUGC)

HMP_BODY_SITE body site where the sample was collected

HMP_BODY_SUBSITE body subsite where the sample was collected

SRS_SAMPLE_ID a sample identifier to be used when comparing 16S rRNA samples to whole metagenome shotgun (WMS) samples

rowData:

CONSENSUS_LINEAGE the most detailed lineage description shared by the sequences within an OTU

SUPERKINGDOM superkingdom taxonomy, assumed to be Bacteria

PHYLUM phylum taxonomy parsed from CONSENSUS_LINEAGE

CLASS calss taxonomy parsed from CONSENSUS_LINEAGE

ORDER order taxonomy parsed from CONSENSUS_LINEAGE

FAMILY family taxonomy parsed from CONSENSUS_LINEAGE

GENUS genus taxonomy parsed from CONSENSUS_LINEAGE

Value

A SummarizedExperiment object

Note

The "PSN" identifiers were used as the colnames of the SummarizedExperiment object, see source for additional information.

Source

The following source information is derived from the HMP Data Analysis and Coordination Center:

Following a July 2010 16S data freeze, data was downloaded from NCBI SRA projects SRP002395: Human Microbiome Project 16S rRNA Clinical Production Phase I, and SRP002012: Human Microbiome Project 454 Clinical Production Pilot. This dataset corresponds to over 5,700 samples and over 10,000 sequence preps. 16S variable region 3–5 (V35) was sequenced for the entire set of samples, and variable region 1–3 (V13) for a subset of samples.

The QIIME (Quantitative Insights Into Microbial Ecology) software package was used to process HMP 16S data using an OTU-binning strategy to which taxonomic classification is added.

Raw 16S sequence and metadata, available through https://tinyurl.com/y7ev836z, were demultiplexed using QIIME. OTU picking was performed for the V1–3 and V3–5 region sequences using OTUPipe, which includes error correction, chimera checking through UCHIME, and clustering via UCLUST, and postprocessing by picking the optimal representative sequence centroid. Taxonomy was assigned using the RDP classifier version 2.2.

The resulting OTU tables were checked for mislabeling and contamination, as described in the SOP available through https://tinyurl.com/y7ev836z. Alpha and beta diversity for each sample and Procrustes analysis were established using QIIME with default parameters.

All QIIME output files are available through https://tinyurl.com/y7ev836z, for both the V1–3 and V3–5 variable regions, as well as Procrustes summary data. SOPs and custom scripts are also available through https://tinyurl.com/y7ev836z.

If you're interested in joint analysis of 16S and shotgun metagenomic datasets from the HMP, pairing up data from the same microbiome samples can initially seem tricky. The HMP Sample Flow Schematic indicates how these sample IDs are related experimentally, and provides tables joining 16S dataset "SN" and "PSN" identifiers with metagenomic dataset "SRS" identifiers.

Four files were used to construct this SummarizedExperiment-class object.

OTU table file with PSN identifiers: https://tinyurl.com/y9rbpjl7

Subject metadata files with PSN identifiers: https://tinyurl.com/yaz35f22

Subject metadata files with SRS identifiers: https://tinyurl.com/y9xjqm29

Representative sequence phylogenetic trees: https://tinyurl.com/y9exxlgr

See Also

V13

Examples

```
V35()
```

Index