

Package: curatedMetagenomicData (via r-universe)

October 10, 2026

Title Curated Metagenomic Data of the Human Microbiome

Version 3.20.0

Date 2025-9-26

Description The curatedMetagenomicData package provides standardized, curated human microbiome data for novel analyses. It includes gene families, marker abundance, marker presence, pathway abundance, pathway coverage, and relative abundance for samples collected from different body sites. The bacterial, fungal, and archaeal taxonomic abundances for each sample were calculated with MetaPhlAn3, and metabolic functional potential was calculated with HUMAnN3. The manually curated sample metadata and standardized metagenomic data are available as (Tree)SummarizedExperiment objects.

License Artistic-2.0

Depends R (>= 4.1.0), SummarizedExperiment, TreeSummarizedExperiment

Imports AnnotationHub, ExperimentHub, S4Vectors, dplyr, magrittr, mia, purrr, rlang, stringr, tibble, tidyr, tidyselect

Suggests BiocStyle, DT, knitr, readr, rmarkdown, scater, testthat, utils, uwot, vegan, OmicsMLRepoR, lefser

URL <https://github.com/waldronlab/curatedMetagenomicData>

BugReports <https://github.com/waldronlab/curatedMetagenomicData/issues>

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

VignetteBuilder knitr

biocViews ExperimentHub, Homo_sapiens_Data, MicrobiomeData, ReproducibleResearch

Config/pak/sysreqs cmake libglpk-dev make libgs10-dev libicu-dev libpng-dev libuv1-dev libxml2-dev libssl-dev zlib1g-dev

Repository Bioconductor 3.23

Date/Publication 2026-04-28 12:29:56 UTC

RemoteUrl <https://github.com/bioc/curatedMetagenomicData>

RemoteRef RELEASE_3_23

RemoteSha a97706cfad6c5ec60e6c2e0d64bff759a8f5e93b

Contents

curatedMetagenomicData	2
mergeData	3
returnSamples	4
sampleMetadata	6
Index	7

curatedMetagenomicData	<i>Access Curated Metagenomic Data</i>
------------------------	--

Description

To access curated metagenomic data users will use `curatedMetagenomicData()` after "shopping" the `sampleMetadata` data frame for resources they are interested in. The `dryrun` argument allows users to perfect a query prior to returning resources. When `dryrun = TRUE`, matched resources will be printed before they are returned invisibly as a character vector. When `dryrun = FALSE`, a list of resources containing `SummarizedExperiment` and/or `TreeSummarizedExperiment` objects, each with corresponding sample metadata, is returned. Multiple resources can be returned simultaneously and if there is more than one date corresponding to a resource, the most recent one is selected automatically. Finally, if a `relative_abundance` resource is requested and `counts = TRUE`, relative abundance proportions will be multiplied by read depth and rounded to the nearest integer.

Usage

```
curatedMetagenomicData(  
  pattern,  
  dryrun = TRUE,  
  counts = FALSE,  
  rownames = "long"  
)
```

Arguments

pattern	regular expression pattern to look for in the titles of resources available in curatedMetagenomicData; "" will return all resources
dryrun	if TRUE (the default), a character vector of resource names is returned invisibly; if FALSE, a list of resources is returned
counts	if FALSE (the default), relative abundance proportions are returned; if TRUE, relative abundance proportions are multiplied by read depth and rounded to the nearest integer prior to being returned
rownames	the type of rownames to use for relative_abundance resources, one of: "long" (the default), "short" (species name), or "NCBI" (NCBI Taxonomy ID)

Details

Above "resources" refers to resources that exists in Bioconductor's ExperimentHub service. In the context of curatedMetagenomicData, these are study-level (sparse) matrix objects used to create [SummarizedExperiment](#) and/or [TreeSummarizedExperiment](#) objects that are ultimately returned as the list of resources. Only the gene_families dataType (see [returnSamples](#)) is stored as a sparse matrix in ExperimentHub – this has no practical consequences for users and is done to optimize storage. When searching for "resources", users will use the study_name value from the [sampleMetadata](#) data.frame.

Value

if dryrun = TRUE, a character vector of resource names is returned invisibly; if dryrun = FALSE, a list of resources is returned

See Also

[mergeData](#), [returnSamples](#), [sampleMetadata](#)

Examples

```
curatedMetagenomicData("AsnicarF_20.+")

curatedMetagenomicData("AsnicarF_2017.relative_abundance", dryrun = FALSE)

curatedMetagenomicData("AsnicarF_20.+relative_abundance", dryrun = FALSE, counts = TRUE)
```

mergeData

Merge curatedMetagenomicData List

Description

To merge the list elements returned from [curatedMetagenomicData](#) into a single [SummarizedExperiment](#) or [TreeSummarizedExperiment](#) object, users will use [mergeData\(\)](#) provided elements are the same dataType (see [returnSamples](#)). This is useful for analysis across entire studies (e.g. meta-analysis); however, when doing analysis across individual samples (e.g. mega-analysis) [returnSamples](#) is preferable.

Usage

```
mergeData(mergeList)
```

Arguments

`mergeList` a list returned from [curatedMetagenomicData](#) where all of the elements are of the same `dataType` (see [returnSamples](#))

Details

Internally, `mergeData()` must full join assays and `rowData` slots of each [SummarizedExperiment](#) or [TreeSummarizedExperiment](#) object (`colData` is merged slightly more efficiently by row binding). While dplyr methods are used for maximum efficiency, users should be aware that memory requirements can be large when merging many list elements.

Value

when `mergeList` elements are of `dataType` (see [returnSamples](#)) `relative_abundance`, a [TreeSummarizedExperiment](#) object is returned; otherwise, a [SummarizedExperiment](#) object is returned

See Also

[curatedMetagenomicData](#), [returnSamples](#)

Examples

```
curatedMetagenomicData("LiJ_20.+marker_abundance", dryrun = FALSE) |>
  mergeData()

curatedMetagenomicData("LiJ_20.+pathway_abundance", dryrun = FALSE) |>
  mergeData()

curatedMetagenomicData("LiJ_20.+relative_abundance", dryrun = FALSE) |>
  mergeData()
```

returnSamples

Return Samples Across Studies

Description

To return samples across studies, users will use `returnSamples()` along with the [sampleMetadata](#) `data.frame` subset to include only desired samples and metadata. The subset [sampleMetadata](#) `data.frame` will be used to get the desired resources, [mergeData](#) will be used to merge them, and the subset [sampleMetadata](#) `data.frame` will be used again to subset the [SummarizedExperiment](#) or [TreeSummarizedExperiment](#) object to include only desired samples and metadata.

Usage

```
returnSamples(sampleMetadata, dataType, counts = FALSE, rownames = "long")
```

Arguments

sampleMetadata	the sampleMetadata data.frame subset to include only desired samples and metadata
dataType	the data type to be returned; one of the following: • "gene_families" • "marker_abundance" • "marker_presence" • "pathway_abundance" • "pathway_coverage" • "relative_abundance"
counts	if FALSE (the default), relative abundance proportions are returned; if TRUE, relative abundance proportions are multiplied by read depth and rounded to the nearest integer prior to being returned
rownames	the type of rownames to use for relative_abundance resources, one of: "long" (the default), "short" (species name), or "NCBI" (NCBI Taxonomy ID)

Details

At present, `curatedMetagenomicData` resources exists only as entire studies which requires potentially getting many resources for a limited number of samples. Furthermore, because it is necessary to use [mergeData](#) internally, the same caveats detailed under **Details** in [mergeData](#) apply here.

Value

when `dataType = "relative_abundance"`, a [TreeSummarizedExperiment](#) object is returned; otherwise, a [SummarizedExperiment](#) object is returned

Examples

```
sampleMetadata |>
  dplyr::filter(age >= 18) |>
  dplyr::filter(!base::is.na(alcohol)) |>
  dplyr::filter(body_site == "stool") |>
  dplyr::select(where(~ !base::all(base::is.na(.x)))) |>
  returnSamples("relative_abundance")
```

sampleMetadata	<i>Manually Curated Sample Metadata</i>
----------------	---

Description

Manually curated sample metadata for all samples in curatedMetagenomicData.

Usage

sampleMetadata

Format

An object of class data.frame with 22588 rows and 141 columns.

Index

* **datasets**

sampleMetadata, [6](#)

curatedMetagenomicData, [2](#), [3](#), [4](#)

mergeData, [3](#), [3–5](#)

returnSamples, [3](#), [4](#), [4](#)

sampleMetadata, [2–5](#), [6](#)

SummarizedExperiment, [2–5](#)

TreeSummarizedExperiment, [2–5](#)