

Package ‘terraTCGAdata’

December 31, 2024

Type Package

Title OpenAccess TCGA Data on Terra as MultiAssayExperiment

Version 1.10.0

Description Leverage the existing open access TCGA data on Terra with well-established Bioconductor infrastructure. Make use of the Terra data model without learning its complexities. With a few functions, you can copy / download and generate a MultiAssayExperiment from the TCGA example workspaces provided by Terra.

Depends AnVILGCP, MultiAssayExperiment

biocViews Software, Infrastructure, DataImport

Imports AnVIL, BiocFileCache, dplyr, GenomicRanges, methods, RaggedExperiment, readr, S4Vectors, stats, tidy, TCGAutils, utils

Suggests AnVILBase, knitr, rmarkdown, BiocStyle, withr, testthat (>= 3.0.0)

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

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

VignetteBuilder knitr

License Artistic-2.0

Encoding UTF-8

RoxygenNote 7.3.2

Config/testthat/edition 3

Date 2024-08-22

git_url <https://git.bioconductor.org/packages/terraTCGAdata>

git_branch RELEASE_3_20

git_last_commit 5658d40

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-12-30

Author Marcel Ramos [aut, cre] (<<https://orcid.org/0000-0002-3242-0582>>)

Maintainer Marcel Ramos <marcel.ramos@sph.cuny.edu>

Contents

.getWorkspaceTable	2
getAssayData	2
getAssayTable	4
getClinical	5
getClinicalTable	6
getTCGAdatalist	7
sampleTypesTable	8
terraTCGAdata	9
terraTCGAworkspace	10
Index	12

.getWorkspaceTable	<i>Obtain the table of datasets from the Terra platform</i>
--------------------	---

Description

The datasets include all TCGA datasets that do not come from the Genomic Data Commons Data Repository because those data use a different data model.

Usage

```
.getWorkspaceTable(project = "^TCGA", cancerCode = ".*")
```

Arguments

project	character(1) A prefix for the regex search across all public projects on the terra platform (default: "^TCGA"). Usually, this does not change.
cancerCode	character(1) Corresponds to the TCGA cancer code (e.g, "ACC" for AdrenoCortical Carcinoma) of interest. The default value of ('.*') provides all available cancer datasets.

Value

A 'tibble' 'data.frame' that match the project in put; by default, TCGA workspaces.

getAssayData	<i>Obtain assay datasets from Terra</i>
--------------	---

Description

Obtain assay datasets from Terra

Usage

```
getAssayData(
  assayName,
  sampleCode = "01",
  tablename = .DEFAULT_TABLENAME,
  workspace = terraTCGAspace(),
  namespace = .DEFAULT_NAMESPACE,
  metacols = .PARTICIPANT_METADATA_COLS,
  sampleIdx = TRUE
)
```

Arguments

assayName	character() The name of the assay dataset column from getAssayTable to import into the current workspace.
sampleCode	character(1) The sample code used to filtering samples e.g., "01" for Primary Solid Tumors, see data("sampleTypes", package = "TCGAutils") for reference
tablename	The Terra data model table from which to extract the clinical data (default: "sample")
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.
metacols	The set of columns that comprise of the metadata columns. See the .PARTICIPANT_METADATA_COLS global variable
sampleIdx	numeric() index or TRUE. Specify an index for subsetting the assay data. This argument is mainly used for example and vignette purposes. To use all the data, use the default value (default: TRUE)

Value

Either a matrix or RaggedExperiment depending on the assay selected

See Also

[getAssayTable\(\)](#)

Examples

```
if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
getAssayData(
  assayName = "protein_exp__mda_rppa_core__mdanderson_org__Level3__protein_normalization__data",
  sampleCode = c("01", "10"),
  workspace = "TCGA_ACC_OpenAccess_V1-0_DATA"
)
```

getAssayTable	<i>Obtain a reference table for assay data in the Terra data model</i>
---------------	--

Description

The column names in the output can be used in the getAssayData function.

Usage

```
getAssayTable(
  tablename = .DEFAULT_TABLENAME,
  metacols = .PARTICIPANT_METADATA_COLS,
  workspace = terraTCGAspace(),
  namespace = .DEFAULT_NAMESPACE
)
```

Arguments

tablename	The Terra data model table from which to extract the clinical data (default: "sample")
metacols	The set of columns that comprise of the metadata columns. See the .PARTICIPANT_METADATA_COLS global variable
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.

Value

A tibble of pointers to resources within the Terra data model

Examples

```
if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
  getAssayTable(workspace = "TCGA_COAD_OpenAccess_V1-0_DATA")
```

getClinical	<i>Obtain clinical data</i>
-------------	-----------------------------

Description

The participant table may contain curated demographic information e.g., sex, age, etc.

Usage

```
getClinical(
  columnName,
  participants = TRUE,
  tablename = .DEFAULT_TABLENAME,
  workspace = terraTCGAspace(),
  namespace = .DEFAULT_NAMESPACE,
  verbose = TRUE,
  metacols = .PARTICIPANT_METADATA_COLS,
  participantIds = NULL
)
```

Arguments

columnName	The name of the column to extract files, see getClinicalTable table. If not provided, the first column in the table will be used to obtain the clinical information.
participants	logical(1) Whether to merge the participant table from avtable("participant") to the clinical data
tablename	The Terra data model table from which to extract the clinical data (default: "sample")
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.
verbose	logical(1) Whether to output additional information regarding the workspace and namespace (default: TRUE).
metacols	The set of columns that comprise of the metadata columns. See the .PARTICIPANT_METADATA_COLS global variable
participantIds	character() TCGA participant identifiers usually in the form of "TCGA-AB-1234". By default, all available participant identifiers will be used. (default: NULL)

Value

A DataFrame with clinical information from TCGA. The metadata i.e., metadata(object) includes the columnName used to obtain the data.

Examples

```

if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
getClinical(
  workspace = "TCGA_ACC_OpenAccess_V1-0_DATA",
  participantIds = c("TCGA-OR-A5J1", "TCGA-OR-A5J2",
    "TCGA-OR-A5J3", "TCGA-OR-A5J4")
)

```

getClinicalTable	<i>Obtain the reference table for clinical data</i>
------------------	---

Description

The column names in the output table can be used in the getClinical function.

Usage

```

getClinicalTable(
  tablename = .DEFAULT_TABLENAME,
  metacols = .PARTICIPANT_METADATA_COLS,
  workspace = terraTCGAspace(),
  namespace = .DEFAULT_NAMESPACE,
  verbose = TRUE
)

```

Arguments

tablename	The Terra data model table from which to extract the clinical data (default: "sample")
metacols	The set of columns that comprise of the metadata columns. See the .PARTICIPANT_METADATA_COLS global variable
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.
verbose	logical(1) Whether to output additional information regarding the workspace and namespace (default: TRUE).

Value

A tibble of Google Storage resource locations e.g., gs://firecloud...

Examples

```

if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
  getClinicalTable(
    workspace = "TCGA_ACC_OpenAccess_V1-0_DATA"
  )

```

getTCGAdatalist

Import Terra TCGA data as a list

Description

Import Terra TCGA data as a list

Usage

```

getTCGAdatalist(
  assayNames,
  sampleCode,
  workspace = terraTCGAworkspace(),
  namespace = .DEFAULT_NAMESPACE,
  tablename = .DEFAULT_TABLENAME,
  sampleIdx = TRUE,
  verbose = TRUE
)

```

Arguments

assayNames	character() A vector of assays selected from the colnames of getAssayTable.
sampleCode	character(1) The sample code used to filtering samples e.g., "01" for Primary Solid Tumors, see data("sampleTypes", package = "TCGAutils") for reference
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAworkspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.
tablename	The Terra data model table from which to extract the clinical data (default: "sample")
sampleIdx	numeric() index or TRUE. Specify an index for subsetting the assay data. This argument is mainly used for example and vignette purposes. To use all the data, use the default value (default: TRUE)
verbose	logical(1L) Whether to output additional details of the data facilitation.

Value

A list of assay datasets

Examples

```

if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
getTCGAdatalist(
  assayNames = c("protein_exp__mda_rppa_core__mdanderson_org__Level_3__protein_normalization__data",
    "snp__genome_wide_snp_6__broad_mit_edu__Level_3__segmented_scna_minus_germline_cnv_hg18__seg"),
  sampleCode = c("01", "10"),
  workspace = "TCGA_COAD_OpenAccess_V1-0_DATA"
)

```

sampleTypesTable

Get an overview of the samples available in the workspace

Description

The function provides an overview of samples from the `avtables("sample")` table for the current workspace. Along with the sample codes and frequencies, the output provides a description for each code and the short letter codes.

Usage

```

sampleTypesTable(
  workspace = terraTCGAworkspace(),
  namespace = .DEFAULT_NAMESPACE,
  tablename = .DEFAULT_TABLENAME,
  verbose = TRUE
)

```

Arguments

workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see <code>terraTCGAworkspace()</code>). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.
tablename	The Terra data model table from which to extract the clinical data (default: "sample")
verbose	logical(1) Whether to output additional information regarding the workspace and namespace (default: TRUE).

Value

A tibble of sample codes and frequency along with their definition and short letter code

Examples

```

if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
)
  sampleTypesTable(workspace = "TCGA_COAD_OpenAccess_V1-0_DATA")

```

terraTCGAdata

Obtain a MultiAssayExperiment from the Terra workspace

Description

Workspaces on Terra come pre-loaded with TCGA Data. The examples in the documentation correspond to the TCGA_COAD_OpenAccess_V1 workspace that can be found on app.terra.bio.

Usage

```

terraTCGAdata(
  clinicalName,
  assays,
  participants = TRUE,
  sampleCode = NULL,
  split = FALSE,
  workspace = terraTCGAspace(),
  namespace = .DEFAULT_NAMESPACE,
  tablename = .DEFAULT_TABLENAME,
  verbose = TRUE,
  sampleIdx = TRUE
)

```

Arguments

clinicalName	character(1) The column name taken from getClinicalTable() and downloaded to be included as the colData.
assays	character() A character vector of assay names taken from getAssayTable()
participants	logical(1) Whether to merge the participant table from avtable("participant") to the clinical data
sampleCode	character() A character vector of sample codes from sampleTypesTable(). By default, (NULL) all samples are downloaded and kept in the data.
split	logical(1L) Whether or not to split the MultiAssayExperiment by sample types using splitAssays helper function (default FALSE).
workspace	character(1) The Terra Data Resources workspace from which to pull TCGA data (default: see terraTCGAspace()). This is set to a package-wide option.
namespace	character(1) The Terra Workspace Namespace that defaults to "broad-firecloud-tcga" and rarely needs to be changed.

tablename	The Terra data model table from which to extract the clinical data (default: "sample")
verbose	logical(1) Whether to output additional information regarding the workspace and namespace (default: TRUE).
sampleIdx	numeric() index or TRUE. Specify an index for subsetting the assay data. This argument is mainly used for example and vignette purposes. To use all the data, use the default value (default: TRUE)

Value

A MultiAssayExperiment object with n number of assays corresponding to the assays argument.

Examples

```
if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP")
)
terraTCGAdata(
  clinicalName = "clin__bio__nationwidechildrens_org__Level_1__biospecimen__clin",
  assays = c("protein_exp__mda_rppa_core__mdanderson_org__Level_3__protein_normalization__data",
    "rnaseqv2__illuminahisec_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data"),
  workspace = "TCGA_COAD_OpenAccess_V1-0_DATA",
  sampleCode = NULL,
  sampleIdx = 1:4,
  split = FALSE
)
```

terraTCGAworkspace	<i>Obtain or set the Terra Workspace Project Dataset</i>
--------------------	--

Description

Terra allows access to about 71 open access TCGA datasets. A dataset workspace can be set using the terraTCGAworkspace function with a projectName input. Use the selectTCGAworkspace function to select a TCGA data workspace from an interactive table.

Usage

```
terraTCGAworkspace(projectName = getOption("terraTCGAdata.workspace", NULL))

selectTCGAworkspace(
  projectName = getOption("terraTCGAdata.workspace", NULL),
  verbose = FALSE,
  ...
)
```

Arguments

projectName	character(1) A project code usually in the form of TCGA_CODE_OpenAccess_V1-0_DATA. See selectTCGAspace to interactively select from a table of project codes.
verbose	logical(1) Whether to provide more informative messages when an the "terraTCGAspace.workspace" option is set.
...	further arguments passed down to lower level functions, not intended for the end user.

Details

Note that GDC workspaces are not supported and are excluded from the search results. GDC workspaces use a Terra workflow to download TCGA data rather than providing Google Bucket storage locations for easy data retrieval. To reset the terraTCGAspace, use terraTCGAspace(NULL) and you will be prompted to select from a list of TCGA workspaces. You may also check the current active workspace by running terraTCGAspace() without any inputs.

Value

A Terra TCGA Workspace name

Functions

- selectTCGAspace(): Function to interactively select from the available TCGA data workspaces in Terra. The 'projectName' argument and 'terraTCGAspace.workspace' option must be 'NULL' to enable the interactive gadget.

Examples

```
if (
  AnVILGCP::gcloud_exists() &&
  identical(AnVILBase::avplatform_namespace(), "AnVILGCP") &&
  nzchar(AnVILGCP::avworkspace_name())
) {
  selectTCGAspace()
  terraTCGAspace()
}
```

Index

* **internal**

.getWorkspaceTable, 2

.getWorkspaceTable, 2

getAssayData, 2

getAssayTable, 4

getAssayTable(), 3

getClinical, 5

getClinicalTable, 6

getTCGAdatalist, 7

sampleTypesTable, 8

selectTCGAspace

(terraTCGAspace), 10

terraTCGAspace, 9

terraTCGAspace, 10