Package ‘terraTCGAdatata’

March 9, 2024

Type Package
Title OpenAccess TCGA Data on Terra as MultiAssayExperiment
Version 1.6.0
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 R (>= 4.2.0), AnVIL, MultiAssayExperiment
biocViews Software, Infrastructure, DataImport
Imports BiocFileCache, dplyr, GenomicRanges, methods, RaggedExperiment, readr, S4Vectors, stats, tidyr, TCGAutils, utils
Suggests 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 7.2.3
Config/testthat/edition 3
git_url https://git.bioconductor.org/packages/terraTCGAdata
git_branch RELEASE_3_18
git_last_commit 4cdd117
git_last_commit_date 2023-10-24
Repository Bioconductor 3.18
Date/Publication 2024-03-08
Author Marcel Ramos [aut, cre] (<https://orcid.org/0000-0002-3242-0582>)
Maintainer Marcel Ramos <marcel.ramos@roswellpark.org>
R topics documented:

.getWorkspaceTable .................................................. 2
getAssayData ............................................................. 3
getAssayTable ............................................................ 4
getClinical ................................................................. 5
getClinicalTable .......................................................... 6
getTCGAdatalist ........................................................... 7
sampleTypesTable .......................................................... 8
terraTCGAdata .............................................................. 9
terraTCGAworkeplace ...................................................... 10

Index 12

---

.getWorkspaceTable  Obtain the table of datasets from the Terra platform

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**

**Obtain assay datasets from Terra**

---

**Description**

Obtain assay datasets from Terra

**Usage**

```r
getAssayData(
  assayName,
  sampleCode = "01",
  tablename = .DEFAULT_TABLENAME,
  workspace = terraTCGAworkspace(),
  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 `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.
- **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()`
getAssayTable

Obtain a reference table for assay data in the Terra data model

Description

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

Usage

getAssayTable(
  tablename = .DEFAULT_TABLENAME,
  metacols = .PARTICIPANT_METADATA_COLS,
  workspace = terraTCGAworkspace(),
  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 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.

Value

A tibble of pointers to resources within the Terra data model

Examples

if (AnVIL::gcloud_exists())
  getAssayTable(workspace = "TCGA_COAD_OpenAccess_V1-0_DATA")
getClinical  Obtain clinical data

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

Usage
getClinical(
  columnName,
  participants = TRUE,
  tablename = .DEFAULT_TABLENAME,
  workspace = terraTCGAworkspace(),
  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 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.
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.
getClinicalTable

Obtain the reference table for clinical data

Description

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

Usage

```r
getClinicalTable(
  tablename = .DEFAULT_TABNAME,
  metacols = .PARTICIPANT_METADATA_COLS,
  workspace = terraTCGAworkspace(),
  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 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.
- **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...
getTCGAdatalist

Examples

```r
if (AnVIL::gcloud_exists())
  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**

```r
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.
sampleTypesTable

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 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")

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 (AnVIL::gcloud_exists())
  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"
  )
Examples

```r
if (AnVIL::gcloud_exists())
    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**

```r
terraTCGAdata(
    clinicalName,
    assays,
    participants = TRUE,
    sampleCode = NULL,
    split = FALSE,
    workspace = terraTCGAworkspace(),
    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 `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.
**terraformworkspace**

**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**

```r
if (AnVIL::gcloud_exists())
  terraTCGAdata(
    clinicalName = "clin_bio_nationwidechildrens_org_Level_1_biospecimen_clin",
    assays = c("protein_exp_mda_rppa_core_mdanderson_org_Level_3_protein_normalization_data",
               "rnaseq2_illuminahiseq_rnaseq2_unc_edu_Level_3_RSEM_genes_normalized_data"),
    workspace = "TCGA_COAD_OpenAccess_V1-0_DATA",
    sampleCode = NULL,
    sampleIdx = 1:4,
    split = FALSE
  )
```

**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**

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

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

Arguments

`projectName` character(1) A project code usually in the form of TCGA_CODE_OpenAccess_V1-0_DATA. See `selectTCGAworlpace` to interactively select from a table of project codes.

`verbose` logical(1) Whether to provide more informative messages when the "terraTCGAdata.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 `terraTCGAworlpace`, use `terraTCGAworlpace(NULL)` and you will be prompted to select from a list of TCGA workspaces. You may also check the current active workspace by running `terraTCGAworlpace()` without any inputs.

Value

A Terra TCGA Workspace name

Functions

- `selectTCGAworlpace()`: Function to interactively select from the available TCGA data workspaces in Terra. The ‘projectName’ argument and ‘terraTCGAdata.workspace’ option must be ‘NULL’ to enable the interactive gadget.

Examples

```r
if (AnVIL::gcloud_exists() && interactive()) {
  selectTCGAworlpace()
  terraTCGAworlpace()
}
```
Index

* internal
  .getWorkspaceTable, 2
  .getWorkspaceTable, 2
  getAssayData, 3
  getAssayTable, 4
  getAssayTable(), 3
  getClinical, 5
  getClinicalTable, 6
  getTCGAdatalist, 7
  sampleTypesTable, 8
  selectTCGAwkspacex
    (terraTCGAwkspacex), 10
  terraTCGAdata, 9
  terraTCGAwkspacex, 10