Package: CPTACdata (via r-universe)

November 18, 2026

Title Append CPTAC Proteomics Data to MultiAssayExperiment Objects

Version 0.99.10

Description Integrates proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) into existing MultiAssayExperiment objects containing The Cancer Genome Atlas (TCGA) patient barcodes. The package queries the Proteomic Data Commons (PDC) GraphQL API to resolve studies by disease type and fetch quantitative protein log2 ratio matrices and biospecimen identifiers. It harmonizes TCGA barcodes with CPTAC case identifiers, constructs SummarizedExperiment objects whose colData holds the matched biospecimen identifiers and sample types, and appends the proteomic assays to the MultiAssayExperiment. It also downloads the CPTAC-3 and CPTAC-2 Prospective cohorts, which use their own patients, as MultiAssayExperiment objects with one experiment per PDC study and the PDC clinical data as colData.

License Artistic-2.0

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

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

biocViews Software, Proteomics, BiomedicalInformatics, Network

Depends R (>= 4.5.0)

Imports BiocBaseUtils, httr2, methods, MultiAssayExperiment, S4Vectors, stats, SummarizedExperiment, TCGAutils

Suggests BiocStyle, curatedTCGAData, knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

Encoding UTF-8

Roxygen list(markdown = TRUE)

Config/roxygen2/version 8.1.1

Config/pak/sysreqs make libbz2-dev libicu-dev liblzma-dev libpng-dev libxml2-dev libssl-dev libx11-dev xz-utils zlib1g-dev

Repository https://biocstaging.r-universe.dev

Date/Publication 2026-11-16 20:54:31 UTC

RemoteUrl https://github.com/BiocStaging/CPTACdata

RemoteRef HEAD

RemoteSha 61b98f1300c09ea3ba2ad2a27239aaa198f0da4f

Contents

appendCPTAC 2
cptacCohorts 4
cptacProjects 6
getCPTAC 7

Index **10**

`appendCPTAC` *Append CPTAC proteomics from the PDC to a TCGA MultiAssayExperiment*

Description

Takes a `MultiAssayExperiment` from `curatedTCGAData`, finds the CPTAC studies in the Proteomic Data Commons (PDC) that were run on the same TCGA tumors, and adds them as new experiments with a correct `sampleMap`.

Usage

```
appendCPTAC(
  mae,
  project = NULL,
  fraction = c("Proteome", "Phosphoproteome", "Glycoproteome"),
  replicates = c("all", "first"),
  dry_run = FALSE
)
```

Arguments

`mae`	A `MultiAssayExperiment` for a single TCGA project, such as one from `curatedTCGAData`.
`project`	TCGA project, e.g. `"OV"` or `"TCGA-OV"`. If `NULL`, it is read from the `admin.disease_code` column of `colData(mae)`.
`fraction`	The assay type to add: one of `"Proteome"` (default), `"Phosphoproteome"` or `"Glycoproteome"`. Call `appendCPTAC` once per type to add several.

`replicates` How to handle technical replicates, which the PDC marks with a `.1`, `.2` suffix on the matrix column. `"all"` (default) keeps every replicate as its own column; `"first"` keeps only the first replicate of each sample.

`dry_run` If `TRUE`, download only sample metadata and return a one-row-per-study summary of what would be added.

Details

Only patients already in `colData(mae)` are added. Specimens are kept if they carry a TCGA barcode, are "Qualified" and not pooled. Each new experiment has one column per column of the PDC matrix, including every technical replicate unless `replicates = "first"`. Missing values are `NA`. The quantification is the PDC Common Data Analysis Pipeline value given by the `data_type` column of cptacProjects (log2 ratios to a common reference).

Biospecimen tables are cached for the R session, so a `dry_run` followed by a real run downloads them only once.

Value

If `dry_run = FALSE`, the input `mae` with one new experiment per PDC study, named as in the `experiment` column of cptacProjects. Each experiment is a `SummarizedExperiment` with:

- One assay, named after the `data_type` of the study (e.g. `"log2_ratio"`), holding the PDC quantification matrix for the matched samples. Rows are genes or protein sites, and missing values are `NA`.
- Columns named by the PDC matrix label, with `TCGA-` added where the PDC leaves it out. Proteome labels are sample barcodes (e.g. `TCGA-BH-A18U-01A`) and phosphoproteome labels are aliquot barcodes (e.g. `TCGA-24-1555-01A-32-2224-29`). The PDC marks technical replicates with a `.1`, `.2` suffix, which is kept (e.g. `TCGA-C8-A131-01A.1` and `TCGA-C8-A131-01A.2`).
- `colData` with `case_submitter_id` (TCGA patient barcode), `sample_submitter_id` (TCGA sample barcode, shared by the replicates of a sample), `aliquot_submitter_id`, `sample_type`, and `pdc_column` (the column label in the PDC matrix).
- `metadata` with `pdc_study_id`, `pdc_name`, `data_type`, `source` (the PDC GraphQL URL), `pdc_study_version` and `pdc_study_uuid` (the PDC version of the study and its identifier, which pins that version), `retrieved` (the time the matrix was downloaded) and `CPTACdata_version` (the package version that downloaded it). Read them with `S4Vectors::metadata(mae[["CPTAC_Proteome"]])`.

The new `sampleMap` links every new column to its TCGA patient. Studies that are skipped, because they have no matrix, no matching patients or are already in `mae`, add no experiment and give a warning or message.

If `dry_run = TRUE`, a `data.frame` with one row per study in cptacProjects for the requested project and fraction, and these columns:

- `experiment`: experiment name the study would be added as.
- `pdc_study_id`: PDC study ID.
- `quant_available`: `TRUE` if the PDC has a quantification matrix for the study.
- `mae_patients`: number of patients in `colData(mae)`. Same value for every row.

- pdc_patients: number of distinct TCGA patients in the PDC study that pass the specimen filters. NA if there is no matrix.
- matched_patients: number of those patients that are also in colData(mae). NA if there is no matrix.

See Also

cptacProjects

Examples

```
## The example queries the PDC API, so it runs only after you have accepted
## the PDC Data Use Agreement with options(CPTACdata.acceptDUA = TRUE).
## A small object stands in for a curatedTCGAData object here.
if (isTRUE(getOption("CPTACdata.acceptDUA"))) {
patients <- c("TCGA-BH-A18U", "TCGA-BH-A0C0", "TCGA-A8-A08Z")
mae <- MultiAssayExperiment::MultiAssayExperiment(
    experiments = list(
        RNA = SummarizedExperiment::SummarizedExperiment(
            matrix(
                0, nrow = 1L, ncol = 3L,
                dimnames = list("gene", patients)
            )
        )
    ),
    colData = S4Vectors::DataFrame(
        admin.disease_code = rep("brca", 3L), row.names = patients
    ),
    sampleMap = S4Vectors::DataFrame(
        assay = factor(rep("RNA", 3L)),
        primary = patients,
        colname = patients
    )
)
appendCPTAC(mae, fraction = "Proteome", dry_run = TRUE)
mae <- appendCPTAC(mae, fraction = "Proteome")
}
```

---

cptacCohorts *CPTAC cohorts in the PDC*

Description

The CPTAC patient cohorts in the Proteomic Data Commons (PDC) that getCPTAC() can download, with one row per PDC study. A cohort, such as the CPTAC-3 clear cell renal cell carcinoma discovery cohort, is profiled in several PDC studies, one for each experiment (proteome, phosphoproteome, acetylome, ...). getCPTAC() combines them in one MultiAssayExperiment.

Usage

```
data(cptacCohorts)
```

Format

A `data.frame` with one row per PDC study and 8 columns:

- `cohort`: Cohort name, the PDC study name without the experiment (e.g. `"CPTAC CCRCC Discovery Study"`), or a curated name for the CPTAC-2 Prospective studies (e.g. `"Prospective Breast"`).
- `experiment`: Name of the experiment in the `MultiAssayExperiment`, from the rest of the study name (e.g. `"Proteome"`, `"DIA_Proteome"`).
- `pdc_study_id`: PDC study ID.
- `pdc_name`: PDC study name.
- `pdc_project`: PDC project.
- `analytical_fraction`: PDC analytical fraction, e.g. Proteome or Phosphoproteome.
- `experiment_type`: PDC experiment type, e.g. `"TMT10"` or `"Label Free"`.
- `data_type`: `quantDataMatrix` data type that is downloaded; `NA` if the PDC has no matrix for the study.

Details

The table covers the PDC projects "CPTAC3 Discovery and Confirmatory" and "CPTAC-2" (the Prospective collections). Studies of reference samples only ("CompRef") and method studies are left out. Cohorts run on TCGA tumors are in cptacProjects and are added to TCGA data with `appendCPTAC()`.

Some studies, such as most metabolome and lipidome studies, have no quantification matrix available through the PDC API. They are listed with `data_type = NA` and are skipped by `getCPTAC()`.

Source

Built from the PDC GraphQL API (`https://pdc.cancer.gov/graphql`) by `data-raw/cptacCohorts.R` in the package sources. The cohort and experiment names of the CPTAC-2 studies, and the excluded studies, are curated in that script.

Examples

```
data("cptacCohorts", package = "CPTACdata")
unique(cptacCohorts[["cohort"]])
cptacCohorts[
    cptacCohorts[["cohort"]] == "CPTAC CCRCC Discovery Study",
    c("experiment", "pdc_study_id", "data_type")
]
```

cptacProjects *TCGA projects with matching CPTAC proteomics in the PDC*

Description

A crosswalk between TCGA projects (e.g. `"TCGA-BRCA"`) and the Proteomic Data Commons (PDC) studies holding CPTAC data generated on the *same TCGA tumors* (the PDC "CPTAC-TCGA" project). It is used by `appendCPTAC()` to select studies.

Usage

```
data(cptacProjects)
```

Format

A `data.frame` with one row per TCGA project and PDC study (8 rows) and 8 columns:

- `tcga_project`: TCGA project ID.
- `pdc_study_id`: PDC study ID.
- `pdc_name`: PDC study name.
- `experiment`: Name given to the experiment added to the `MultiAssayExperiment`.
- `analytical_fraction`: Proteome, Phosphoproteome or Glycoproteome.
- `data_type`: `quantDataMatrix` data type that is downloaded; `NA` if the PDC has no matrix for the study.
- `pdc_disease_type`: PDC disease label used to split studies that span several TCGA projects; `NA` if not needed.
- `verified`: `TRUE` if every qualified TCGA patient in the study was found in the GDC project when the data were built.

Details

Cancer types profiled by CPTAC-3 (e.g. LUAD, ccRCC) and the "Prospective" CPTAC-2 collections are not included: those cohorts use new patients rather than TCGA samples, so there is nothing to match against a TCGA MultiAssayExperiment.

One study can cover more than one TCGA project (PDC000111 holds colon and rectum tumors); `pdc_disease_type` separates them. The JHU and PNNL ovarian proteomes share 33 patients and are separate experiments.

Some TCGA-derived studies (the colon proteome, PDC000111, and the ovarian glycoproteome, PDC000112) have biospecimen records in the PDC but no quantification matrix available through its API. They are listed with `data_type = NA` and are skipped by `appendCPTAC()`.

Source

Built from the PDC GraphQL API (`https://pdc.cancer.gov/graphql`) and the GDC API (`https://api.gdc.cancer.gov`) by `data-raw/cptacProjects.R` in the package sources. Experiment names and data types are curated in that script.

Examples

```
data("cptacProjects", package = "CPTACdata")
cptacProjects
```

`getCPTAC` *Get a CPTAC cohort from the PDC as a MultiAssayExperiment*

Description

Downloads the studies of one CPTAC cohort listed in cptacCohorts, such as the CPTAC-3 clear cell renal cell carcinoma discovery cohort, and combines them in a `MultiAssayExperiment`, with one experiment per study and the PDC clinical data of the cohort's patients as `colData`.

Usage

```
getCPTAC(
  cohort,
  experiments = NULL,
  replicates = c("all", "first"),
  dry_run = FALSE
)
```

Arguments

`cohort`	The cohort to download, a value of the `cohort` column of cptacCohorts, e.g. `"CPTAC CCRCC Discovery Study"`.
`experiments`	The experiments to include, values of the `experiment` column of cptacCohorts for that cohort, e.g. `c("Proteome", "Phosphoproteome")`. `NULL` (default) includes every study of the cohort with a quantification matrix.
`replicates`	How to handle technical replicates, which the PDC marks with a `.1`, `.2` suffix on the matrix column. `"all"` (default) keeps every replicate as its own column; `"first"` keeps only the first replicate of each sample.
`dry_run`	If `TRUE`, download only sample metadata and return a one-row-per-study summary of what would be downloaded.

Details

Specimens are kept if they are "Qualified", not pooled, and come from a patient: reference channels such as the internal reference pool, QC samples and the NCI-7 cell lines (PDC disease type "Other") are dropped. Columns are named by the PDC matrix label, which for these cohorts is the `aliquot_submitter_id`; technical replicates keep their `.1`, `.2` suffix. The quantification is the PDC Common Data Analysis Pipeline value given by the `data_type` column of cptacCohorts.

Biospecimen and clinical tables are cached for the R session, so a `dry_run` followed by a real run downloads them only once.

Value

If dry_run = FALSE, a MultiAssayExperiment with:

- One experiment per study, named as in the experiment column of cptacCohorts. Each is a SummarizedExperiment with one assay, named after the data_type of the study, holding the PDC quantification matrix (missing values are NA). Its colData has case_submitter_id, sample_submitter_id, aliquot_submitter_id, sample_type and pdc_column (the column label in the PDC matrix), and its metadata has the same provenance fields as the experiments added by appendCPTAC().
- colData with one row per patient (case_submitter_id), holding the PDC disease type, demographics (e.g. gender, vital_status, days_to_birth) and diagnosis (e.g. primary_diagnosis, tumor_grade, ajcc_pathologic_stage, age_at_diagnosis). Values are kept as the PDC reports them, as character. When a patient has several diagnosis records, their distinct values are joined with "; ".
- A sampleMap linking every column to its patient.

If dry_run = TRUE, a data.frame with one row per selected study and these columns:

- experiment: experiment name.
- pdc_study_id: PDC study ID.
- quant_available: TRUE if the PDC has a quantification matrix for the study.
- patients: number of distinct patients in the study that pass the specimen filters.
- aliquots: number of their aliquots.

See Also

cptacCohorts, appendCPTAC()

Examples

```
## Cohort and experiment names come from cptacCohorts, which ships with
## the package and needs no network access
data("cptacCohorts", package = "CPTACdata")
unique(cptacCohorts[["cohort"]])

## Experiments of one cohort; those with data_type NA have no matrix
ccrccStudies <- cptacCohorts[
    cptacCohorts[["cohort"]] == "CPTAC CCRCC Discovery Study",
    c("experiment", "pdc_study_id", "data_type")
]
ccrccStudies

## Downloading queries the PDC API, so it runs only after you have
## accepted the PDC Data Use Agreement with
## options(CPTACdata.acceptDUA = TRUE).
if (isTRUE(getOption("CPTACdata.acceptDUA"))) {
getCPTAC("CPTAC CCRCC Discovery Study", dry_run = TRUE)
ccrcc <- getCPTAC(
    "CPTAC CCRCC Discovery Study",
```

```
    experiments = c("Proteome", "Phosphoproteome")
)
ccrcc
}
```

Index

∗ **datasets**
cptacCohorts, 4
cptacProjects, 6

appendCPTAC, 2
appendCPTAC(), *5, 6*, *8*

cptacCohorts, 4, *7, 8*
cptacProjects, *3–5*, 6

getCPTAC, 7
getCPTAC(), *4, 5*