# CPTAC Data Portal & Proteomic Data Commons

*Imaging SIG*

**Ratna Rajesh Thangudu, Ph.D.**

July 1, 2019

# Outline

- CPTAC Data Coordinating Center
- Proteomic Data Commons

# CPTAC Data Coordinating Center

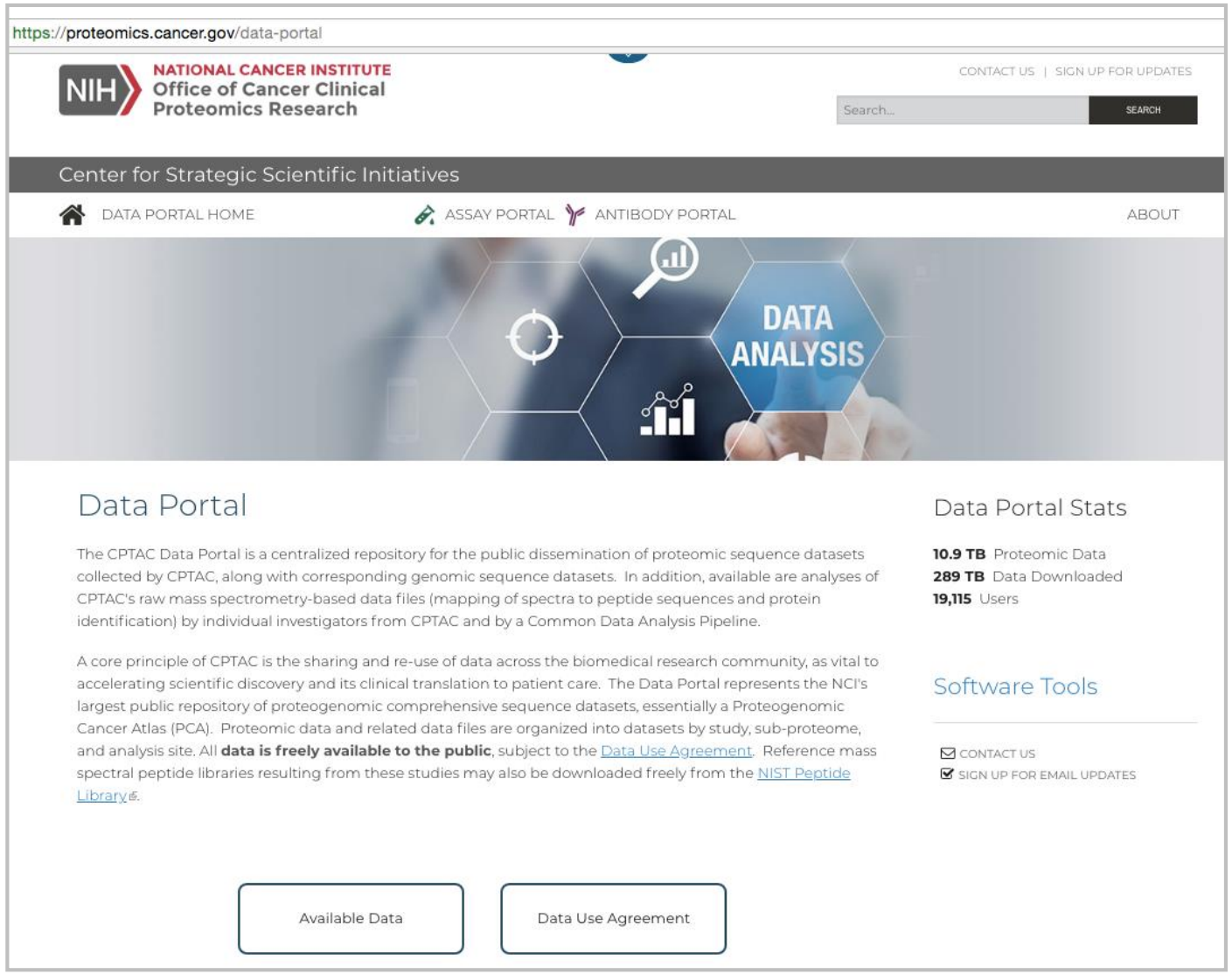
**CPTAC Public Portal (https://proteomics.cancer.gov/data-portal)**

- A centralized repository for the public dissemination of CPTAC proteomic datasets
- Analyze all of the CPTAC data through a Common Data Analysis Pipeline (CDAP) for public release
- Enable high speed transfer through UDP technology (Aspera)
- Provide support to the user community

**CPTAC Assay Portal (https://proteomics.cancer.gov/assay-portal)**

- The CPTAC Assay Portal serves as a centralized public repository of "fit-for-purpose," multiplexed quantitative mass spectrometry-based proteomic targeted assays.

# CPTAC Public Data Portal




https://proteomics.cancer.gov/data-portal

# CPTAC Public Data Portal

## Portal Summary

Overview of CPTAC data available, number of portal visitors, and amount of proteomics data downloaded

### Summary Statistics Jul-01-2019

**CPTAC Data Available**

20.51 TB
Amount of Data

| | |
|---|---|
| Amount of Data (in TB) | 20.51 |
| Number of Files | 85,099 |

View More »

**Portal Visits**

438,454
Number of Page Views

| | |
|---|---|
| Total Number of Data Portal Page Views | 438,454 |
| Total Number of Sessions | 141,318 |
| Total Number of Users | 70,916 |
| Returning Visits | 70,402 |
| Distinct countries | 119 |

View More »

**Total Data Downloaded**

363.31 TB
Amount of Data

| | |
|---|---|
| Amount of Data (in TB) | 363.31 |
| Number of Files | 1,851,555 |

View More »

# CPTAC Data Portal: Releases

| Study Name | Description | Publications |
|---|---|---|
| Pediatric Brain Cancer Pilot Study<br>new | A pediatric brain cancer cohort of 199 patients was used for a proteogenomic pilot study. Global proteomic and phosphoproteomic mass spectrometry using the 11-plexed isobaric tandem mass tags (TMT-11) was used to characterize 219 brain tumor samples across seven histologies: Low Grade Glioma, High Grade Glioma, Ependymoma, Ganglioglioma, Craniopharyngioma, Atypical Teratoid Rhabdoid Tumor (ATRT), Medulloblastoma. (Twenty patients from the cohort of 199 had tumor samples from 2 clinical events, totaling 219 tumors) | |
| CPTAC LUAD Discovery Study<br>new | A Lung Adenocarcinoma (LUAD) discovery cohort of 111 tumor samples was analyzed by global proteomic and phosphoproteomic mass spectrometry using the 10-plexed isobaric tandem mass tags (TMT-10) following the CPTAC reproducible workflow protocol published by Mertins et al., (2018 Nature Protocols). This data release contains raw mass spectrometry data and analysis from the CPTAC Common Data Analysis Pipeline (CDAP). | |
| Colon Cancer Therapeutic Opportunities<br>new | Proteogenomic study on a prospectively collected colon cancer cohort with 110 paired tumor and normal adjacent tissues. | |
| Proteogenomics of Gastric Cancer | Proteogenomic analysis was performed on a cohort of 80 patients with early-onset gastric cancer recruited from the Korean population. | |
| CPTAC UCEC Discovery Study | A Uterine Corpus Endometrial Carcinoma (UCEC) discovery cohort of 100 tumor samples was analyzed by global proteomic and phosphoproteomic mass spectrometry using the 10-plexed isobaric tandem mass tags (TMT-10) following the CPTAC reproducible workflow protocol published by Mertins et al., (2018 Nature Protocols). This data release contains raw mass spectrometry data and analysis from the CPTAC Common Data Analysis Pipeline (CDAP). | |
| CPTAC CCRCC Discovery Study | Tumor samples from 110 patients with Clear Cell Renal Cell Carcinoma (CCRCC) were analyzed by global proteomic and phosphoproteomic mass spectrometry using the 10-plexed isobaric tandem mass tags (TMT-10) following the CPTAC reproducible workflow protocol published by Mertins et al., (2018 Nature Protocols). This data release contains raw mass spectrometry data and analysis from the CPTAC Common Data Analysis Pipeline (CDAP). | |

# CPTAC Data Portal: Releases

PRINT

## Data Use Agreement

*Data users must click "Accept" to access data (bottom of page)*

### Responsible Use of CPTAC Data

CPTAC requests that data users abide by the same principles that were previously established in the Fort Lauderdale and Amsterdam meetings. The recommendations from the Fort Lauderdale meeting (2003) on best practices and principles for sharing large-scale genomic data address the roles and responsibilities of data producers, data users and funders of community resource projects. The aim of the recommendations is to establish and maintain an appropriate balance between the interests that data users have in rapid access to data and the needs that data producers have to publish and receive recognition for their work. The conclusion of the attendees at the Fort Lauderdale meeting was that a "responsible use" approach for secondary data users would be sufficient to ensure that the efforts of data producers will be recognized. "Responsible use" was defined as allowing the data producers to have the opportunity to publish the initial global analyses of the data within a reasonable period of time.

In 2008, the National Cancer Institute OCCPR organized a workshop to discuss how and when proteomics data should be released. The result was the Amsterdam Principles, that established guidelines for the timing of data release, comprehensiveness of a dataset, data format, deposition to repositories, quality metrics, and responsibility for proteomic data release. Participants agreed that mass spectrometry output data files should be available to support the claims of proteomics publications. In 2010, the National Cancer Institute OCCPR convened a follow-on workshop to address quality metrics for proteomics, with an emphasis on mass spectrometry. As a sign of solidarity for these principles, four peer-reviewed journals simultaneously published the corollary to Amsterdam Principles.

Agreeing to abide by these principles and the CPTAC Publication Guidelines is required to gain access to CPTAC data.

### Publication Guidelines and Embargo Period

CPTAC is a community resource project and data are made available rapidly after generation for community research use. To act in accord with the Fort Lauderdale principles and support the continued prompt public release of large-scale proteomic data prior to publication, researchers who plan to prepare manuscripts or presentations involving CPTAC data and journal editors who receive such manuscripts, are encouraged to coordinate their independent reports with CPTAC's publication schedule. This may be done by contacting the CPTAC network at cancer.proteomics@mail.nih.gov.

CPTAC defines a global analysis publication as the first marker paper, authored by one or more members of the CPTAC, which includes analysis of the existing CPTAC data generated on the tumor type or sample set at the time of a data freeze. Specifically, these manuscripts report on the comprehensive, integrated analyses of multiple CPTAC datasets which may include: characterization of global proteome and/or PTMs such as phosphorylation, acetylation, glycosylation and ubiquitination, and reverse-phase protein analysis. Prior to a global analysis publication on a specific tumor type, available datasets should be considered pre-publication data subject to the standard principles of scientific etiquette regarding publication of findings using data obtained from other sources.

The CPTAC program has established the following policy to clarify freedom of CPTAC and non-CPTAC users to publish findings using CPTAC data. There are no limitations on submitting manuscripts to a journal and subsequent publications containing analyses using any CPTAC data set if the data set meets one of the following three freedom-to-publish criteria:

1. A global analysis publication paper has been published on that tumor type or sample set; or
2. 22 months (embargo period) after the final samples of a given tumor type has been delivered to the Proteome Characterization Center for that tumor type; or
3. The author or presenter receives specific approval from the CPTAC Steering Committee.

The specific status of each tumor dataset is displayed on the study page.

**Citing CPTAC Data in Publications**
The CPTAC program requests that publications using data from this program include the following statement:
"Data used in this publication were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC)."

If you have questions, do not hesitate to contact cancer.proteomics@mail.nih.gov.

Accept Decline

# CPTAC Data Portal: Releases

## Biospecimens

Clinical Data for CPTAC Ovarian Cancer Confirmatory Study
CPTAC Ovarian Cancer Confirmatory Study Specimens

## Data Types Available for Download

(ALL): Selection of this box downloads all data in the row
(raw): The original mass spectrometry(MS) instrument files
(mzML): HUPO-PSI standard raw data files generated from the original MS instrument files
(PSM): Peptide-Spectrum Match data
(prot): Protein assembly data and protein relative abundance
(meta): Clinical data files, mapping of biospecimens to iTRAQ labels or TMT10 labels (where applicable), folder and file naming conventions
Checksum files are included in all downloads for verification.

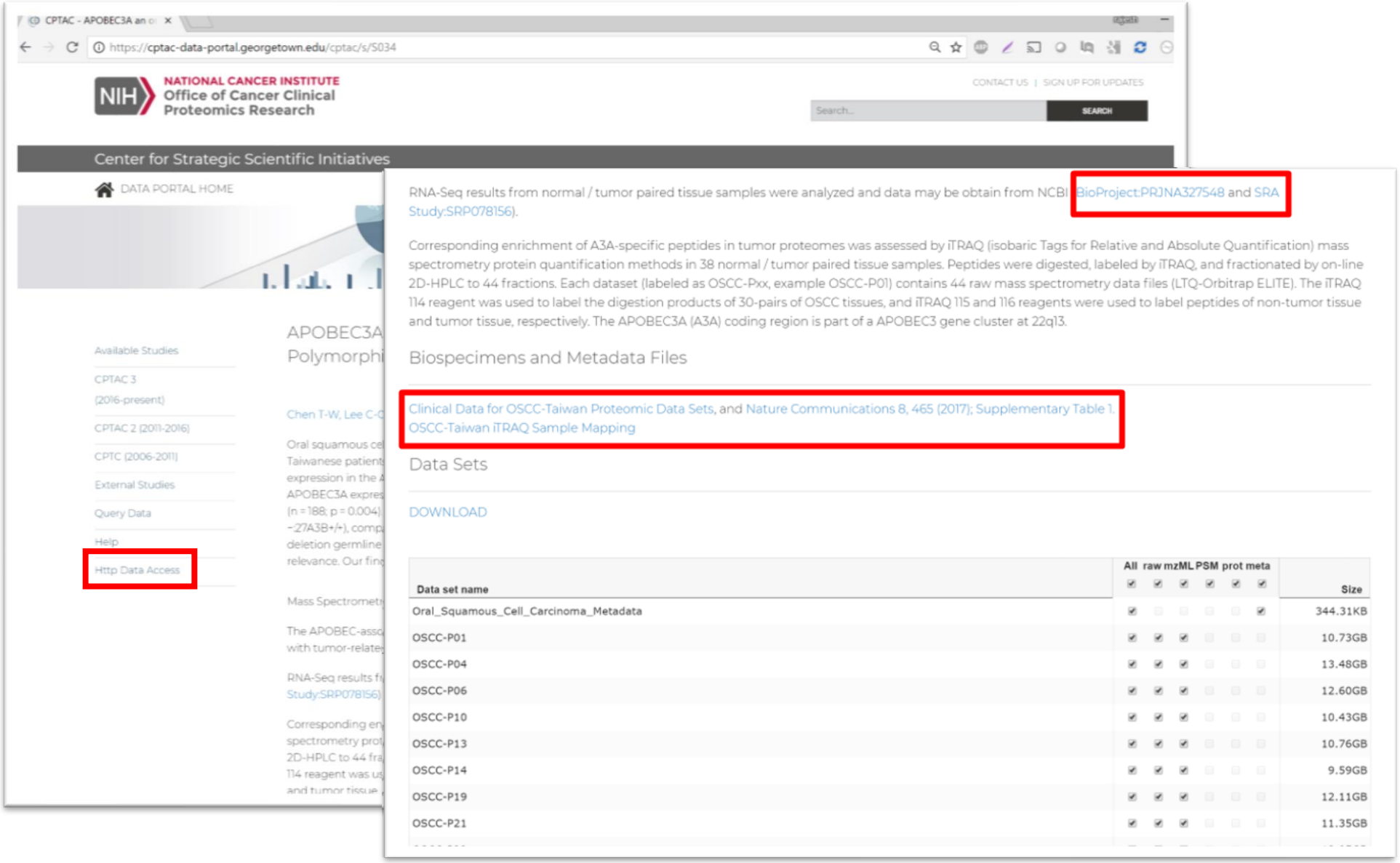
## Data Sets

DOWNLOAD

Mass Spectrometry Site: Select an Option  Analytical Fraction: Select an Option

| Data set name | All | raw | mzML | PSM | prot | meta | Size |
|---|---|---|---|---|---|---|---|
| CPTAC_OVprospective_metadata | ✓ | | | | | ✓ | 222.12KB |
| CPTAC_OVprospective_PNNL_Phosphoproteome_CDAP_Protein_Report.r1 | ✓ | | | | ✓ | ✓ | 111.34MB |
| 01CPTAC_OVprospective_Proteome_JHU_20161209 | ✓ | ✓ | ✓ | ✓ | | | 25.49GB |
| 02CPTAC_OVprospective_Proteome_JHU_20161104 | ✓ | ✓ | ✓ | ✓ | | | 25.67GB |
| 03CPTAC_OVprospective_Proteome_JHU_20161110 | ✓ | ✓ | ✓ | ✓ | | | 27.45GB |
| 04CPTAC_OVprospective_Proteome_JHU_20161111 | ✓ | ✓ | ✓ | ✓ | | | 26.02GB |

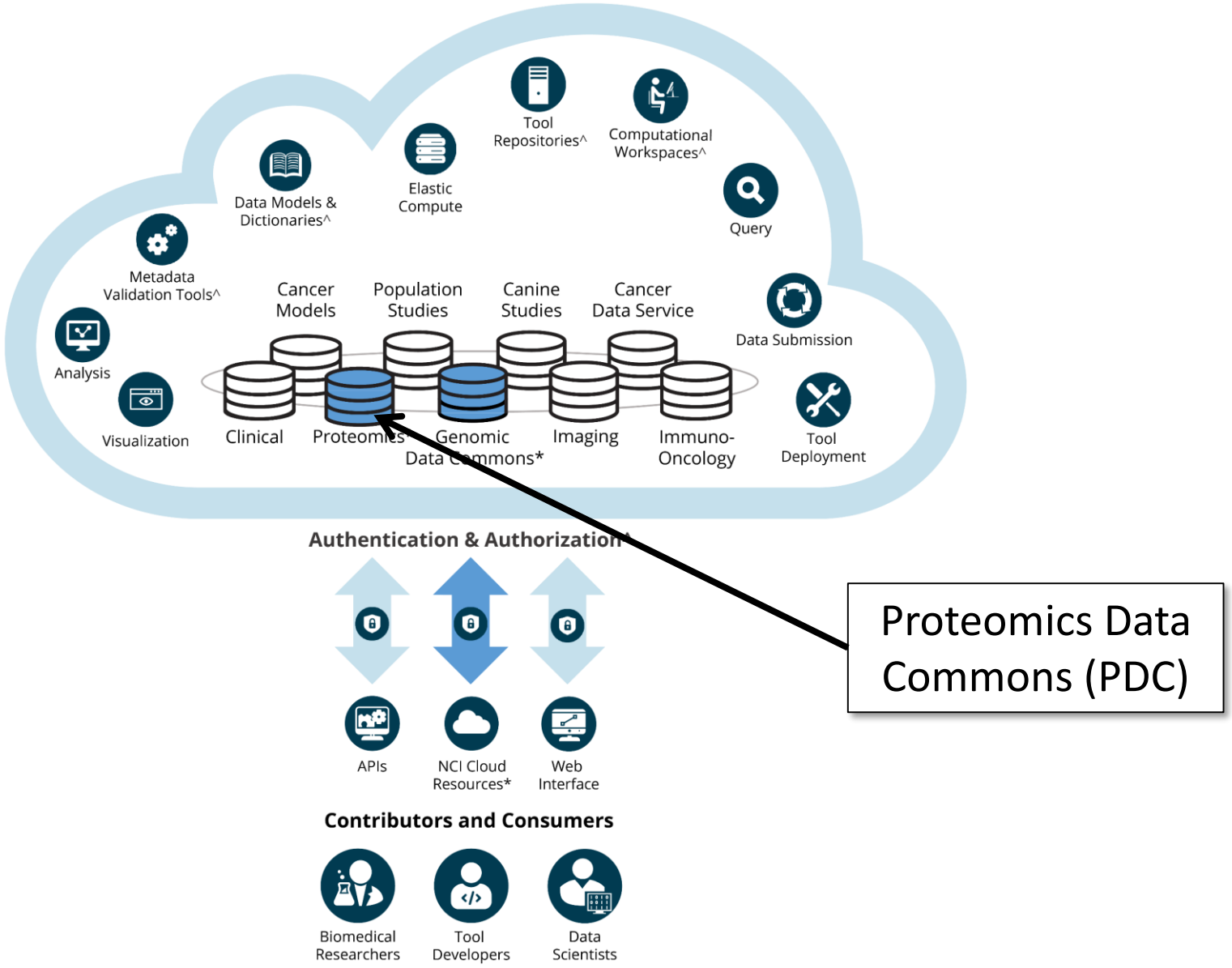
# CPTAC Data Portal: Releases



# Outline

- CPTAC Data Coordinating Center
- **Proteomic Data Commons**

# NCI Cancer Research Data Commons (CRDC)




* The Genomics Data Commons, Proteomics Data Commons, and NCI Cloud Resources are in production and available to the community

^ Components of the Data Commons Framework

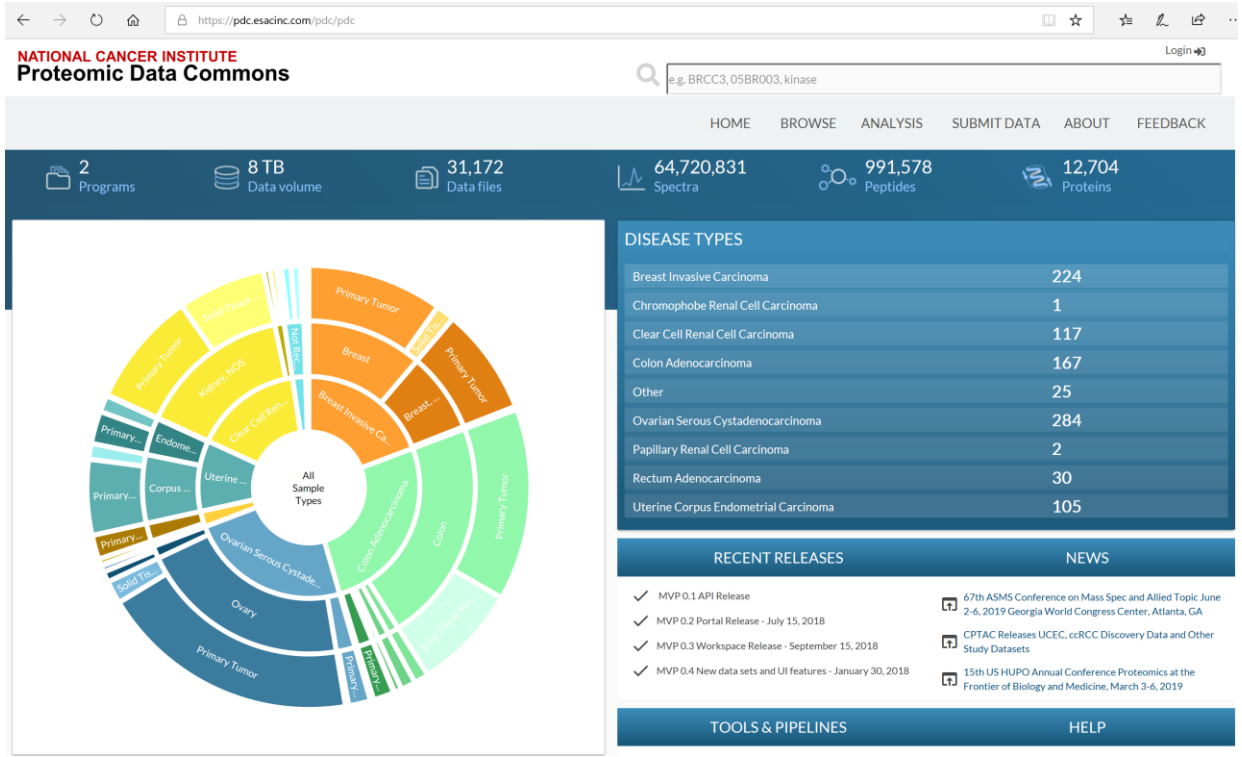
# Proteomics Data Commons: High Level Goals

- Unsilo mass spectrometry data.  Bring data into a common location that satisfy Findability, Accessibility, Interoperability and Reusability
- Move from a situation where people move data to local tools to where people move their tools to the data.
- Shift from a ‘data graveyard model’ to a ‘data workspace model’
- Make it feasible for pipelines to be released with data during publication to improve reproducibility
- Improve meta-data annotations. Ensure data is annotated well using common vocabularies but that the process is non-onerous.

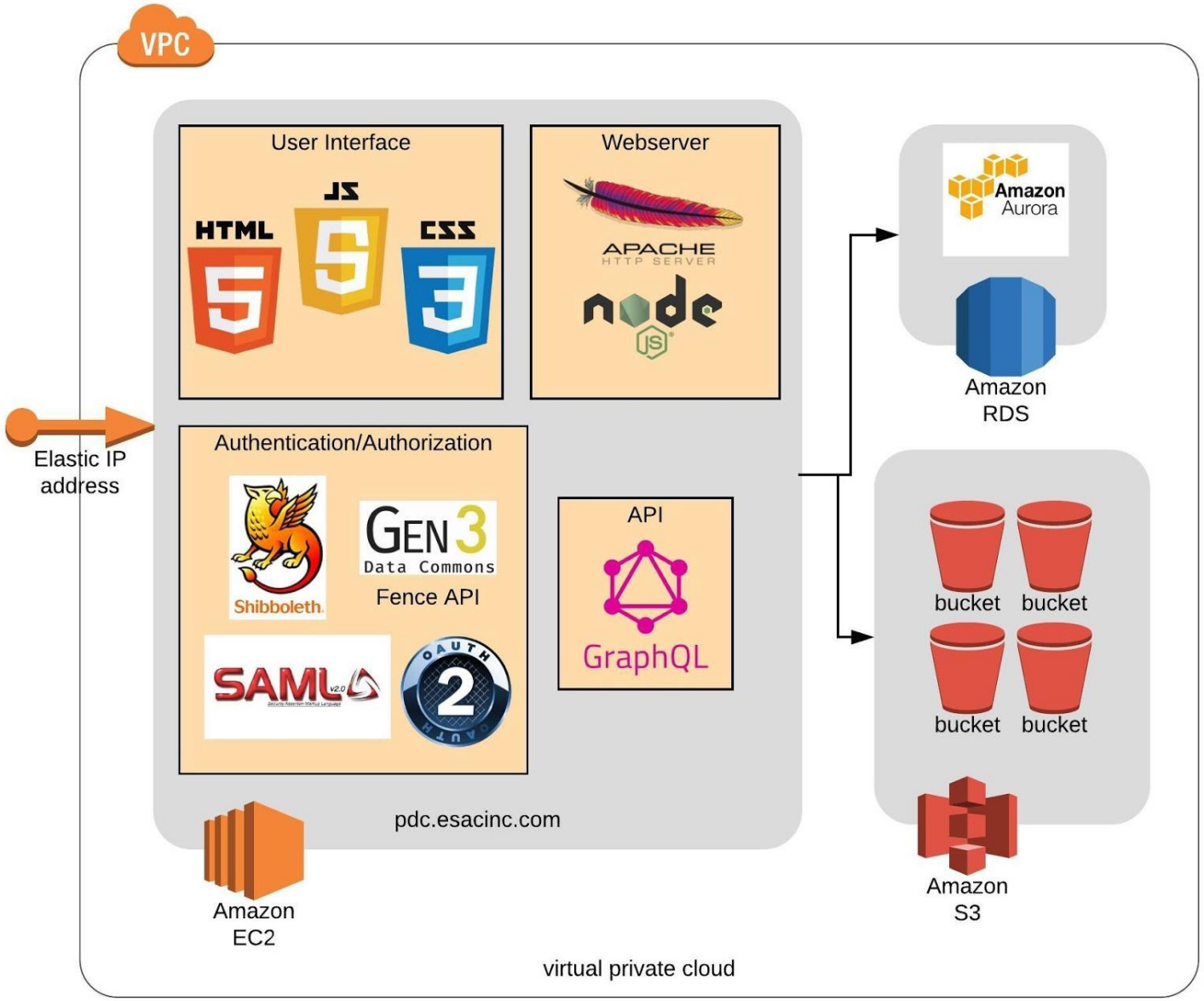
# Proteomic Data Commons: now available as beta program

https://pdc.esacinc.com

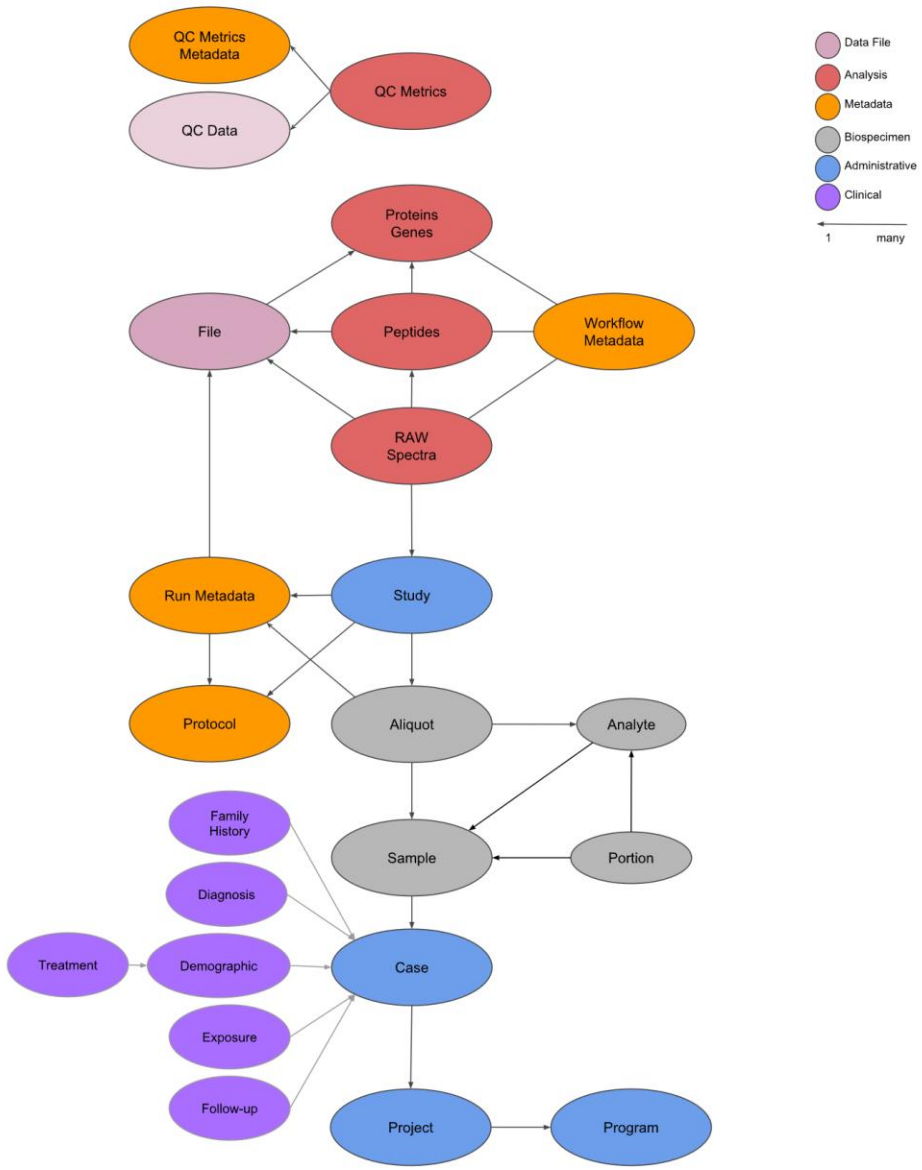
- Open to all users, no login required
- Enabled deep linking to connect from CPTAC data portal
- Released UCEC and CCRCC datasets from CPTAC3 program.
- Preparing to accept data from APOLLO, ICPC, etc



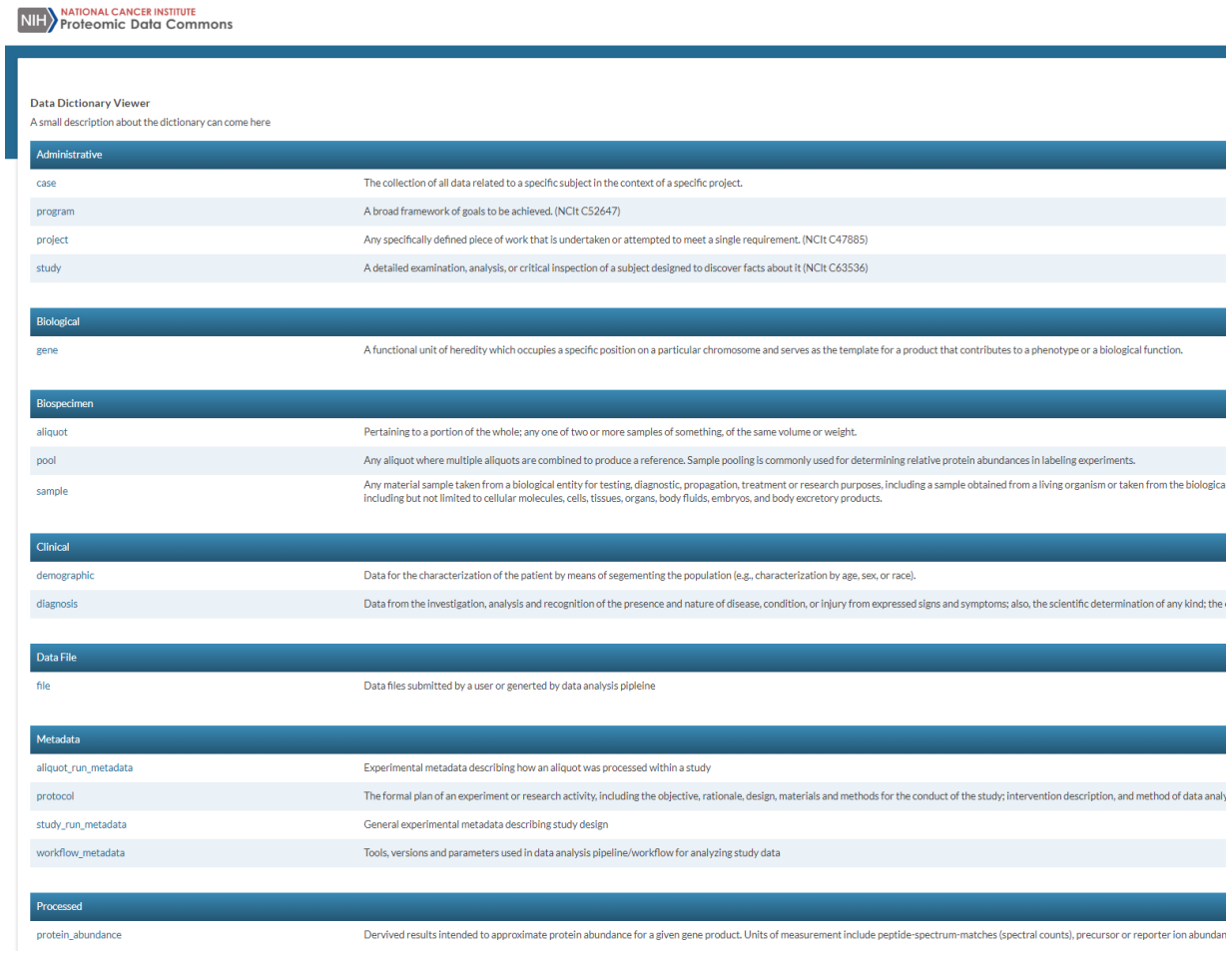
# PDC: Cloud architecture



# PDC: Data Model & Dictionary




- cancer Data Standards Registry and Repository (**caDSR**)
- NCI Thesaurus (NCIt)
- PSI MS Controlled Vocabularies and data formats



NIH NATIONAL CANCER INSTITUTE Proteomic Data Commons

**Data Dictionary Viewer**

A small description about the dictionary can come here

| Administrative | |
|---|---|
| case | The collection of all data related to a specific subject in the context of a specific project. |
| program | A broad framework of goals to be achieved. (NCIt C52647) |
| project | Any specifically defined piece of work that is undertaken or attempted to meet a single requirement. (NCIt C47885) |
| study | A detailed examination, analysis, or critical inspection of a subject designed to discover facts about it (NCIt C63536) |

| Biological | |
|---|---|
| gene | A functional unit of heredity which occupies a specific position on a particular chromosome and serves as the template for a product that contributes to a phenotype or a biological function. |

| Biospecimen | |
|---|---|
| aliquot | Pertaining to a portion of the whole; any one of two or more samples of something, of the same volume or weight. |
| pool | Any aliquot where multiple aliquots are combined to produce a reference. Sample pooling is commonly used for determining relative protein abundances in labeling experiments. |
| sample | Any material sample taken from a biological entity for testing, diagnostic, propagation, treatment or research purposes, including a sample obtained from a living organism or taken from the biologica including but not limited to cellular molecules, cells, tissues, organs, body fluids, embryos, and body excretory products. |

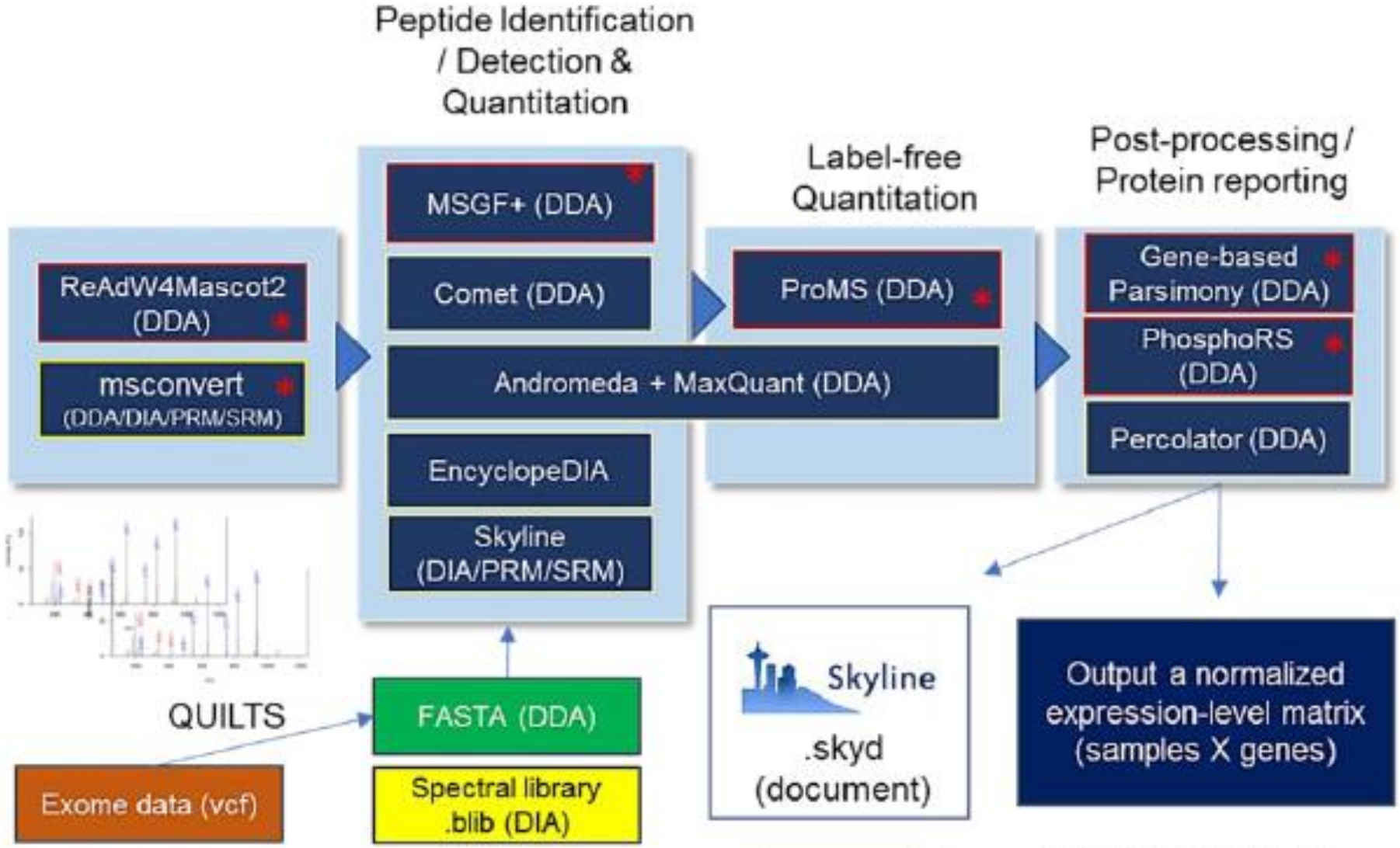
| Clinical | |
|---|---|
| demographic | Data for the characterization of the patient by means of segementing the population (e.g., characterization by age, sex, or race). |
| diagnosis | Data from the investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the |

| Data File | |
|---|---|
| file | Data files submitted by a user or generted by data analysis pipleine |

| Metadata | |
|---|---|
| aliquot_run_metadata | Experimental metadata describing how an aliquot was processed within a study |
| protocol | The formal plan of an experiment or research activity, including the objective, rationale, design, materials and methods for the conduct of the study; intervention description, and method of data analy |
| study_run_metadata | General experimental metadata describing study design |
| workflow_metadata | Tools, versions and parameters used in data analysis pipeline/workflow for analyzing study data |

| Processed | |
|---|---|
| protein_abundance | Dervived results intended to approximate protein abundance for a given gene product. Units of measurement include peptide-spectrum-matches (spectral counts), precursor or reporter ion abundan |

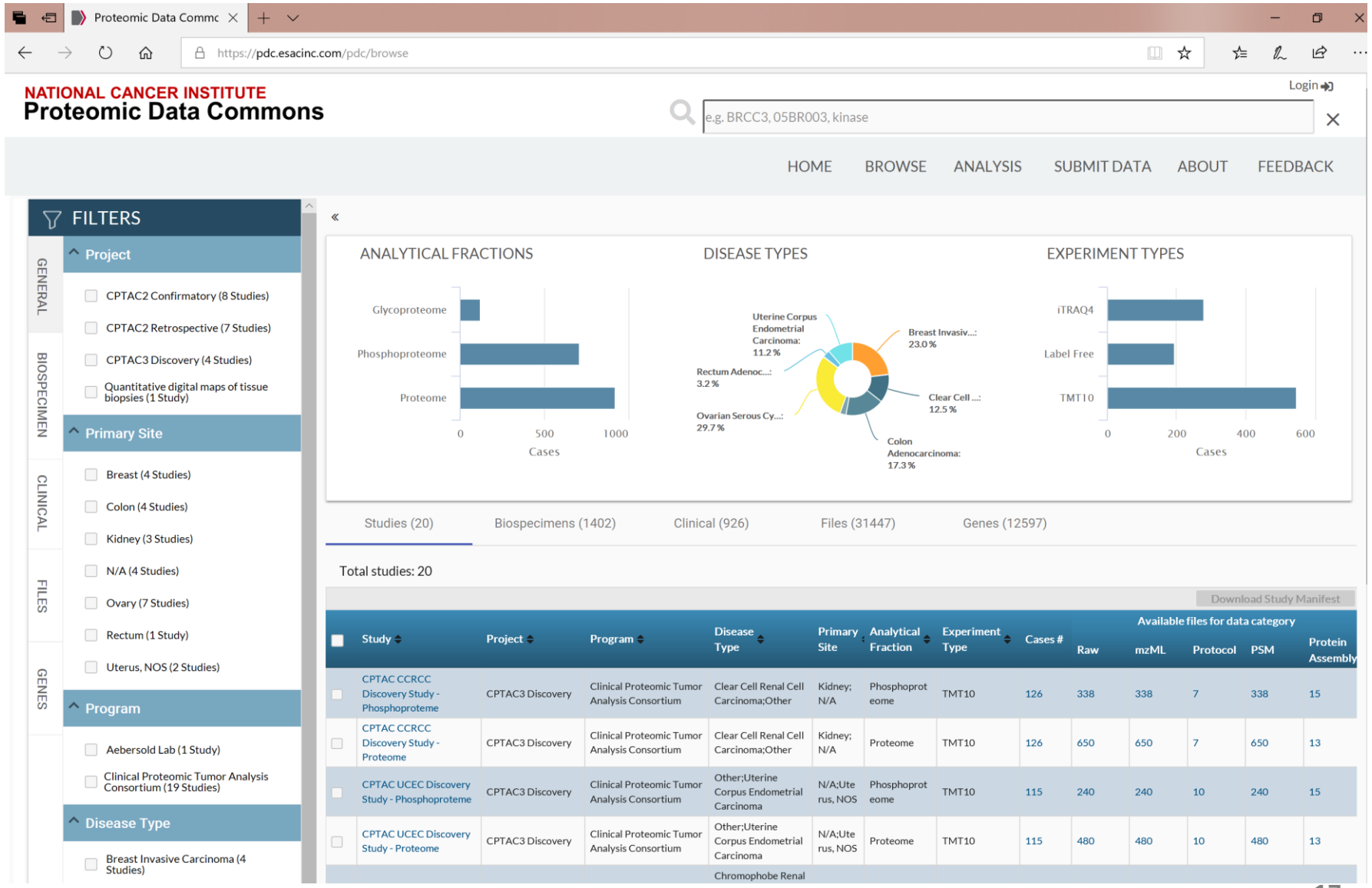
# Proteomics Data Commons: Common Data Analysis Pipeline



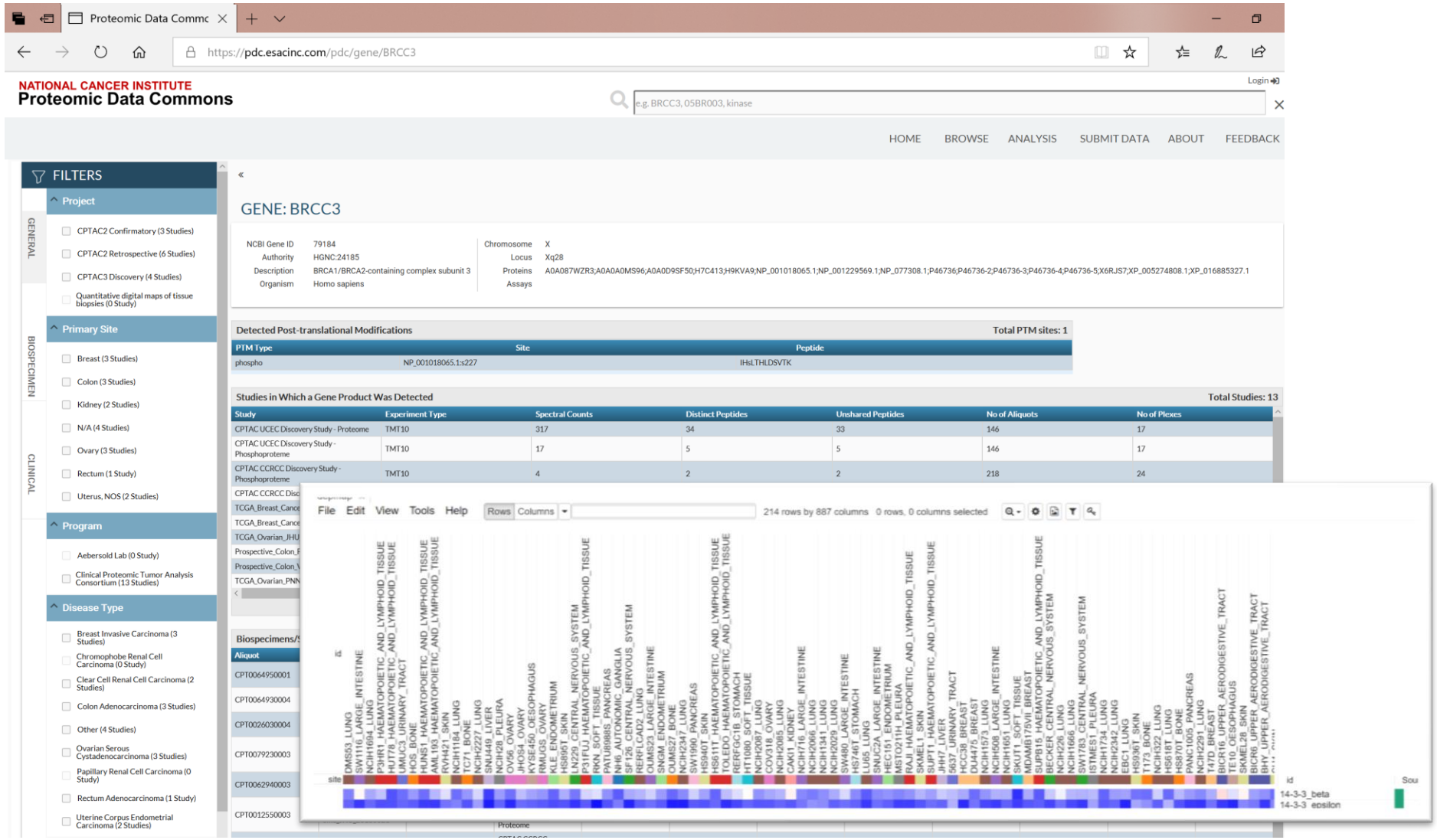

* component of current CPTAC CDAP Pipeline.

# PDC: Browse Data

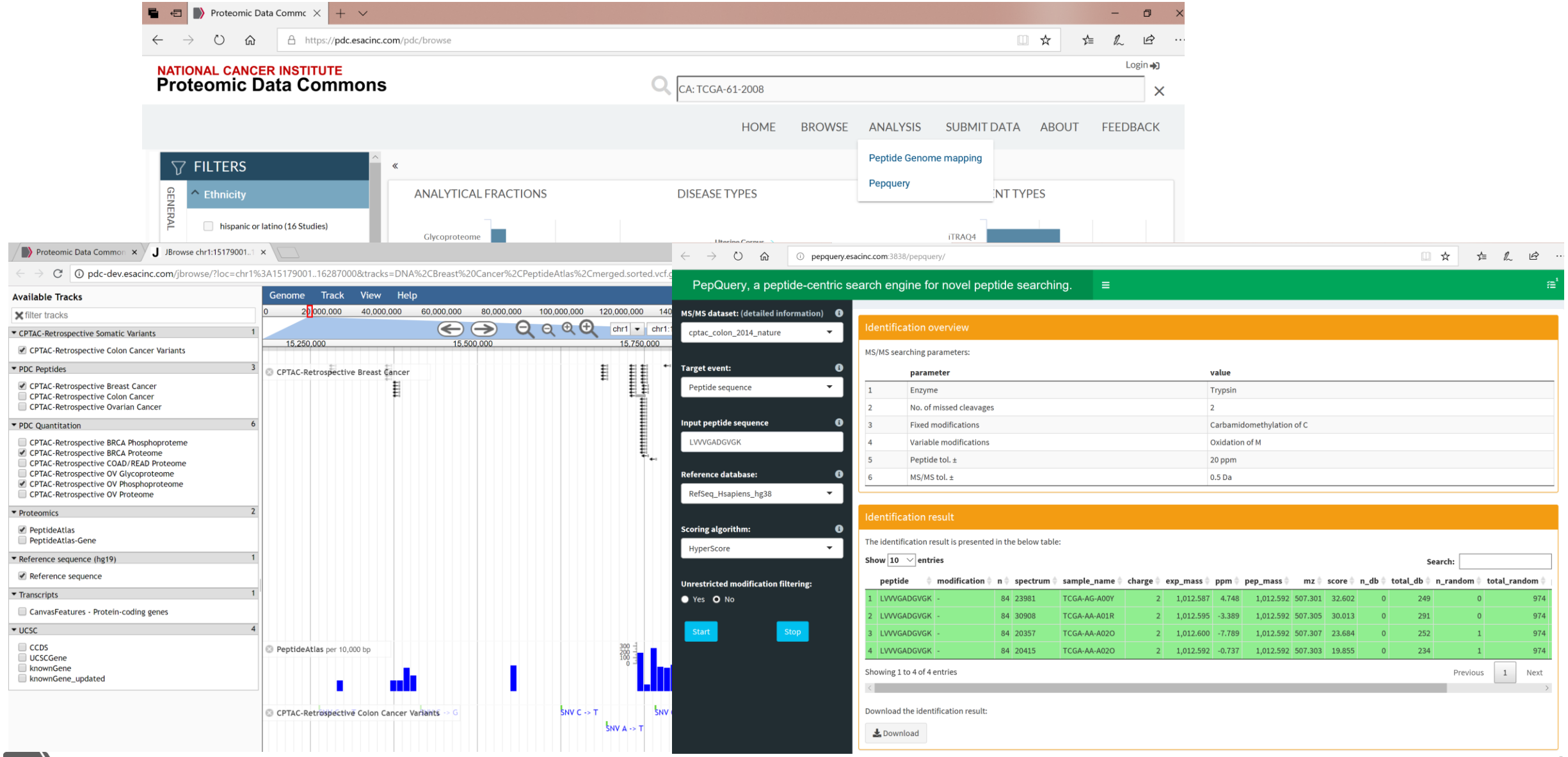
The data, including protein expression data, can be browsed interactively using a series of filters and accessed by API.



# PDC: Gene Summary

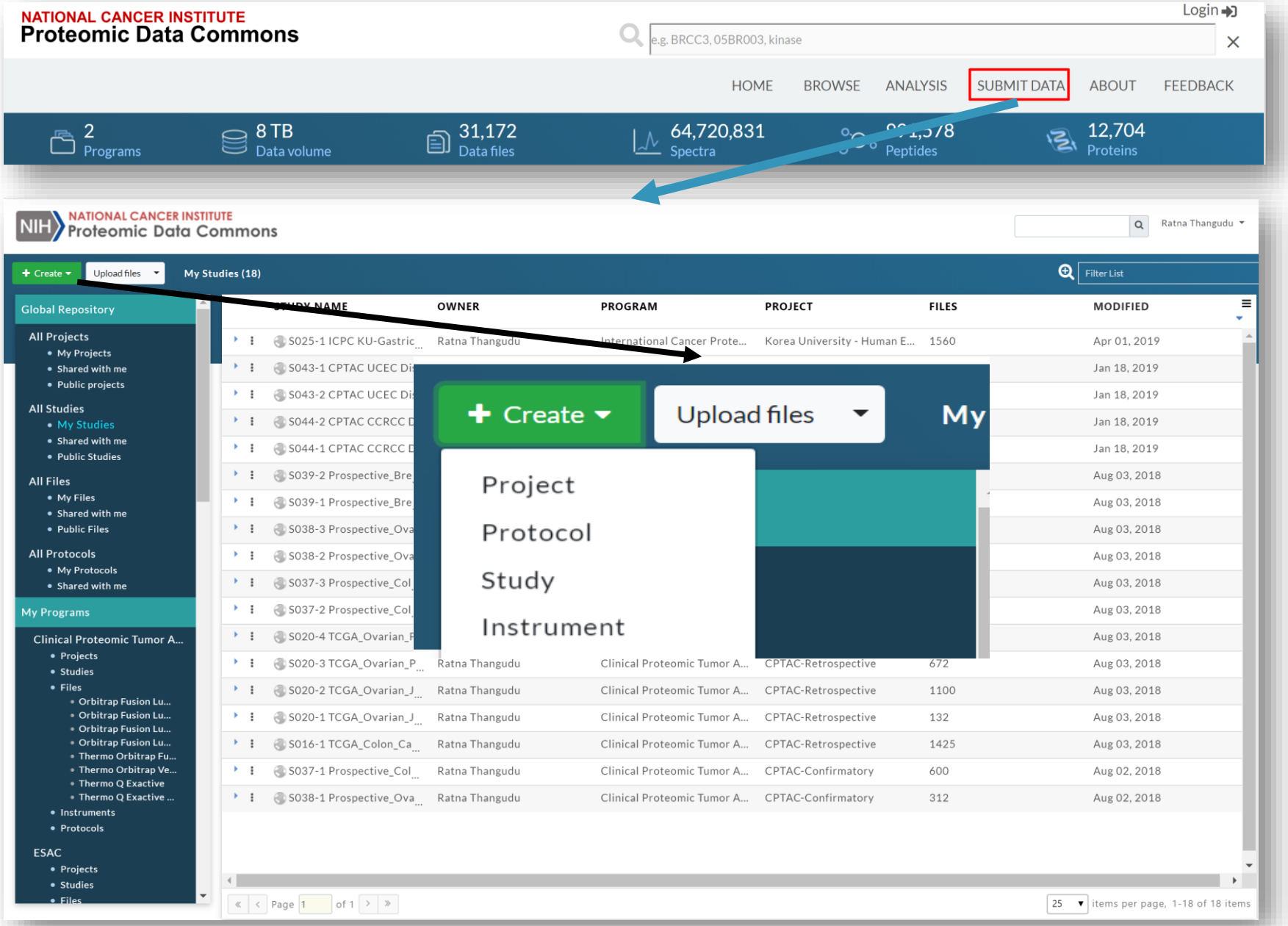


# PDC: Proteogenomic Integration



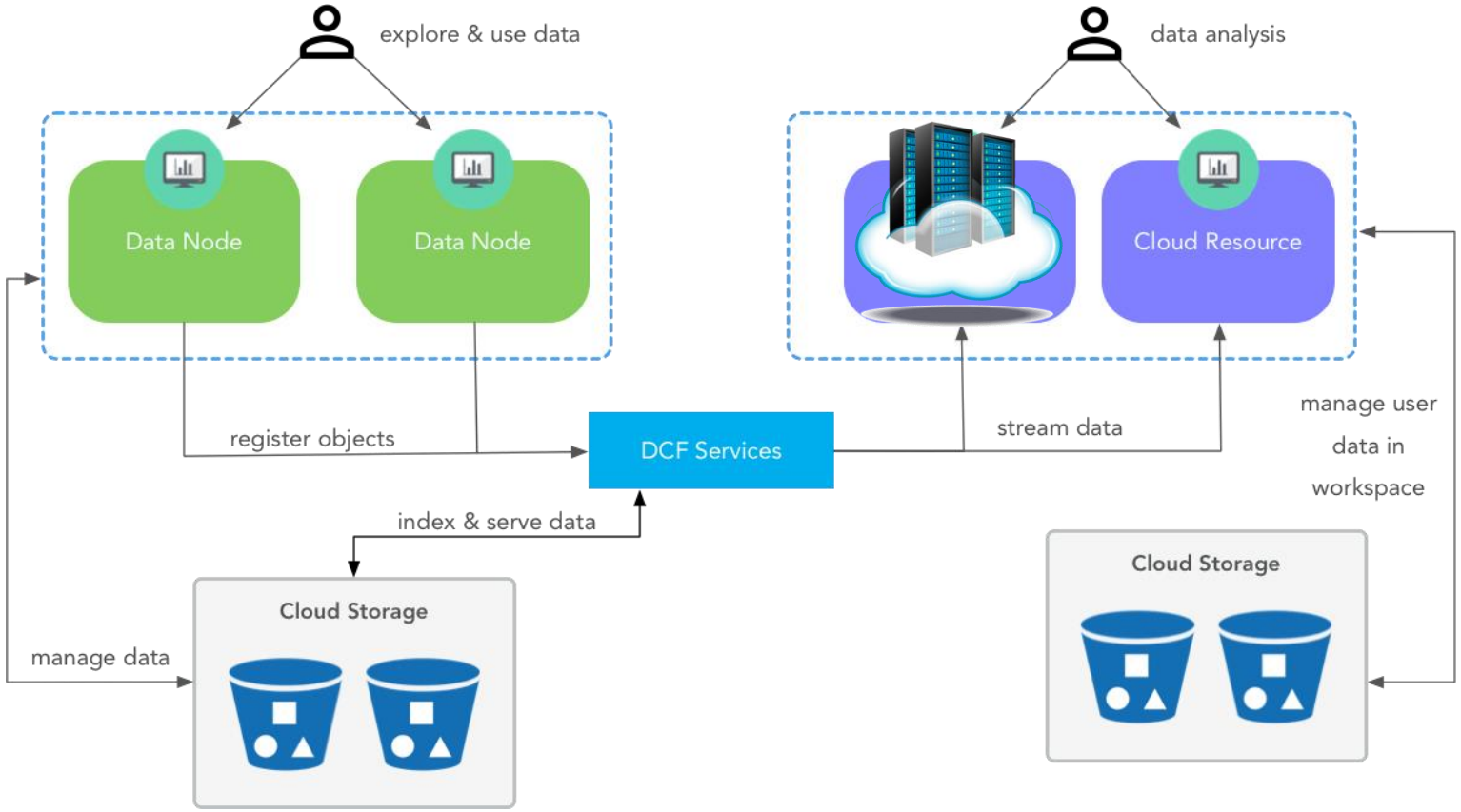
# PDC: Workspace

- Private user data repository
- Data submission portal
- Environment for analysis workflows.



# PDC: Interoperability with CRDC

Provide ability to access and stream data from across the CRDC resources with a single sign-on

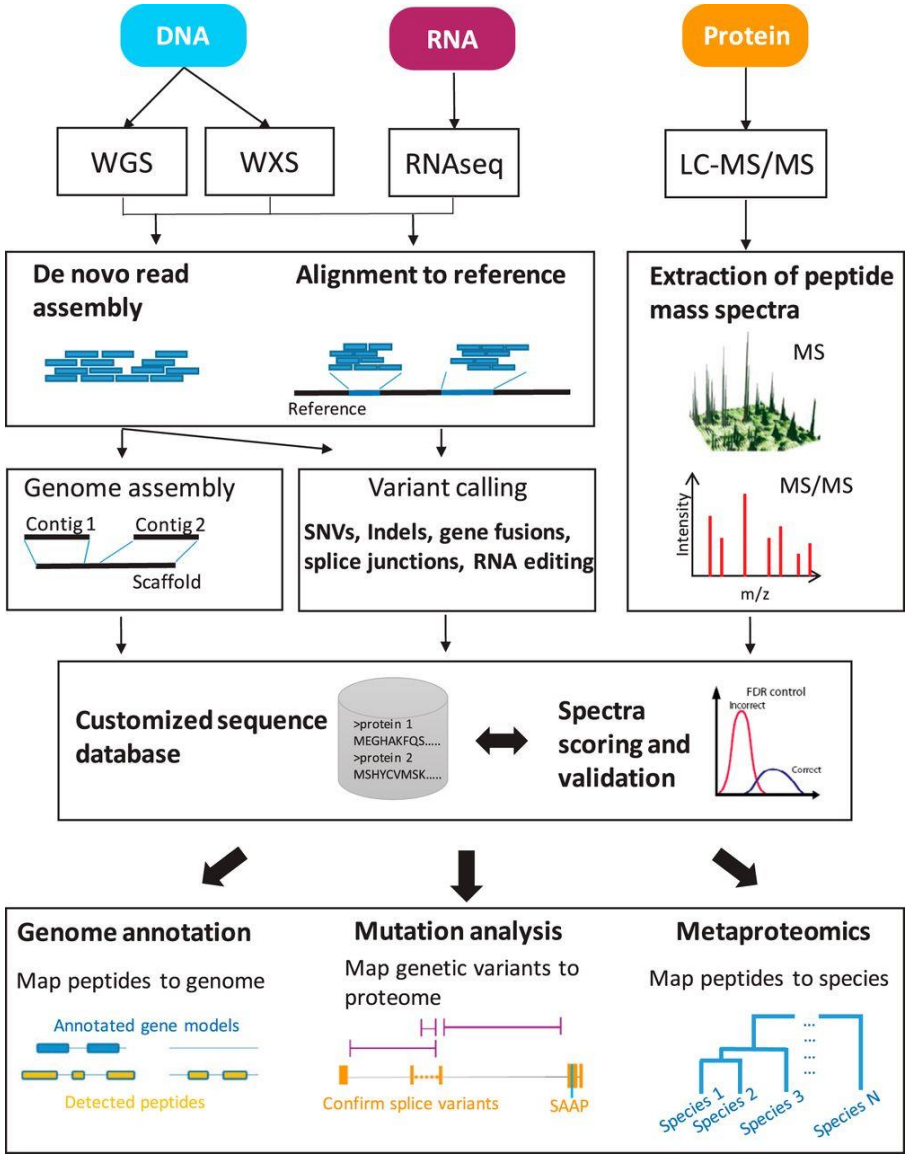



**Cloud Resources**
- **SBG CGC**
- **Broad FireCloud**
- **ISB CGC**

Source: CRDC DCF

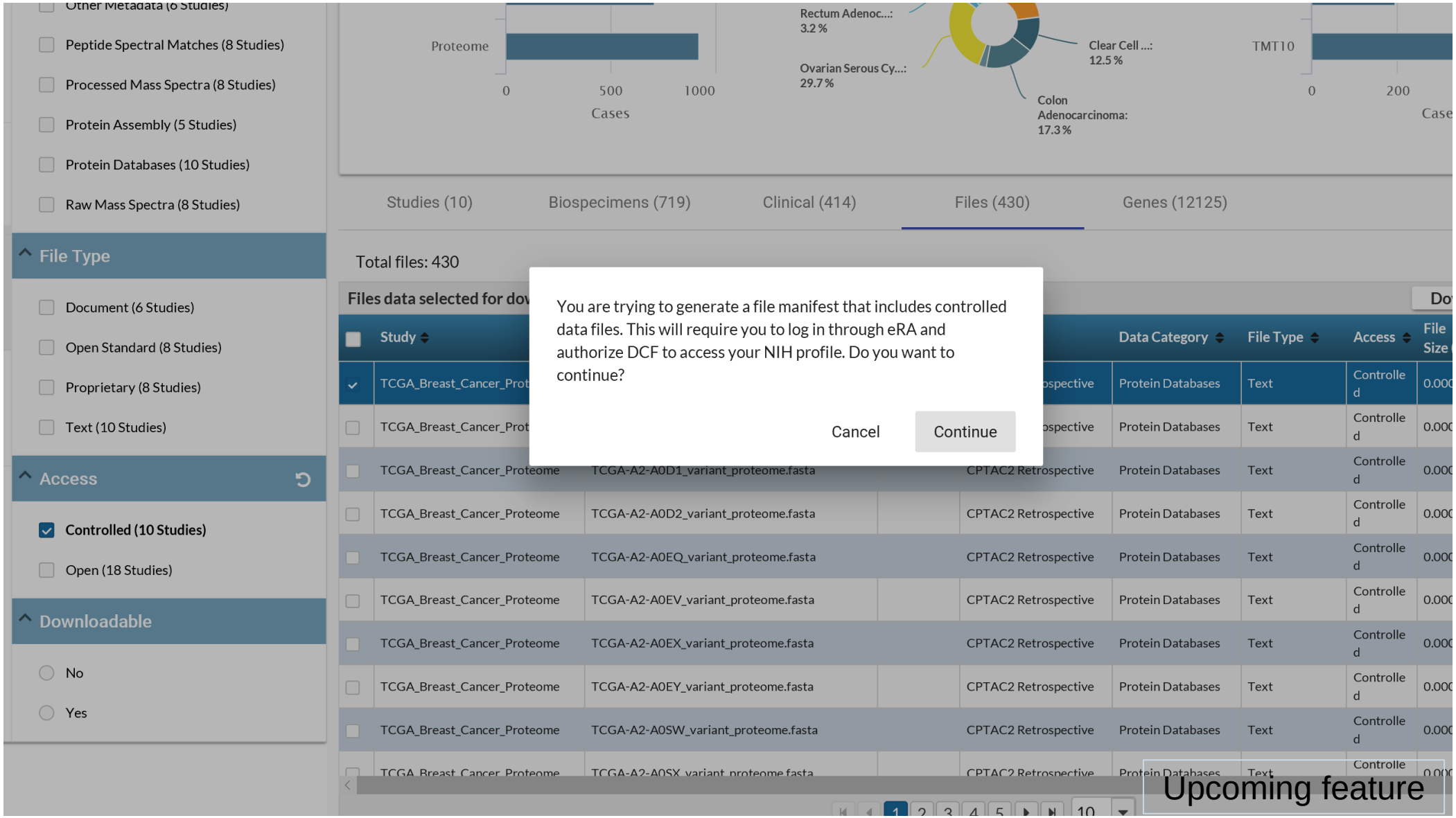
# Proteogenomics Across the CRDC

Run full proteogenomic pipelines using data stored in the Commons




Ruggles et al. 2017, MCP 16(6) 959-981

# PDC: Controlled Access Data



# PDC: Application Programming Interface

URL

https://pdc.esacinc.com/swagger.json

TOKEN

Enter api key or token

API REFERENCE

- Case
- Program
- Gene
  - Get spectral counts of available projec…
  - Find genes by partial gene_name
  - **Get spectral counts of availabl…**
- Disease
- Files
- Paginated Records
- General
- Project
- Protein

## PDC API

This is a PDC data server.

BASE URL: /graphql API VERSION: 1.0.0

## Get available info of diseases

GET ?query={diseasesAvailable {disease_type tissue_or_organ_of_origin project_submitter_id cases_count}}

Returns a list of diseases.

Parameters:

- disease_type
- **tissue_or_organ_of_origin:** Text term that describes the anatomic site of the tumor or disease. caDSR: 3427536, example: Breast
- project_submitter_id
- cases_count

Test this endpoint

TRY

Response Type application/json

Response Messages

| | |
|---|---|
| 401 | Unauthorized |


```
{
  "diagnosis_id": "string",
  "diagnosis_submitter_id": "string",
  "case_id": "string",
  "case_submitter_id": "string",
  "gdc_case_id": "string",
  "project_submitter_id": "string",
  "age_at_diagnosis": "string",
  "classification_of_tumor": "string",
  "days_to_last_follow_up": "string",
  "days_to_last_known_disease_status": "string",
  "days_to_recurrence": "string",
  "last_known_disease_status": "string",
  "morphology": "string",
  "progression_or_recurrence": "string",
  "site_of_resection_or_biopsy": "string",
  "tumor_grade": "string",
  "tumor_stage": "string",
  "vital_status": "string",
  "days_to_birth": "string",
  "days_to_death": "string",
  "prior_malignancy": "string",
  "ajcc_clinical_m": "string",
  "ajcc_clinical_n": "string",
  "ajcc_clinical_stage": "string",
  "ajcc_clinical_t": "string",
  "ajcc_pathologic_m": "string",
  "ajcc_pathologic_n": "string",
  "ajcc_pathologic_stage": "string",
  "ajcc_pathologic_t": "string",
  "ann_arbor_b_symptoms": "string",
  "ann_arbor_clinical_stage": "string",
  "ann_arbor_extranodal_involvement": "string",
  "ann_arbor_pathologic_stage": "string",
  "best_overall_response": "string",
  "burkitt_lymphoma_clinical_variant": "string",
  "cause_of_death": "string",
  "circumferential_resection_margin": "string",
  "colon_polyps_history": "string",
  "days_to_best_overall_response": "string",
  "days_to_diagnosis": "string",
  "days_to_hiv_diagnosis": "string",
  "days_to_new_event": "string",
  "figo_stage": "string",
```


Top

# Jupyter Notebook Example

NATIONAL CANCER INS'
**Proteomic Data (**

This notebook attempts to demonstrate the following:

1. Using the Proteome Data Commons (PDC) API to retreive relative protein express
   Common Data Analysis Pipeline (CDAP). More information on the PDC implmenta
2. Using the PDC API to retrieve the associated clinical metadata.
3. Formatting the data for analysis.
4. Clustering the data using the Seaborn clustermap package.
5. Visualizing the clustermap / heatmap.

The results are intended to help idenify clusters of samples (tumors) displaying similar

These are the required imports. Install them using pip if needed.

In [1]:
```
import requests
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Next, set up the query parameters.

The first one is **study_submitter_id**. These can be retrieved using an API like this one

In [2]:
```
study_submitter_id = 'S015-1' # S015-1 is TCGA-Breast(iTRAQ4)
```

Next, select the data type to retrieve for the given study. A table of data_types is availal
here. In brief, these values are log2 transformed ratio of the sample to the control chan
normalization.

In [3]:
```
data_type = 'log2_ratio'  # Retrieves CDAP iTRAQ or TMT data
```

Next, set the number of samples to retrieve. Samples are identified by their aliquot_sut
currently recommended during the initial PDC development period. Higher values may

In [4]:
```
max_aliquots = 25
```

Next, the expression data GraphQL query is set up. Adding the **study_submitter_id** a

In [5]:
```
exp_data_query = '''
{
    paginatedDataMatrix(study_submitter_id: "''' + study_submitter_id
      offset: 0 limit: ''' + str(max_aliquots) + ''') {
        total
        dataMatrix
        pagination {
            count
            sort
            from
            page
            total
            pages
            size
        }
    }
}'''
```

Get the TMT

Let's do the same thing for the clinical data.

In [6]:
```
metadata_query = '''
    {
        clinicalMetadata(study_submitter_id: "''' + study_submitter_i
            aliquot_submitter_id
            morphology
            primary_diagnosis
            tumor_grade
            tumor_stage
        }
    }
'''
```

Get the clinic

Now we can define a function to make the GraphQL Post query. This will get called onc
new to GraphQL, you can also try your queries here.

In [7]:
```
def query_pdc(query):
    URL = 'https://pdc-dev.esacinc.com/graphql'
    # Send the POST graphql query
    print('Sending query.')
    pdc_response = requests.post(URL, json={'query': query})

    # Set up a data structure for the query result
    decoded = dict()

    # Check the results
    if pdc_response.ok:
        # Decode the response
        decoded = pdc_response.json()
    else:
        # Response not OK, see error
        pdc_response.raise_for_status()
    return decoded
```

Retrieve the expression data and convert it into a pandas dataframe.

In [8]:
```
decoded = query_pdc(exp_data_query)
matrix = decoded['data']['paginatedDataMatrix']['dataMatrix']

# Aliquots are first row, gene names are first column
ga = pd.DataFrame(matrix[1:], columns=matrix[0]).set_index('Gene/Aliquot')
print('Created a dataframe of these dimensions: {}'.format(ga.shape))
```

```
Sending query.
Created a dataframe of these dimensions: (10625, 25)
```

Since the expression values are returned as strings, we need to convert those to floats and deal with missing data.

In [9]:
```
for col in ga.keys():
    ga[col] = pd.to_numeric(ga[col], errors='coerce')
```

The clustermap module within the Seaborn package does not allow for NaN values. So we must create a mask value that does not interfere much with the clustering and is likely to be unique. Not imputation is used. Missing data is a particularly tough challenge for proteomics data, particularly for phosphorylation studies. By using a value close to 0, we are saying that these are unchanged between samples. Better solutions may be used.

In [11]:
```
decoded = query_pdc(metadata_query)
matrix = decoded['data']['clinicalMetadata']
metadata = pd.DataFrame(matrix, columns=matrix[0]).set_index
print('Created a dataframe of these dimensions: {}'.format(m
```

```
Sending query.
Created a dataframe of these dimensions: (111, 4)
```

We can then set up a color mapping function for the clinical annotations.

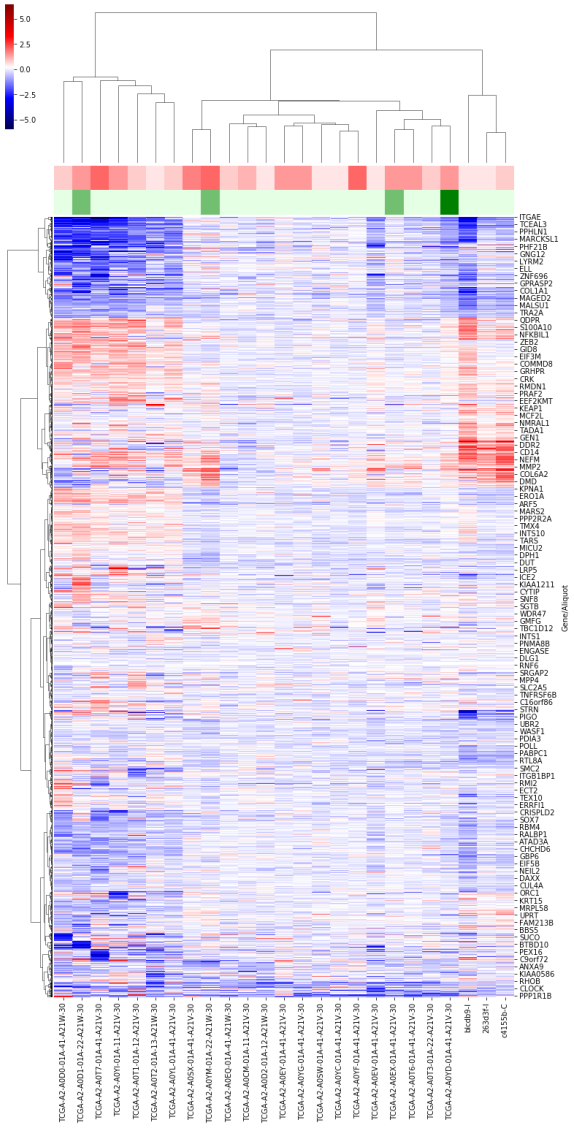
In [12]:
```
def get_colors(df, name, color) -> pd.Series:
    s = df[name]
    su = s.unique()
    colors = sns.light_palette(color, len(su))
    lut = dict(zip(su, colors))
    return s.map(lut)
```

Next, call get_colors() to map the tumor_stage and primary_diagnosis attrib

In [13]:
```
stage_col_colors = get_colors(metadata, 'tumor_stage', 'red'
diagnosis_col_colors = get_colors(metadata, 'primary_diagnos
```

And, finally, generate the large clustermap.

In [14]:
```
sns.clustermap(ga, metric='euclidean', method='ward', cmap='
               col_colors=[stage_col_colors, diagnosis_col_c
plt.show()
```



# Summary

- The PDC is being developed as a resource for democratized proteomics data (MS and harmonized processed data plus rich metadata)
- Current public datasets are from CPTAC, others to follow
- The first full-release product will have both a portal and workspace
- Tool integration is underway with the NCI Cloud Resource partners
- It is currently under active development, with limited features available

- Community feedback is being actively solicited – please **send your thoughts**

Feedback: **nci.pdc.help@esacinc.com**

# Acknowledgements


**ESAC, Inc**

- Rajesh Thangudu
- Anand Basu
- Michael Holck
- Deepak Singhal
- Karen Ketchum
- Lei Ma
- Maya Zuhl
- Yi Xin
- Padimini Chilappagari
- Ngoc Nguyen

**University of Washington**

- Michael MacCoss

**Spectragen Informatics**

- Paul Rudnick

**Georgetwon University**

- Nathan Edwards

**NCI**

- Chris Kinsinger
- Erika Kim
- Tara Hiltke

**Scientific Advisory Committee**

- Nuno Bandeira – UCSD
- Tony Blau - UW
- David Fenyo - NYU
- Brad Gibson - Amgen
- Jake Jaffe – Broad
- Mark Musen - Stanford
- Michael Noble – Broad
- Sam Payne - PNNL
- Adam Resnick - CHOP
- Dan Spellman - Merck


**Feedback : nci.pdc.help@esacinc.com**