

CPTAC Data Portal & Proteomic Data Commons

Imaging SIG

Ratna Rajesh Thangudu, Ph.D.



GEORGETOWN UNIVERSITY
Georgetown University Medical Center



July 1, 2019

Outline

- CPTAC Data Coordinating Center
- Proteomic Data Commons

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

The screenshot shows the CPTAC Public Data Portal website. At the top, the URL is <https://proteomics.cancer.gov/data-portal>. The header includes the NIH logo and the text "NATIONAL CANCER INSTITUTE Office of Cancer Clinical Proteomics Research". There are links for "CONTACT US" and "SIGN UP FOR UPDATES", and a search bar. Below the header is a navigation bar with "Center for Strategic Scientific Initiatives" and links for "DATA PORTAL HOME", "ASSAY PORTAL", "ANTIBODY PORTAL", and "ABOUT". The main content area features a large banner with the text "DATA ANALYSIS" and icons representing data analysis. Below the banner, there is a "Data Portal" section with a description of the portal's purpose and a "Data Portal Stats" section showing "10.9 TB Proteomic Data", "289 TB Data Downloaded", and "19,115 Users". There is also a "Software Tools" section with checkboxes for "CONTACT US" and "SIGN UP FOR EMAIL UPDATES". At the bottom, there are two buttons: "Available Data" and "Data Use Agreement".

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

NIH NATIONAL CANCER INSTITUTE
Office of Cancer Clinical
Proteomics Research

CONTACT US | SIGN UP FOR UPDATES

Search... SEARCH

Center for Strategic Scientific Initiatives

DATA PORTAL HOME ASSAY PORTAL ANTIBODY PORTAL ABOUT

Data Portal

The CPTAC Data Portal is a centralized repository for the public dissemination of proteomic sequence datasets collected by CPTAC, along with corresponding genomic sequence datasets. In addition, available are analyses of CPTAC's raw mass spectrometry-based data files (mapping of spectra to peptide sequences and protein identification) by individual investigators from CPTAC and by a Common Data Analysis Pipeline.

A core principle of CPTAC is the sharing and re-use of data across the biomedical research community, as vital to accelerating scientific discovery and its clinical translation to patient care. The Data Portal represents the NCI's largest public repository of proteogenomic comprehensive sequence datasets, essentially a Proteogenomic Cancer Atlas (PCA). Proteomic data and related data files are organized into datasets by study, sub-proteome, and analysis site. All **data is freely available to the public**, subject to the [Data Use Agreement](#). Reference mass spectral peptide libraries resulting from these studies may also be downloaded freely from the [NIST Peptide Library](#).

Data Portal Stats

10.9 TB Proteomic Data
289 TB Data Downloaded
19,115 Users

Software Tools

CONTACT US
 SIGN UP FOR EMAIL UPDATES

Available Data Data Use Agreement

<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 new | <p>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)</p> | |
| CPTAC LUAD Discovery Study new | <p>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).</p> | |
| Colon Cancer Therapeutic Opportunities new | <p>Proteogenomic study on a prospectively collected colon cancer cohort with 110 paired tumor and normal adjacent tissues.</p> |  |
| Proteogenomics of Gastric Cancer | <p>Proteogenomic analysis was performed on a cohort of 80 patients with early-onset gastric cancer recruited from the Korean population.</p> |  |
| CPTAC UCEC Discovery Study | <p>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).</p> | |
| CPTAC CCRCC Discovery Study | <p>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).</p> | |

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.

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: Analytical Fraction:

| Data set name | All raw mzML PSM prot meta | | | | | | Size |
|---|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|----------|
| | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | |
| CPTAC_OVprospective_metadata | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | 222.12KB |
| CPTAC_OVprospective_PNNL_Phosphoproteome_CDAP_Protein_Report.r1 | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | 111.34MB |
| 01CPTAC_OVprospective_Proteome_JHU_20161209 | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 25.49GB |
| 02CPTAC_OVprospective_Proteome_JHU_20161104 | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 25.67GB |
| 03CPTAC_OVprospective_Proteome_JHU_20161110 | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 27.45GB |
| 04CPTAC_OVprospective_Proteome_JHU_20161111 | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 26.02GB |
| 05CPTAC_OVprospective_Proteome_JHU_20161115 | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 24.74GB |

CPTAC Data Portal: Releases

The screenshot displays the CPTAC Data Portal interface. The top navigation bar includes the NIH logo, the text 'NATIONAL CANCER INSTITUTE Office of Cancer Clinical Proteomics Research', and a search bar. Below this is the 'Center for Strategic Scientific Initiatives' banner and a 'DATA PORTAL HOME' link.

On the left side, there is a sidebar with 'Available Studies' including 'CPTAC 3 (2016-present)', 'CPTAC 2 (2011-2016)', 'CPTC (2006-2011)', 'External Studies', 'Query Data', 'Help', and 'Http Data Access' (highlighted with a red box).

The main content area features a section for 'APOBEC3A Polymorphi' with a description of oral squamous cell carcinoma in Taiwanese patients. A red box highlights a link to 'BioProject:PRJNA327548 and SRA' in the text. Below this is a section for 'Biospecimens and Metadata Files' with a red box around a link to 'Clinical Data for OSCC-Taiwan Proteomic Data Sets, and Nature Communications 8, 465 (2017); Supplementary Table 1, OSCC-Taiwan iTRAQ Sample Mapping'.

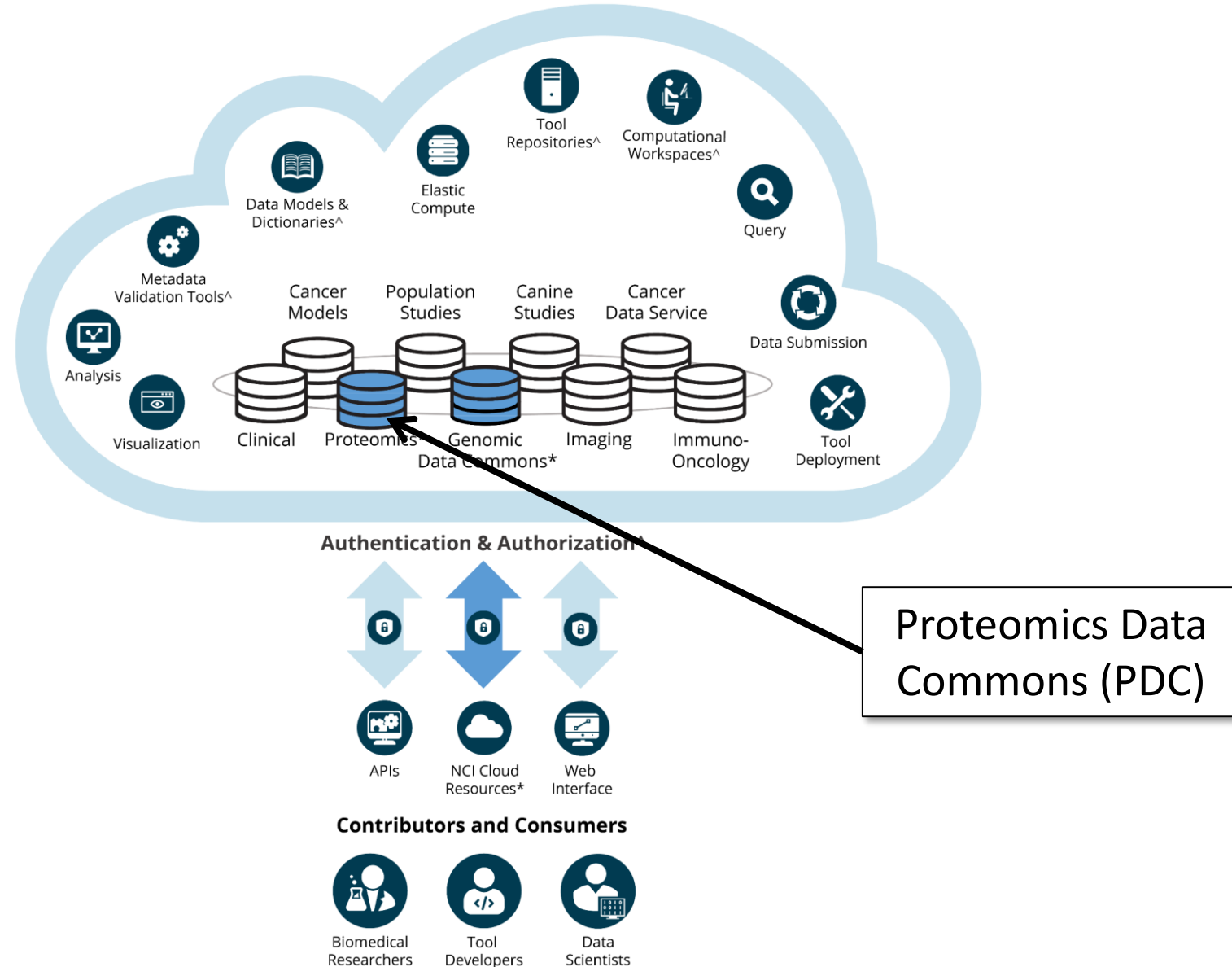
A 'Data Sets' section contains a 'DOWNLOAD' button and a table of data sets. The table has columns for 'Data set name', 'All raw mzML PSM prot meta' (checkboxes), and 'Size'.

| Data set name | All raw mzML PSM prot meta | Size |
|---------------------------------------|---|----------|
| Oral_Squamous_Cell_Carcinoma_Metadata | <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 344.31KB |
| OSCC-P01 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 10.73GB |
| OSCC-P04 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 13.48GB |
| OSCC-P06 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 12.60GB |
| OSCC-P10 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 10.43GB |
| OSCC-P13 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 10.76GB |
| OSCC-P14 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 9.59GB |
| OSCC-P19 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 12.11GB |
| OSCC-P21 | <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> | 11.35GB |

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

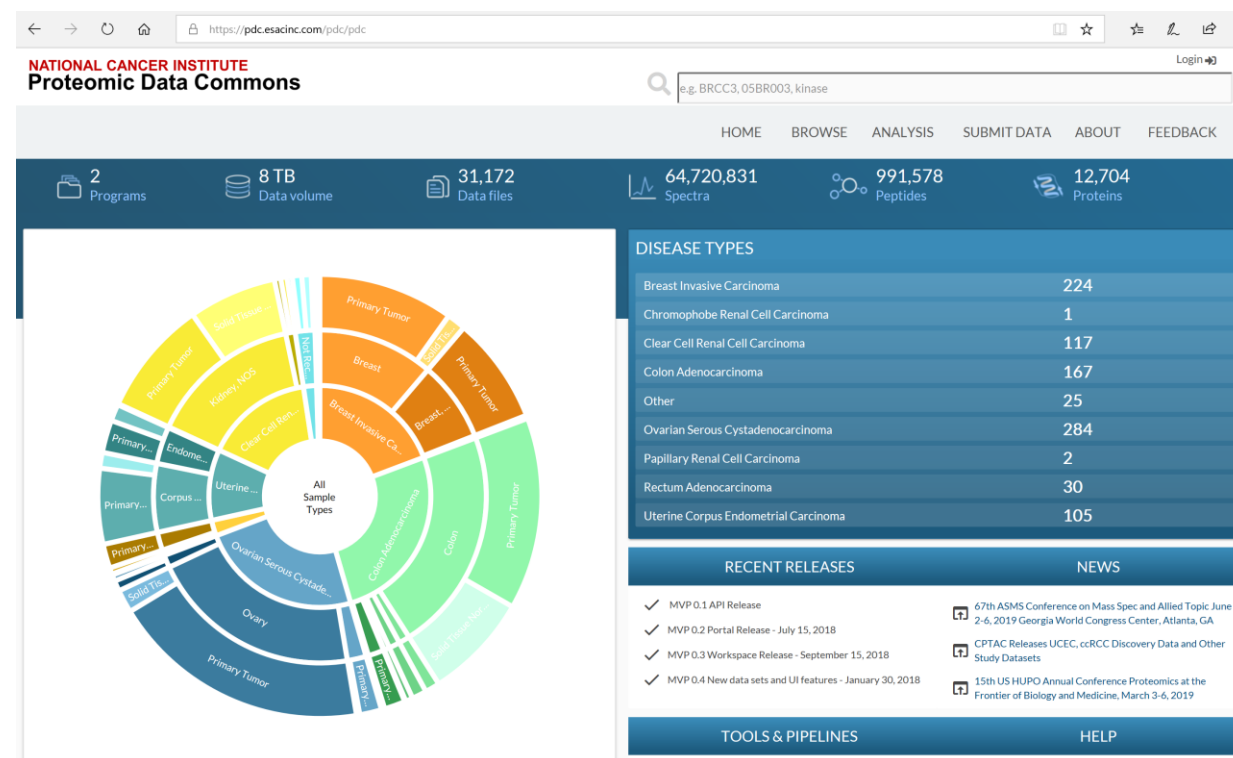
Proteomics Data Commons: High Level Goals

- 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.

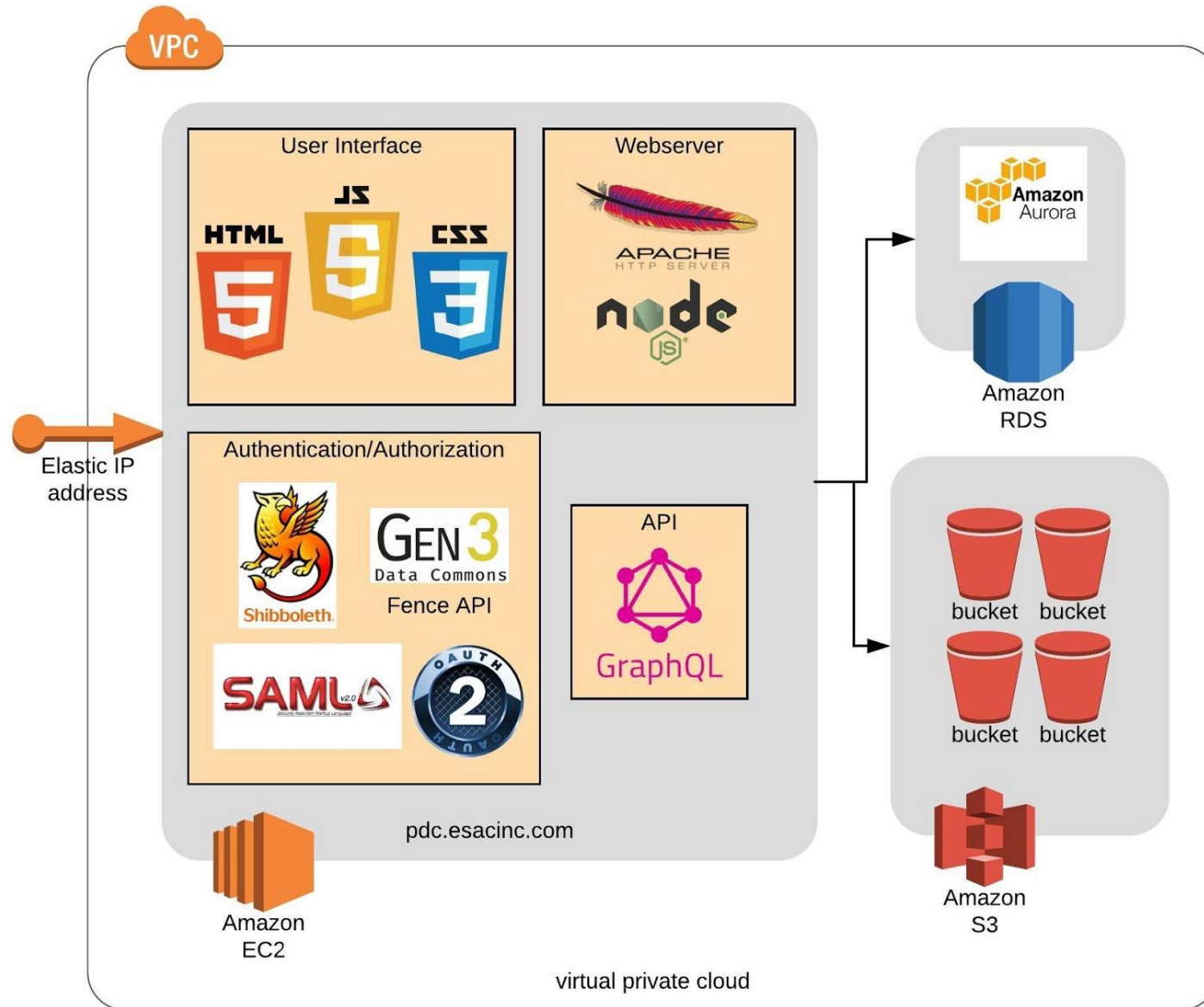
Proteomic Data Commons: now available as beta program

- 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

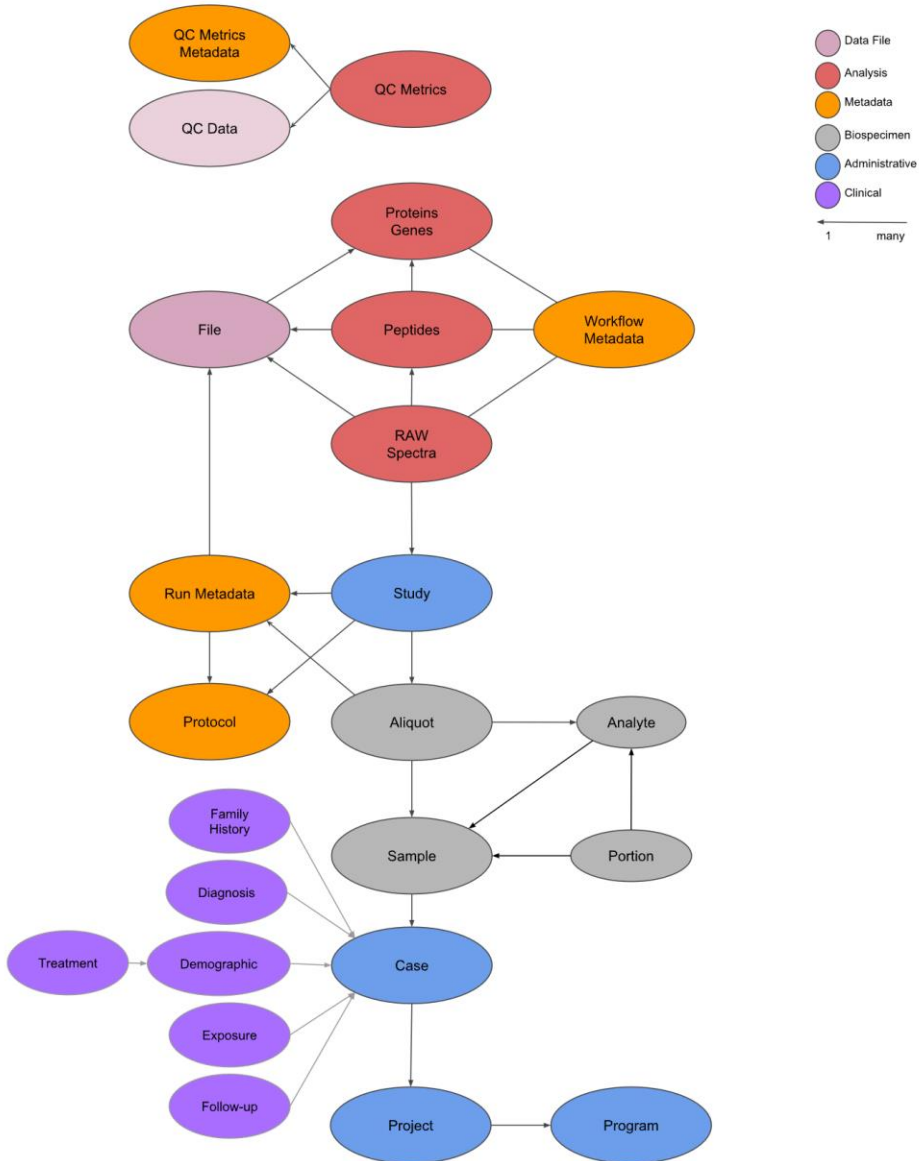
<https://pdc.esacinc.com>



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 labelling 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 biological entity 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 segmenting 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 generated by data analysis pipeline |
|------|---|

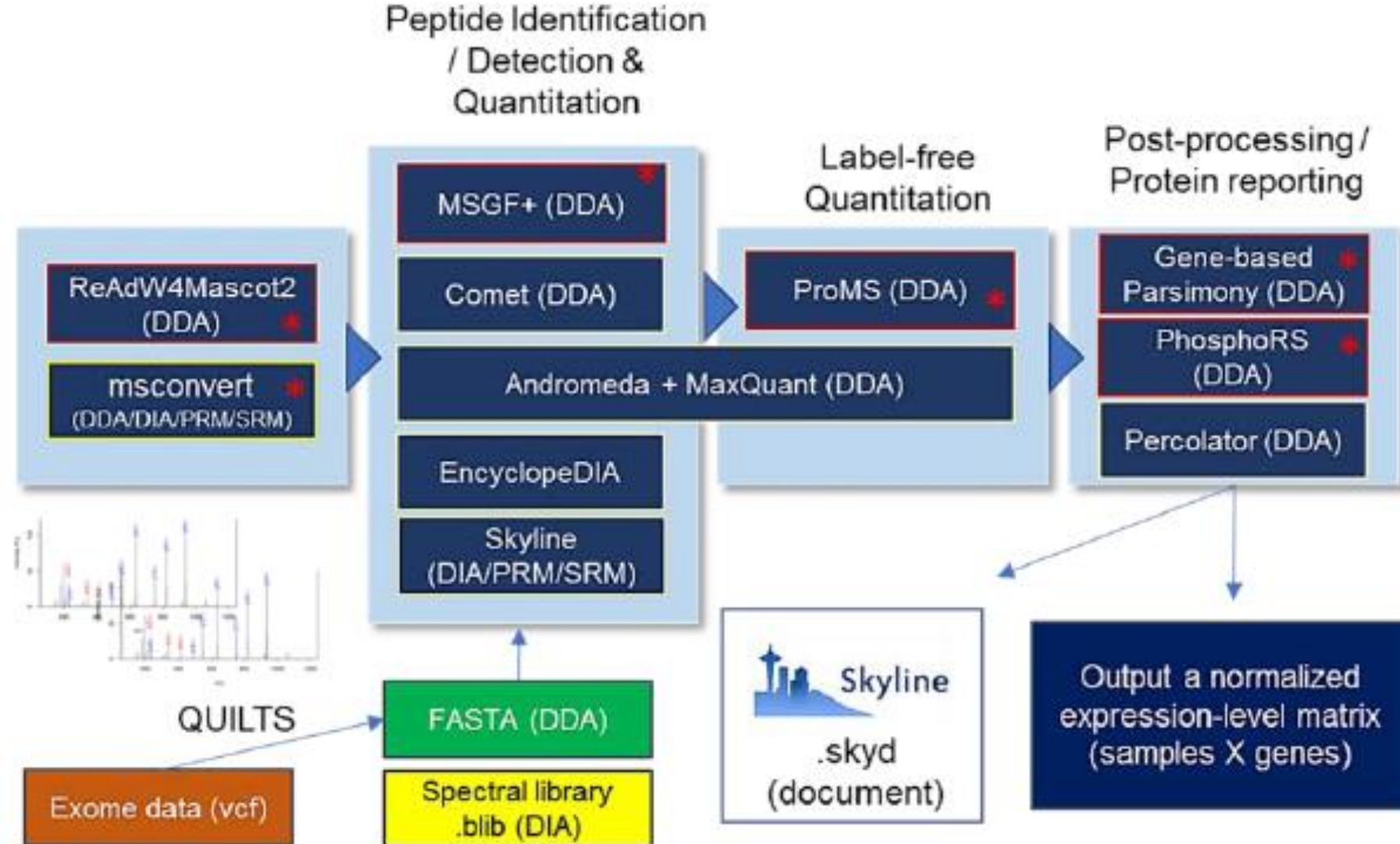
Metadata

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

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.

NATIONAL CANCER INSTITUTE Proteomic Data Commons

Search: e.g. BRCC3, 05BR003, kinase

HOME BROWSE ANALYSIS SUBMIT DATA ABOUT FEEDBACK

FILTERS

- Project**
 - CPTAC2 Confirmatory (8 Studies)
 - CPTAC2 Retrospective (7 Studies)
 - CPTAC3 Discovery (4 Studies)
 - Quantitative digital maps of tissue biopsies (1 Study)
- Primary Site**
 - Breast (4 Studies)
 - Colon (4 Studies)
 - Kidney (3 Studies)
 - N/A (4 Studies)
 - Ovary (7 Studies)
 - Rectum (1 Study)
 - Uterus, NOS (2 Studies)
- Program**
 - Aebersold Lab (1 Study)
 - Clinical Proteomic Tumor Analysis Consortium (19 Studies)
- Disease Type**
 - Breast Invasive Carcinoma (4 Studies)

ANALYTICAL FRACTIONS

| Fraction | Cases |
|-----------------|-------|
| Glycoproteome | ~100 |
| Phosphoproteome | ~700 |
| Proteome | ~900 |

DISEASE TYPES

| Disease Type | Percentage |
|--------------------------------------|------------|
| Uterine Corpus Endometrial Carcinoma | 11.2% |
| Breast Invasiv... | 23.0% |
| Rectum Adenoc... | 3.2% |
| Ovarian Serous Cy... | 29.7% |
| Clear Cell ... | 12.5% |
| Colon Adenocarcinoma | 17.3% |

EXPERIMENT TYPES

| Experiment Type | Cases |
|-----------------|-------|
| iTRAQ4 | ~250 |
| Label Free | ~180 |
| TMT10 | ~550 |

Studies (20) Biospecimens (1402) Clinical (926) Files (31447) Genes (12597)

Total studies: 20 [Download Study Manifest](#)

| Study | Project | Program | Disease Type | Primary Site | Analytical Fraction | Experiment Type | Cases # | Raw | mzML | Protocol | PSM | Protein Assembly |
|---|------------------|--|--|-----------------|---------------------|-----------------|---------|-----|------|----------|-----|------------------|
| <input type="checkbox"/> CPTAC CCRC Discovery Study - Phosphoproteome | CPTAC3 Discovery | Clinical Proteomic Tumor Analysis Consortium | Clear Cell Renal Cell Carcinoma;Other | Kidney; N/A | Phosphoproteome | TMT10 | 126 | 338 | 338 | 7 | 338 | 15 |
| <input type="checkbox"/> CPTAC CCRC Discovery Study - Proteome | CPTAC3 Discovery | Clinical Proteomic Tumor Analysis Consortium | Clear Cell Renal Cell Carcinoma;Other | Kidney; N/A | Proteome | TMT10 | 126 | 650 | 650 | 7 | 650 | 13 |
| <input type="checkbox"/> CPTAC UCEC Discovery Study - Phosphoproteome | CPTAC3 Discovery | Clinical Proteomic Tumor Analysis Consortium | Other;Uterine Corpus Endometrial Carcinoma | N/A;Uterus, NOS | Phosphoproteome | TMT10 | 115 | 240 | 240 | 10 | 240 | 15 |
| <input type="checkbox"/> CPTAC UCEC Discovery Study - Proteome | CPTAC3 Discovery | Clinical Proteomic Tumor Analysis Consortium | Other;Uterine Corpus Endometrial Carcinoma | N/A;Uterus, NOS | Proteome | TMT10 | 115 | 480 | 480 | 10 | 480 | 13 |

PDC: Gene Summary

Proteomic Data Commons | <https://pdc.esacinc.com/pdc/gene/BRCC3>

NATIONAL CANCER INSTITUTE
Proteomic Data Commons

Search: e.g. BRCC3, 05BR003, kinase

HOME BROWSE ANALYSIS SUBMIT DATA ABOUT FEEDBACK

FILTERS

Project

- CPTAC2 Confirmatory (3 Studies)
- CPTAC2 Retrospective (6 Studies)
- CPTAC3 Discovery (4 Studies)
- Quantitative digital maps of tissue biopsies (0 Study)

Primary Site

- Breast (3 Studies)
- Colon (3 Studies)
- Kidney (2 Studies)
- N/A (4 Studies)
- Ovary (3 Studies)
- Rectum (1 Study)
- Uterus, NOS (2 Studies)

Program

- Aebersold Lab (0 Study)
- Clinical Proteomic Tumor Analysis Consortium (13 Studies)

Disease Type

- Breast Invasive Carcinoma (3 Studies)
- Chromophobe Renal Cell Carcinoma (0 Study)
- Clear Cell Renal Cell Carcinoma (2 Studies)
- Colon Adenocarcinoma (3 Studies)
- Other (4 Studies)
- Ovarian Serous Cystadenocarcinoma (3 Studies)
- Papillary Renal Cell Carcinoma (0 Study)
- Rectum Adenocarcinoma (1 Study)
- Uterine Corpus Endometrial Carcinoma (2 Studies)

GENE: BRCC3

NCBI Gene ID: 79184 | Authority: HGNC:24185 | Description: BRCA1/BRCA2-containing complex subunit 3 | Organism: Homo sapiens

Chromosome: X | Locus: Xq28 | Assays: A0A087WZR3, A0A0A0MS96, A0A0D9SF50, H7C413, H9KVA9, NP_001018065.1, NP_001229569.1, NP_077308.1, P46736, P46736-2, P46736-3, P46736-4, P46736-5, X6RJ57, XP_005274808.1, XP_016885327.1

Detected Post-translational Modifications | Total PTM sites: 1

| PTM Type | Site | Peptide |
|----------|---------------------|-------------|
| phospho | NP_001018065.1:s227 | IHLTHLDSVTK |

Studies in Which a Gene Product Was Detected | Total Studies: 13

| Study | Experiment Type | Spectral Counts | Distinct Peptides | Unshared Peptides | No of Aliquots | No of Ploxes |
|--|-----------------|-----------------|-------------------|-------------------|----------------|--------------|
| CPTAC UCEC Discovery Study - Proteome | TMT10 | 317 | 34 | 33 | 146 | 17 |
| CPTAC UCEC Discovery Study - Phosphoproteome | TMT10 | 17 | 5 | 5 | 146 | 17 |
| CPTAC CCRC Discovery Study - Phosphoproteome | TMT10 | 4 | 2 | 2 | 218 | 24 |

Biospecimens/ Aliquot

| Aliquot | Site | Proteome |
|---------------|---|----------------|
| CPT0064930001 | DM583_LUNG | 14-3-3_beta |
| CPT0064930004 | NCI1094_LUNG | 14-3-3_epsilon |
| CPT0026030004 | P3H1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE | |
| CPT0079230003 | HUT78_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE | |
| CPT0062940003 | LUMUC3_URINARY_TRACT | |
| CPT0012550003 | HOS_BONE | |
| | NCI1174_LUNG | |
| | NCI1191_LUNG | |
| | NCI1227_LUNG | |
| | NCI1448_LIVER | |
| | NCI1428_PLEURA | |
| | OV56_OVARY | |
| | JH054_OVARY | |
| | KY5680_OESOPHAGUS | |
| | KLE_ENDOMETRIUM | |
| | HS9951_SKIN | |
| | LN229_CENTRAL_NERVOUS_SYSTEM | |
| | P31FUJ_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE | |
| | RKN_SOFT_TISSUE | |
| | PATU9985_PANCREAS | |
| | NHG_AUTONOMIC_GANGLIA | |
| | REB1_CENTRAL_NERVOUS_SYSTEM | |
| | REB1_C22_LUNG | |
| | OU6823_LARGE_INTESTINE | |
| | SNGM_ENDOMETRIUM | |
| | OU6827_BONE | |
| | NCI12347_LUNG | |
| | SW1990_PANCREAS | |
| | HS9401_SKIN | |
| | POS117_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE | |
| | REB1_CENTRAL_NERVOUS_SYSTEM | |
| | REB1_C22_LUNG | |
| | HT1080_SOFT_TISSUE | |
| | NCI12087_LUNG | |
| | COV318_OVARY | |
| | NCI12085_LUNG | |
| | CAK1_KIDNEY | |
| | NCI1716_LARGE_INTESTINE | |
| | NCI1181_LUNG | |
| | NCI1341_LUNG | |
| | SW480_LARGE_INTESTINE | |
| | HS7467_STOMACH | |
| | LU65_LUNG | |
| | SNUC2A_LARGE_INTESTINE | |
| | HEC101_ENDOMETRIUM | |
| | NCI1593_LUNG | |
| | NCI1691_LUNG | |
| | SKUT1_SOFT_TISSUE | |
| | MDAMB175VIL_BREAST | |
| | SUPB15_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE | |
| | BECKER_CENTRAL_NERVOUS_SYSTEM | |
| | NCI1226_LUNG | |
| | NCI1166_LUNG | |
| | NCI1174_LUNG | |
| | ISTMES1_PLEURA | |
| | NCI1174_LUNG | |
| | NCI12342_LUNG | |
| | EBCT_LUNG | |
| | T173_BONE | |
| | NCI1322_LUNG | |
| | HS9107_BONE | |
| | HS9107_BONE | |
| | PAN1005_PANCREAS | |
| | NCI12291_LUNG | |
| | T47D_BREAST | |
| | BICR16_UPPER_AERODIGESTIVE_TRACT | |
| | TE10_OESOPHAGUS | |
| | SKMEL28_SKIN_AERODIGESTIVE_TRACT | |
| | SKMEL28_SKIN_AERODIGESTIVE_TRACT | |
| | IBI_UPPER_AERODIGESTIVE_TRACT | |

PDC: Proteogenomic Integration

| parameter | value |
|---------------------------|---------------------------|
| 1 Enzyme | Trypsin |
| 2 No. of missed cleavages | 2 |
| 3 Fixed modifications | Carbamidomethylation of C |
| 4 Variable modifications | Oxidation of M |
| 5 Peptide tol. ± | 20 ppm |
| 6 MS/MS tol. ± | 0.5 Da |

| peptide | modification | n | spectrum | sample_name | charge | exp_mass | ppm | pep_mass | mz | score | n_db | total_db | n_random | total_random |
|--------------|--------------|----|----------|--------------|--------|-----------|--------|-----------|---------|--------|------|----------|----------|--------------|
| 1 LWVGADGVGK | - | 84 | 23981 | TCGA-AG-A00Y | 2 | 1,012.587 | 4.748 | 1,012.592 | 507.301 | 32.602 | 0 | 249 | 0 | 974 |
| 2 LWVGADGVGK | - | 84 | 30908 | TCGA-AA-A01R | 2 | 1,012.595 | -3.389 | 1,012.592 | 507.305 | 30.013 | 0 | 291 | 0 | 974 |
| 3 LWVGADGVGK | - | 84 | 20357 | TCGA-AA-A02O | 2 | 1,012.600 | -7.789 | 1,012.592 | 507.307 | 23.684 | 0 | 252 | 1 | 974 |
| 4 LWVGADGVGK | - | 84 | 20415 | TCGA-AA-A02O | 2 | 1,012.592 | -0.737 | 1,012.592 | 507.303 | 19.855 | 0 | 234 | 1 | 974 |

PDC: Workspace

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

NATIONAL CANCER INSTITUTE Proteomic Data Commons

Search: e.g. BRCC3, 05BR003, kinase

HOME BROWSE ANALYSIS **SUBMIT DATA** ABOUT FEEDBACK

2 Programs 8 TB Data volume 31,172 Data files 64,720,831 Spectra 99,1578 Peptides 12,704 Proteins

NIH NATIONAL CANCER INSTITUTE Proteomic Data Commons

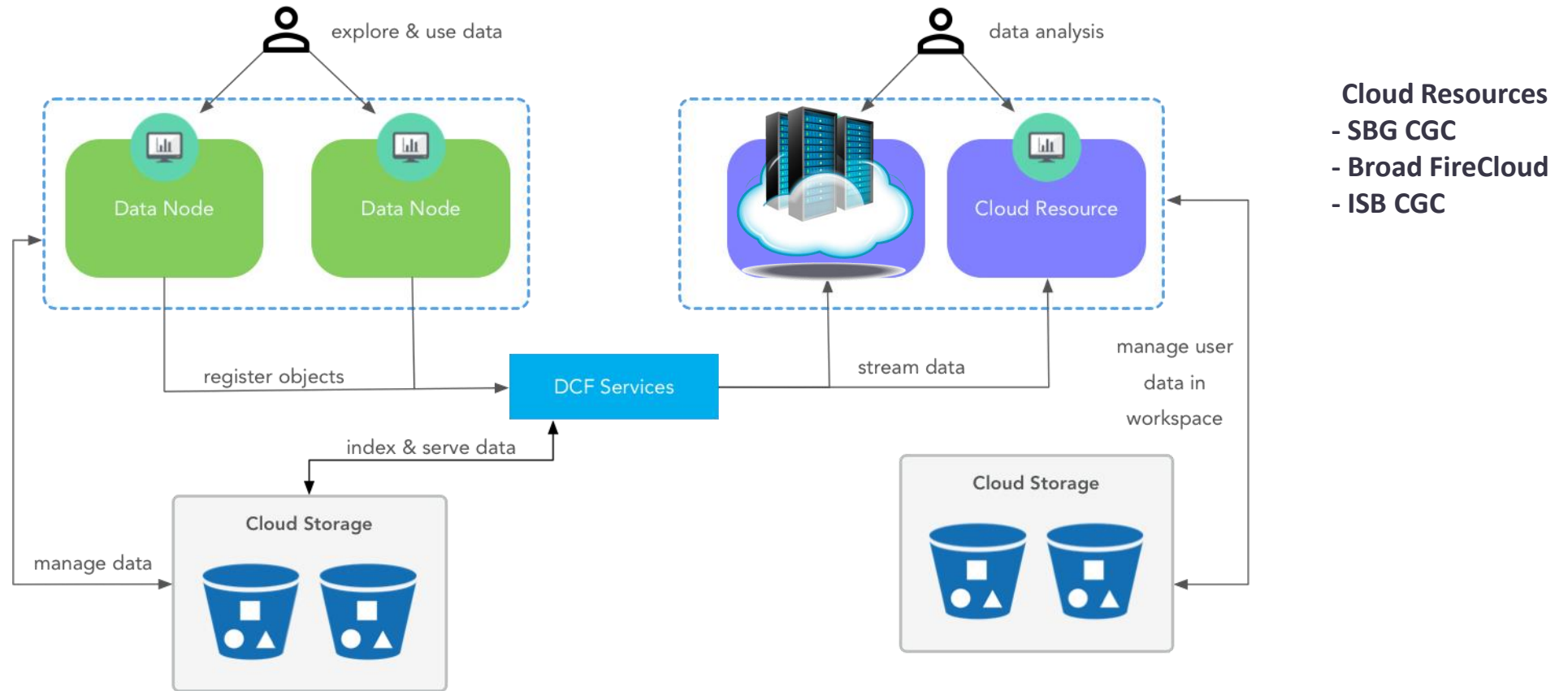
My Studies (18)

| STUDY NAME | OWNER | PROGRAM | PROJECT | FILES | MODIFIED |
|---------------------------|----------------|-------------------------------|-------------------------------|-------|--------------|
| S025-1 ICPC KU-Gastric... | Ratna Thangudu | International Cancer Prote... | Korea University - Human E... | 1560 | Apr 01, 2019 |
| S043-1 CPTAC UCEC D... | | | | | Jan 18, 2019 |
| S043-2 CPTAC UCEC D... | | | | | Jan 18, 2019 |
| S044-2 CPTAC CCRCC D... | | | | | Jan 18, 2019 |
| S044-1 CPTAC CCRCC D... | | | | | Jan 18, 2019 |
| S039-2 Prospective_Bre... | | | | | Aug 03, 2018 |
| S039-1 Prospective_Bre... | | | | | Aug 03, 2018 |
| S038-3 Prospective_Ova... | | | | | Aug 03, 2018 |
| S038-2 Prospective_Ova... | | | | | Aug 03, 2018 |
| S037-3 Prospective_Col... | | | | | Aug 03, 2018 |
| S037-2 Prospective_Col... | | | | | Aug 03, 2018 |
| S020-4 TCGA_Ovarian_F... | | | | | Aug 03, 2018 |
| S020-3 TCGA_Ovarian_P... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Retrospective | 672 | Aug 03, 2018 |
| S020-2 TCGA_Ovarian_J... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Retrospective | 1100 | Aug 03, 2018 |
| S020-1 TCGA_Ovarian_J... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Retrospective | 132 | Aug 03, 2018 |
| S016-1 TCGA_Colon_Ca... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Retrospective | 1425 | Aug 03, 2018 |
| S037-1 Prospective_Col... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Confirmatory | 600 | Aug 02, 2018 |
| S038-1 Prospective_Ova... | Ratna Thangudu | Clinical Proteomic Tumor A... | CPTAC-Confirmatory | 312 | Aug 02, 2018 |

Page 1 of 1 25 items per page, 1-18 of 18 items

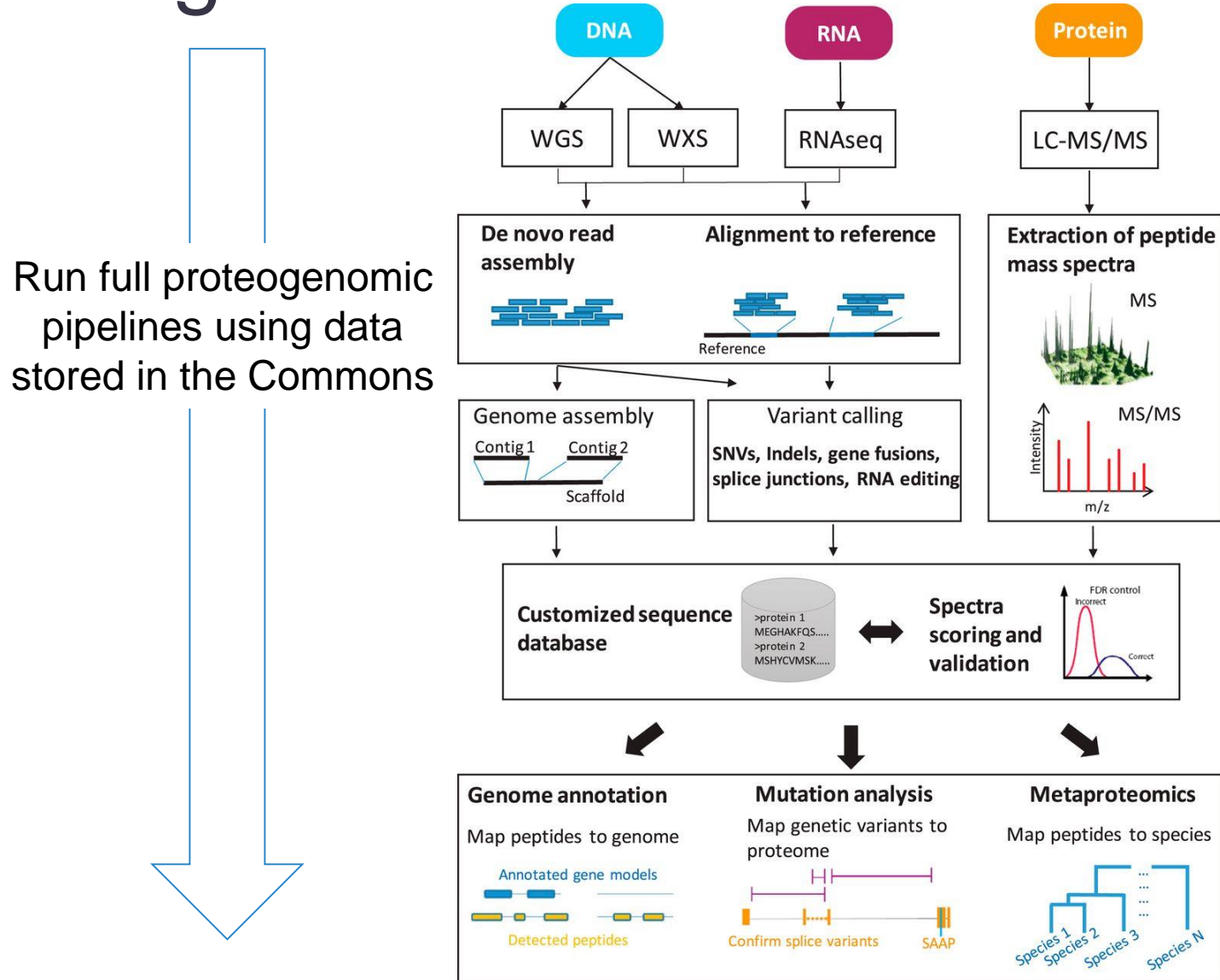
PDC: Interoperability with CRDC

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



Source: CRDC DCF

Proteogenomics Across the CRDC



Run full proteogenomic pipelines using data stored in the Commons

PDC: Controlled Access Data

The screenshot displays a data management interface with several components:

- Left Sidebar:** A list of filters for file types and access. Under "File Type", "Text (10 Studies)" is selected. Under "Access", "Controlled (10 Studies)" is selected. Under "Downloadable", "Yes" is selected.
- Top Charts:** A bar chart for "Proteome" (Cases) and a donut chart for "Cases" with categories: Ovarian Serous Cy... (29.7%), Colon Adenocarcinoma: (17.3%), Clear Cell ... (12.5%), and Rectum Adenoc... (3.2%).
- Summary:** Studies (10), Biospecimens (719), Clinical (414), Files (430), Genes (12125).
- Table:** A table titled "Files data selected for download" with columns: Study, Data Category, File Type, Access, File Size. It lists 10 rows of TCGA Breast Cancer Proteome data.
- Modal Dialog:** A white box with the text: "You are trying to generate a file manifest that includes controlled data files. This will require you to log in through eRA and authorize DCF to access your NIH profile. Do you want to continue?" with "Cancel" and "Continue" buttons.
- Bottom:** A pagination bar showing page 1 of 10.

Upcoming feature

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",
  "hiv_positive": "string"
}
```

Top

Jupyter Notebook Example

NATIONAL CANCER INSTITUTE Proteomic Data Commons

This notebook attempts to demonstrate the following:

1. Using the Proteome Data Commons (PDC) API to retrieve relative protein expression data using the Common Data Analysis Pipeline (CDAP). More information on the PDC implementation is available [here](#).
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 identify clusters of samples (tumors) displaying similar expression patterns.

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 available [here](#). In brief, these values are log2 transformed ratio of the sample to the control channel 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_submitter_id currently recommended during the initial PDC development period. Higher values may be used in the future.

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

Next, the expression data GraphQL query is set up. Adding the `study_submitter_id` and `max_aliquots` to the query.

```
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_id
    + ''' aliquot_submitter_id:
    morphology:
    primary_diagnosis:
    tumor_grade:
    tumor_stage:
  }
},...'''
```

Get the clinical

Now we can define a function to make the GraphQL Post query. This will get called once 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('Aliquot')
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.

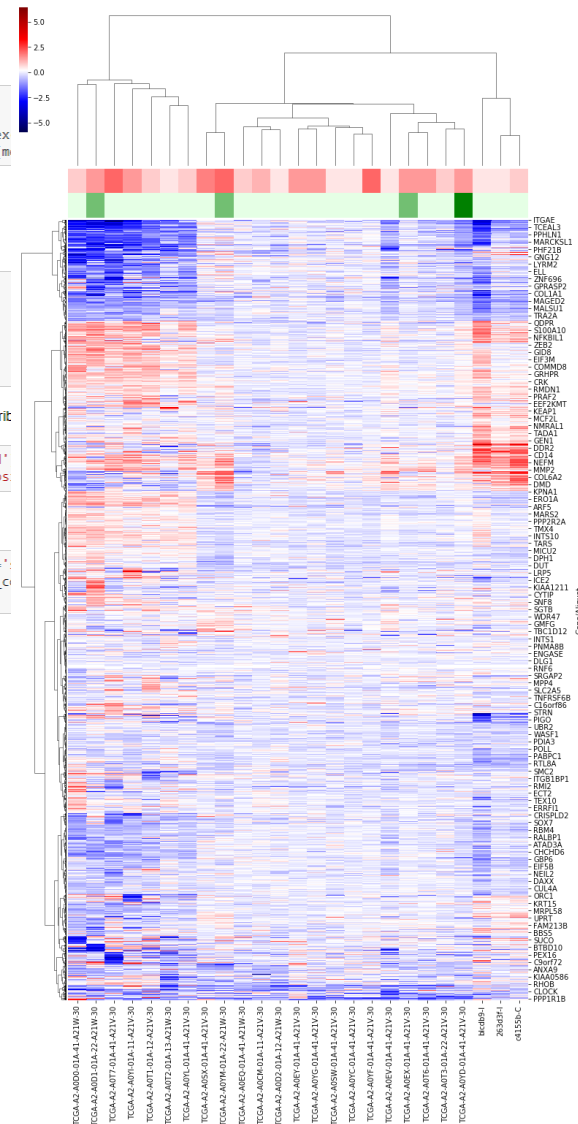
```
In [12]: def get_colors(df, name, color) -> pd.Series:
s = df[name]
su = s.unique()
colors = sns.light_palette(color, len(su))
lut = dict(zip(su, colors))
return s.map(lut)
```

Next, call `get_colors()` to map the `tumor_stage` and `primary_diagnosis` attributes.

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

And, finally, generate the large clustermap.

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