CPTAC: Accelerating Cancer Research Through Multiomic Integration and Data Sharing

Ratna Rajesh Thangudu, Ph.D.

NIH NATIONAL CANCER INSTITUTE

Agenda

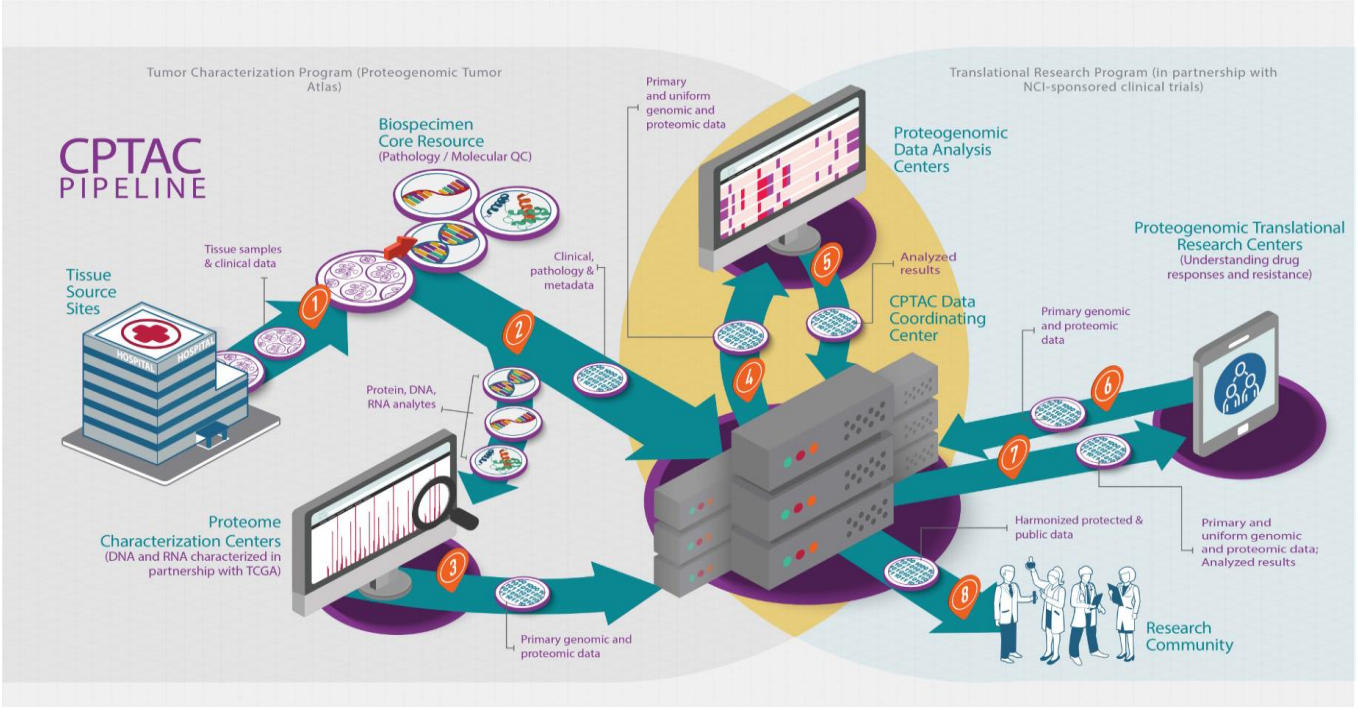
- CPTAC Historical Overview
- OCCPR's Leadership in Data Sharing
- Leveraging the CRDC
- Future of Data Access with PDC

- **CPTAC is** a national effort to accelerate the understanding of the molecular basis of cancer through large-scale proteome and genome analysis, known as proteogenomics.

- **Objective:** To systematically identify proteins that derive from alterations in cancer genomes and related biological processes, in order to understand the molecular basis of cancer that is not fully elucidated or not possible through genomics and to accelerate the translation of molecular findings into the clinic.

CPTAC

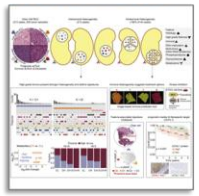
- **Collaborative effort by a multidisciplinary team**
- **Committed to open data sharing in proteomics**
- **Key Milestones**:
  - **2006-2011**: Initial focus on standardizing proteomic technologies for cancer research.
  - **2012-2016**: Illumination of new tumor biology by focusing on TCGA tumor collection.
  - **2016-Now**: Advancing proteogenomics and translational research with prospective tumor collection.



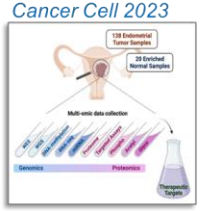
Tumor Characterization Program

CPTAC produces public resources of high-quality proteogenomic data of human tumors for hypothesis-driven science

Kidney
Cancer Cell 2023

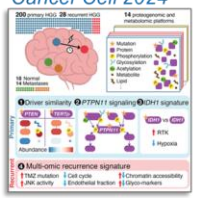


Endometrial
Cancer Cell 2023



Kidney (non-ccRCC)
Cell Rep Med 2024

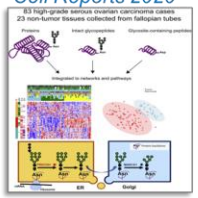
GBM (adult HGG)
Cancer Cell 2024



12 Cancer types

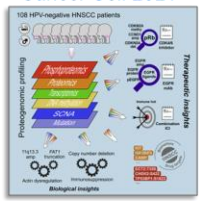
16 Flagship studies

Ovarian
Cell Reports 2020

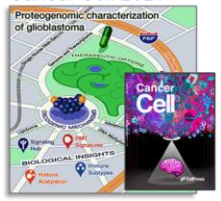


GBM (pediatric)
Cell 2020

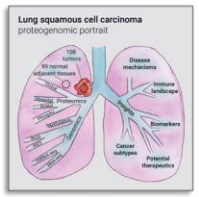
Head & Neck
Cancer Cell 2021



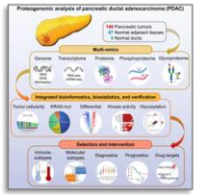
GBM (adult)
Cancer Cell 2021



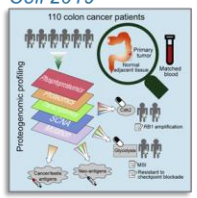
Lung squamous
Cell 2021



Pancreas
Cell 2021



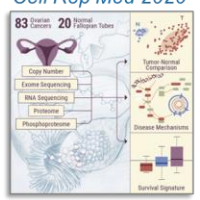
Colon
Cell 2019



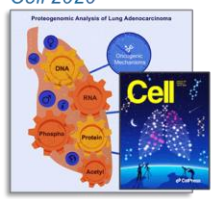
Kidney
Cell 2019

Endometrial
Cell 2020

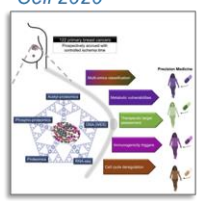
Ovarian
Cell Rep Med 2020

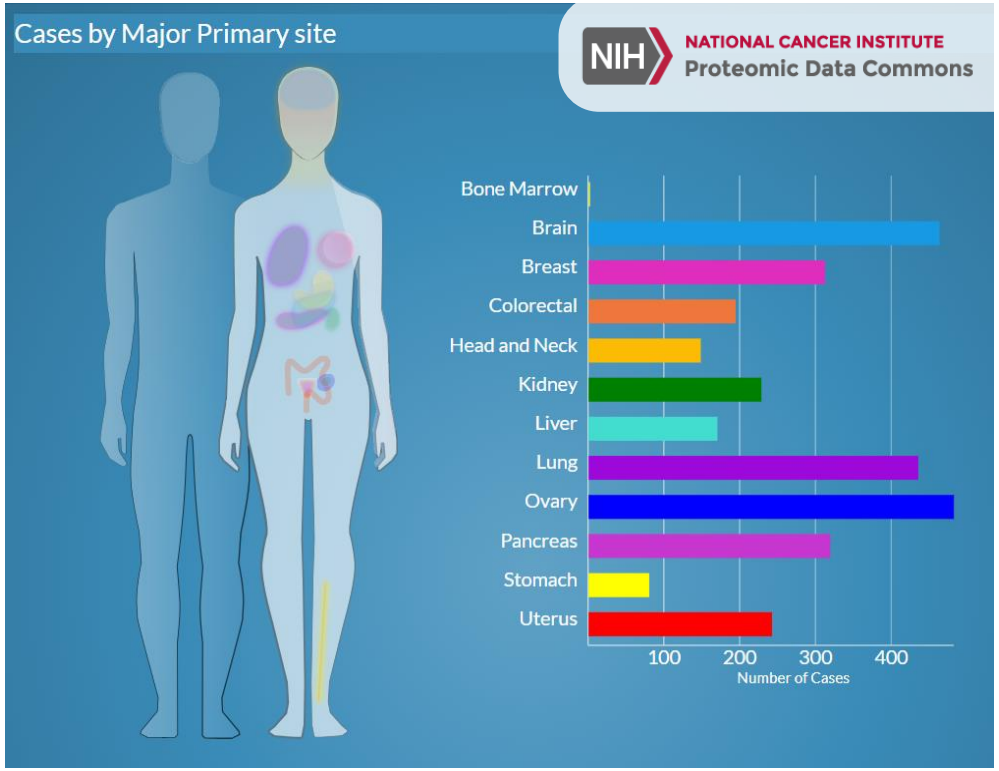


Lung adeno
Cell 2020



Breast
Cell 2020

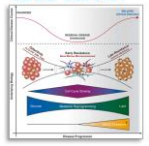




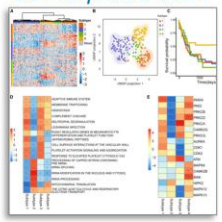
Source: Dr Henry Rodriguez, OCCPR, NCI

Translational Research Program

AML
Cancer Cell 2021

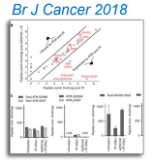


AML
Cell Rep Med 2024

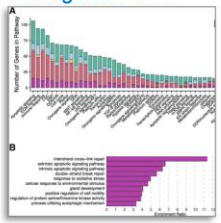


Phase 1b/2 trial: acute myeloid leukemia (BeatAML)
(NCT01728402; *Cell Rep Med* 2024, PMID 38232702)

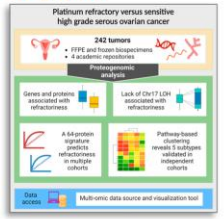
Ovarian
Br J Cancer 2018



Ovarian
Oncogene 2021

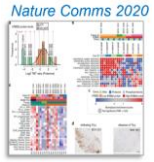


Ovarian
Cell 2023

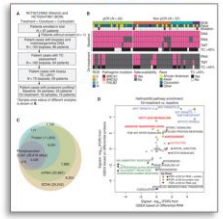


Refractory high grade serous ovarian cancer
(*Cell* 2023, PMID 37541199)
[CT access to be submitted]

Breast
Nature Comms 2020



Breast
Cancer Disc 2022



Phase 3 trial: neoadjuvant of paclitaxel-trastuzumab w/ or wo/ lapatinib in HER2+ breast cancer
(NCT02547987 [Alliance/CALGB 40601]) manuscript submitted

Phase 2 trial: neoadjuvant carboplatin and docetaxel in TNBC (CADENCE)
(NCT02547987 & NCT02124902; *Cancer Discovery* 2022, PMID 36001024)

External Collaborations

(expanding analytical expertise to trials outside the CPTAC network [pilot demonstration studies])

Phase 1 trial: locoregional CAR T cells in glioma
(NCT04185038; *Cancer Discov* 2023, PMID 36259971)

Phase 1/2 trial: ATR inhibitor in relapsed/refractory CLL
(NCT03328273; Ther Adv Hematol 2023, PMID 37273420)

Phase 2 trial: pembrolizumab in lymphoma
(NCT02243579; Front Oncol 2023, PMID 37205196)

Source: Dr Henry Rodriguez, OCCPR, NCI

Ongoing Programs

Tumor Characterization

- LUAD [CPTAC-Taiwan ICPC]
- AML
- GBM AYA
- Oligodendroglioma
- Sarcoma
- Melanoma
- Gastric
- LSCC
- Prostate
- Thyroid
- Liver
- LUAD onco negative

Translational/Clinical

- Melanoma
- Multiple myeloma
- AML (Beat AML)
- NSCLC (ALCHEMIST)

Collaborations

Children's Brain Tumor Network
Until every child is cured

ICPC
INTERNATIONAL CANCER PROTEOGENOME CONSORTIUM

APOLLO

NATIONAL CANCER INSTITUTE
PATIENT-DERIVED MODELS REPOSITORY

CCDI
CHILDHOOD CANCER DATA INITIATIVE

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

LINDONLIGHT COLLECTIVE
TO END PEDIATRIC BRAIN CANCER

Championing Proteomic Data Sharing

OCCPR has championed international data-sharing policies in proteomics since its inception.

Shaped by Key Data Sharing Policies:

- **Bermuda (1996)**: Early push for rapid data release in genomics.
- **Fort Lauderdale (2003)**: Formalized immediate public access to data from large-scale projects.
- **Amsterdam (2008)**: Established for **proteomics**, ensuring open data release without restrictions.
- **Sydney (2010)**: Emphasized open data sharing with additional considerations for clinical trials.

Impact on Cancer Research:

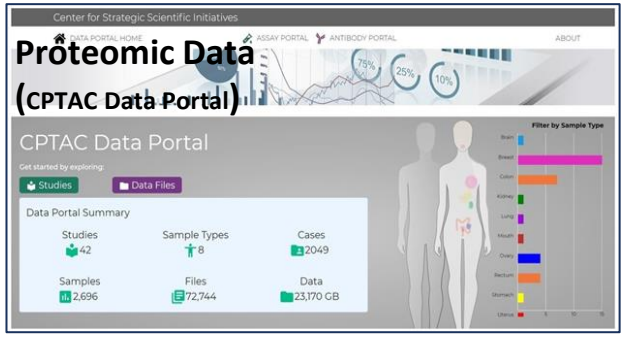
OCCPR's adoption of these principles ensured timely public release of high-quality proteomic data, accelerating cancer research discoveries.

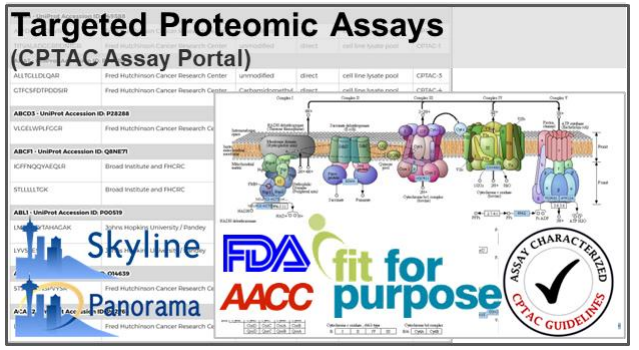
Building the Foundation Data Sharing

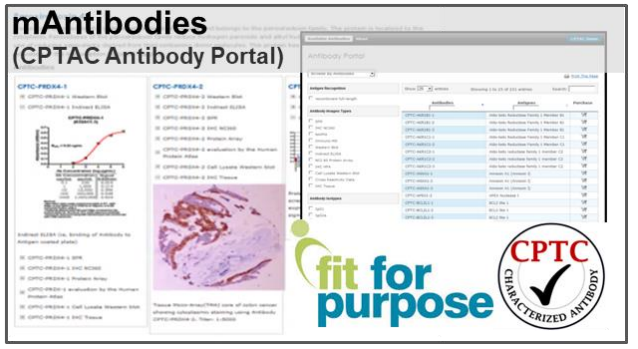
Early OCCPR Platforms: Key resources developed to support open access of proteomic data.

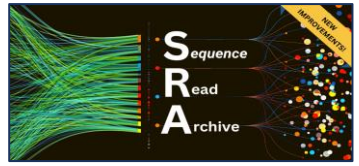
Collaboration with other NCI and NCBI resources: To distribute imaging and genomic data.

Open Science: These resources embodied OCCPR's commitment to transparent, unrestricted data sharing, paving the way for PDC and CRDC.





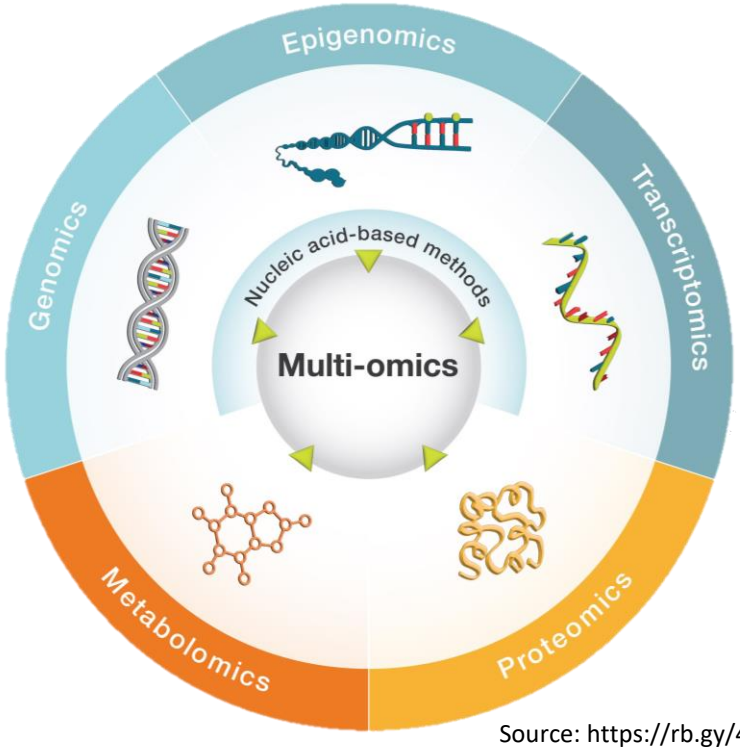




CPTAC Comprehensive Molecular Data

- **Multi-Omics Analyses**:
  - Genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - Lipidomics

Protein Post-Translational Modifications (PTMs)
- Phosphoproteome
- Acetylome
- Glycoproteome
- Ubiquitylome



Source: https://rb.gy/408zi1

Clinical Data: Patient demographics, diagnosis, treatment, exposure and follow-up data.

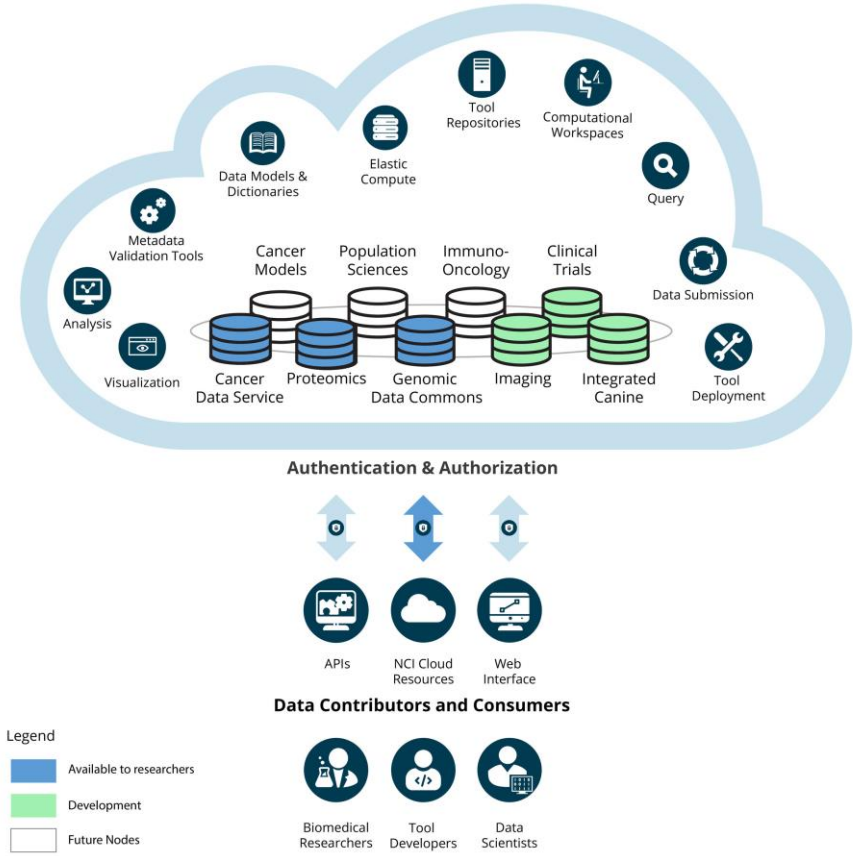
Histology & Radiology Images: Comprehensive imaging data for in-depth tissue and tumor analysis.

Sharing Complex, High-Value Data: The Need for Specialized Resources

Why Specialized Resources Are Needed:

- Handling **complex, large-scale multi-omics** data.
- Ensuring **data standardization** and **harmonization** across studies.
- Providing access to cutting-edge **analytical tools** and **computational infrastructure**.
- Making data **FAIR** (Findable, Accessible, Interoperable, Reusable).
- Supporting the research community in reusing data to drive **scientific breakthroughs**.

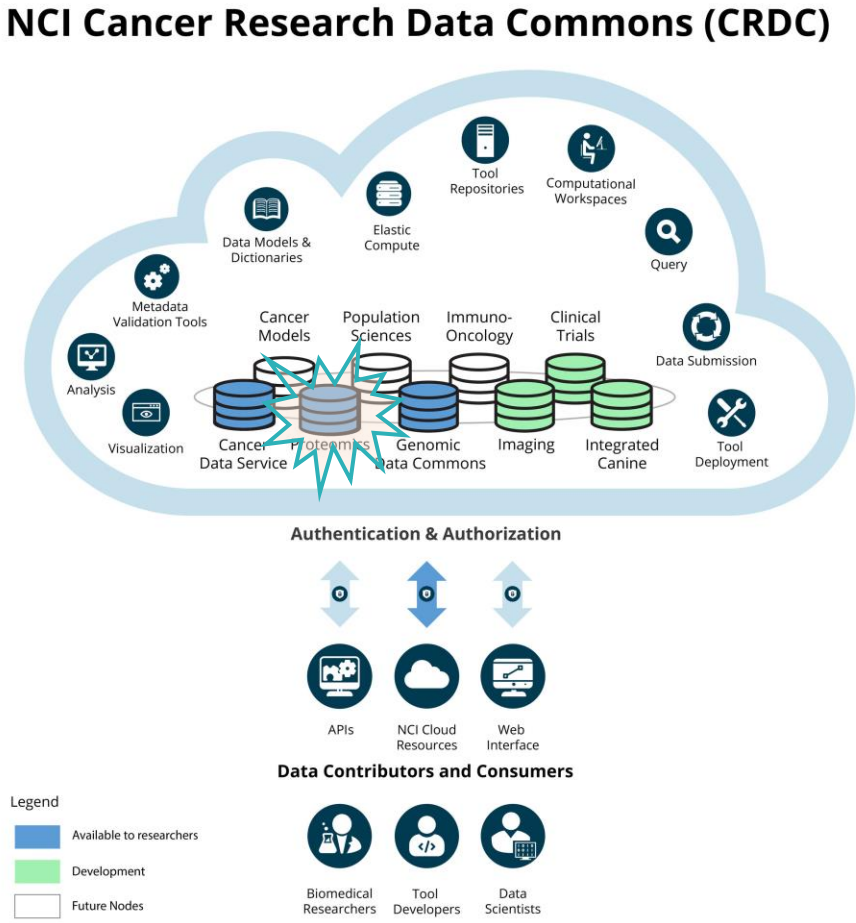
NCI Cancer Research Data Commons (CRDC)



Leveraging the CRDC

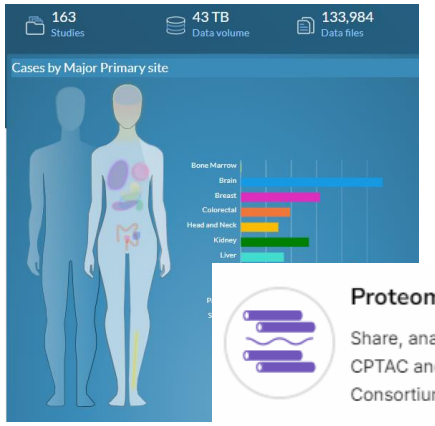
Multiomic Data Resources:

- Established **Proteomic Data Commons** for distributing mass spectrometry based proteomic data, operating as a node with in CRDC
- Migrated all of the historic data from **CPTAC data portal**
- Identified other nodes within CRDC for distribution of multiomics data types
- **Genomic Data Commons** for Genomic and Transcriptomic data
- **Imaging Data Commons** for pathology and radiology images



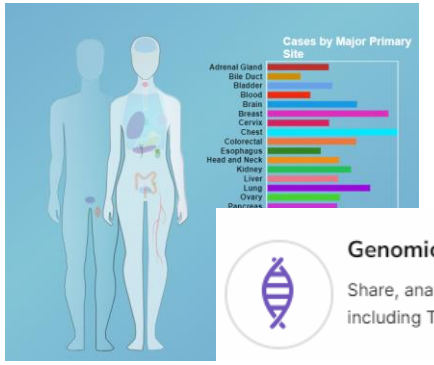
Serving as Knowledgebases

- Adherence to FAIR Guidelines
- Common Data Models & Standardization
- Harmonization Pipelines
- Sophisticated Tools for Data Exploration
- Analysis tools for protein/gene expression clustering, building cohorts, and survival analysis.
- APIs for programmatic access



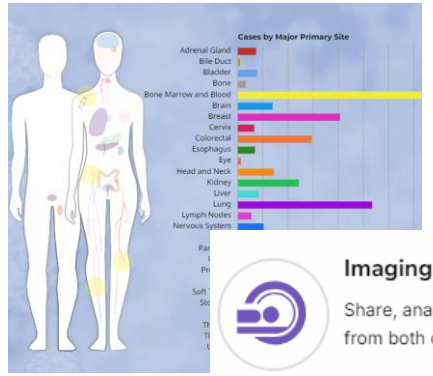
Proteomic Data Commons (PDC)

Share, analyze, and visualize proteomic data, such as CPTAC and The International Cancer Proteogenome Consortium (ICPC).



Genomic Data Commons (GDC)

Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

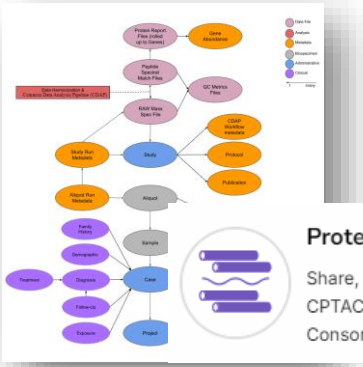


Imaging Data Commons (IDC)

Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.

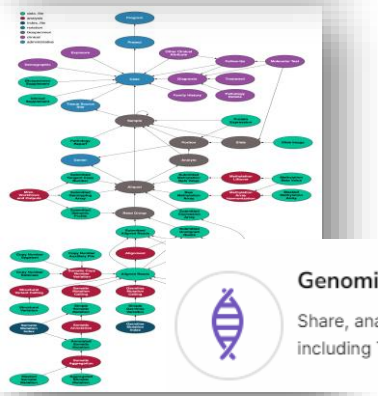
Serving as Knowledgebases

- Adherence to FAIR Guidelines
- Common Data Models & Standardization
- Harmonization Pipelines
- Sophisticated Tools for Data Exploration
- Analysis tools for protein/gene expression clustering, building cohorts, and survival analysis.
- APIs for programmatic access



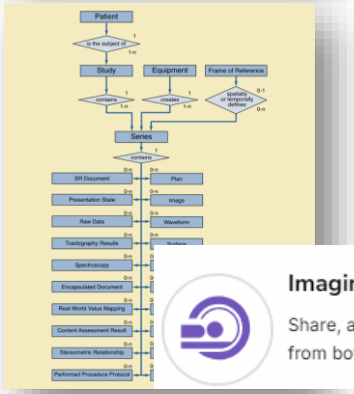
Proteomic Data Commons (PDC)

Share, analyze, and visualize proteomic data, such as CPTAC and The International Cancer Proteogenome Consortium (ICPC).



Genomic Data Commons (GDC)

Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

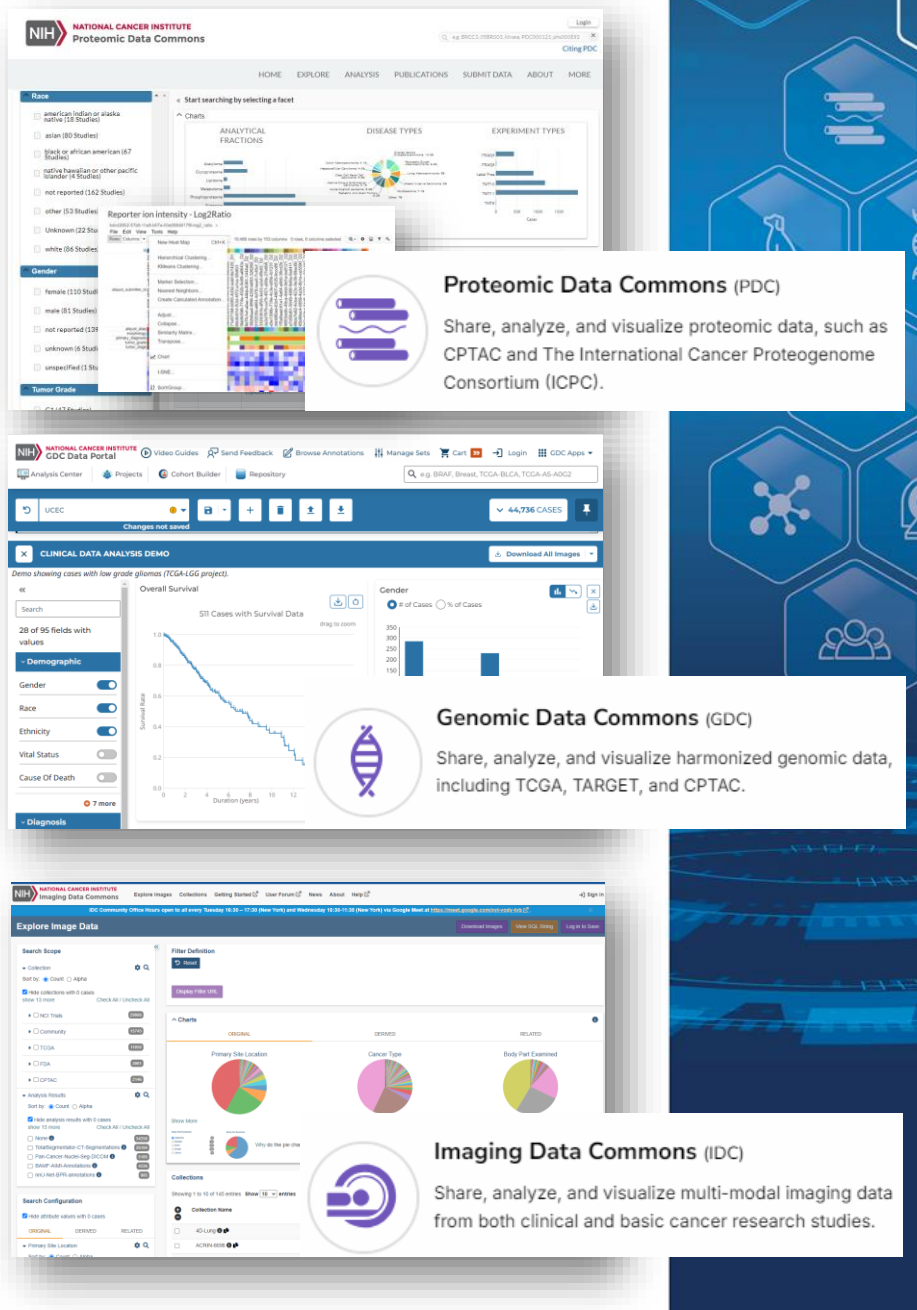


Imaging Data Commons (IDC)

Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.

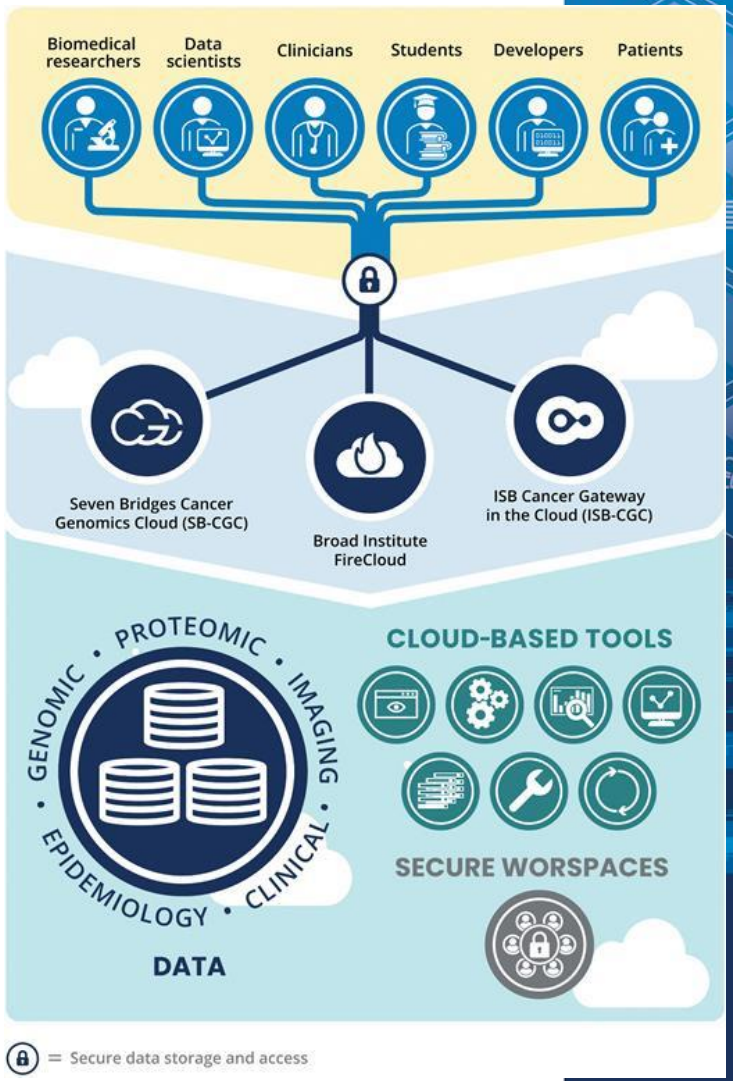
Serving as Knowledgebases

- Adherence to FAIR Guidelines
- Common Data Models & Standardization
- Harmonization Pipelines
- Sophisticated Tools for Data Exploration
- Analysis tools for protein/gene expression clustering, building cohorts, and survival analysis.
- APIs for programmatic access



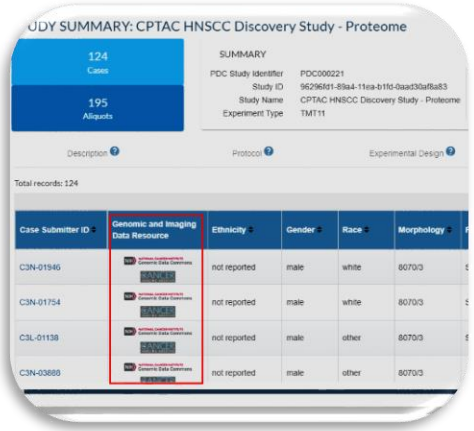
Supporting Computational Analysis

- Simplifying access to complex multiomic data for analysis
  - Primary data analysis – SB-CGC; FireCloud
  - Correlating the processed data – ISB CGC
- Brings CPTAC data closer to the scalable compute infrastructure
- Versatile tools and analysis pipelines



Expanding CPTAC Data Sharing

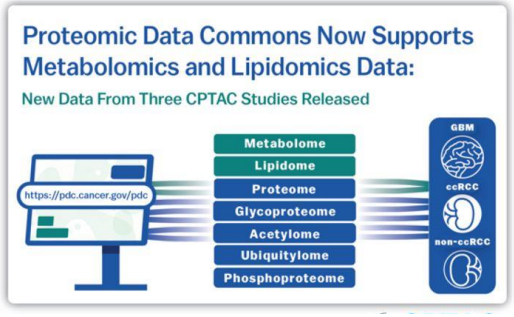
- Annual updates of the clinical data
- Cross-referencing across portals to ensure discoverability of multiomics data.
- Capture highly complex features of the cohorts
  - such as metastatic tumors of different cancer origin
  - bridging samples across different cohorts across international programs,
  - capturing multiple protocols used in studies, to name a few



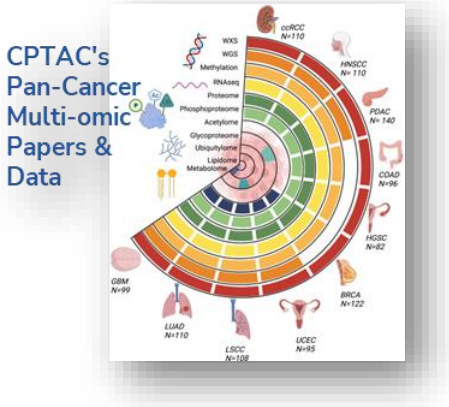
PDC Cross referencing to GDC and TCIA

Expanding CPTAC Data Sharing

- Expanded PDC to supporting CPTAC Metabolomic data Lipidomic Data

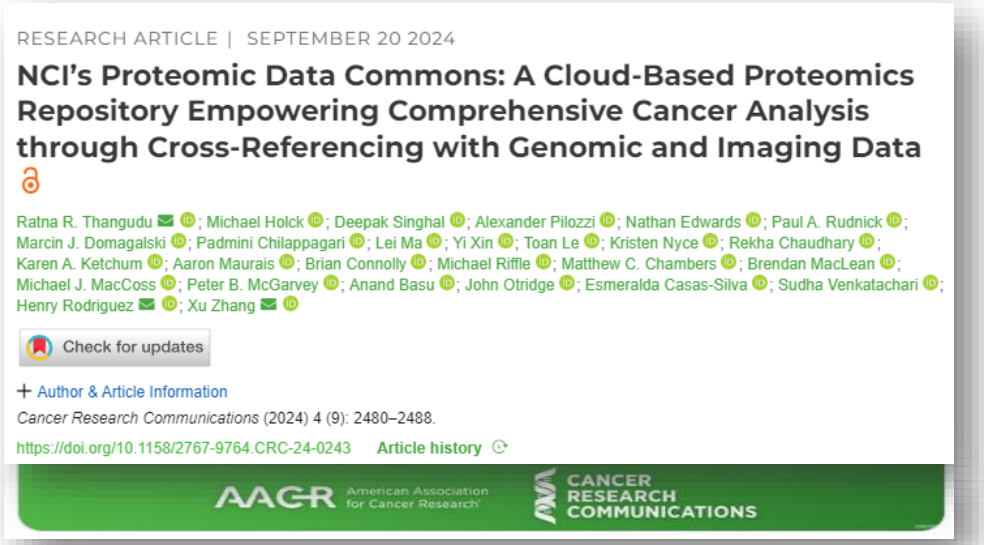


- Established a CPTAC Pan-Cancer resource on PDC to distribute data from the comprehensive data and papers



Expanding CPTAC Data Sharing

https://aacrjournals.org/cancerres/issue/84/9



https://doi.org/10.1158/2767-9764.CRC-24-0243

CPTAC Public Resources (data, protocols, reagents)

PROTEOMICS.CANCER.GOV

Antibody Portal

- >940 fit-for-purpose mAbs
- >6,300 units sold

Proteomic Data Commons

- >1.3 PB downloaded
- >80K users / >152 countries

Genomic Data Commons

- >4.1 PB downloaded

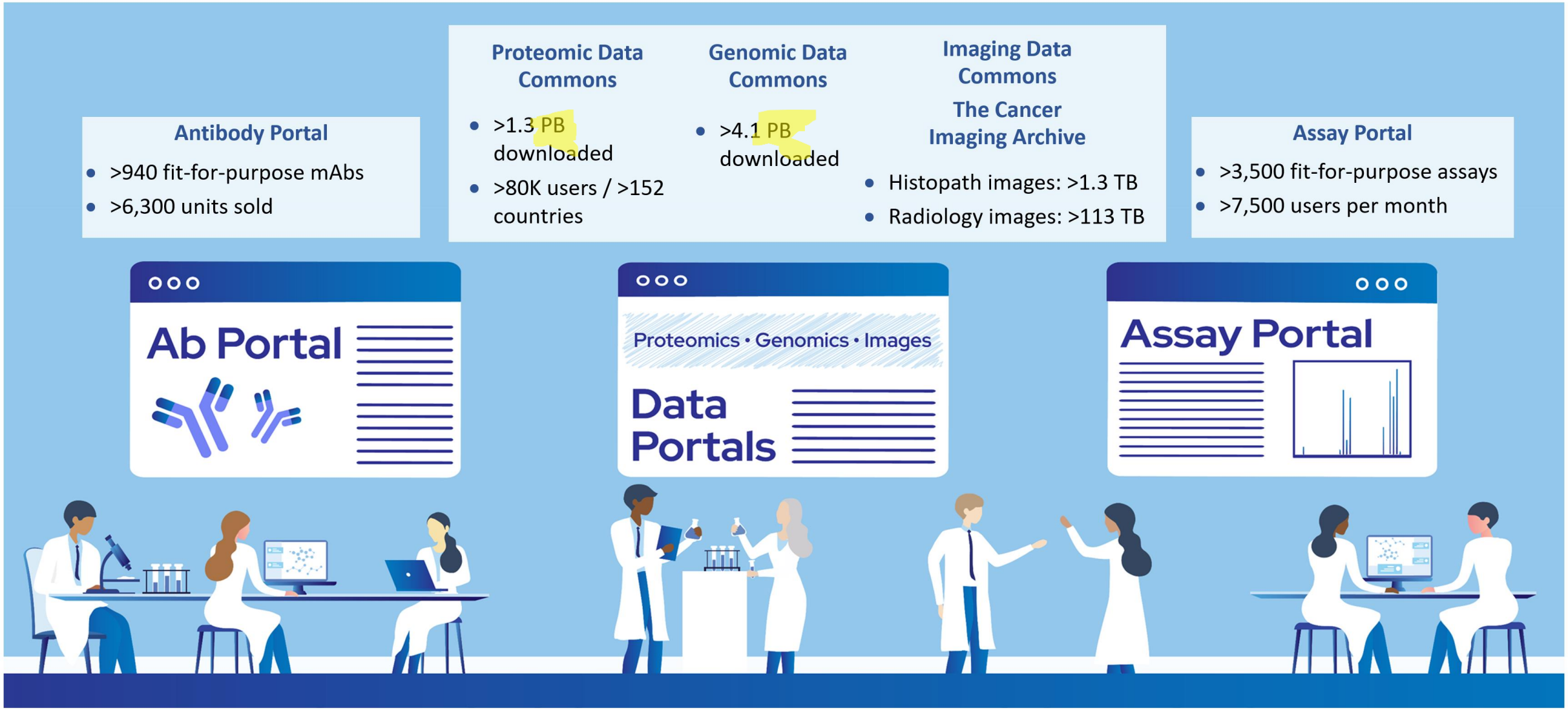
Imaging Data Commons

The Cancer Imaging Archive

- Histopath images: >1.3 TB
- Radiology images: >113 TB

Assay Portal

- >3,500 fit-for-purpose assays
- >7,500 users per month



Source: Dr Henry Rodriguez, OCCPR, NCI

Resources

CPTAC Program	https://proteomics.cancer.gov	cancer.proteomics@mail.nih.gov
Proteomic Data commons	https://pdc.cancer.gov	pdchelpdesk@mail.nih.gov
Genomic Data Commons	https://gdc.cancer.gov	support@nci-gdc.datacommons.io
Imaging Data Commons	https://portal.imaging.datacommons.cancer.gov	support@canceridc.dev
The Cancer Imaging Archive	https://www.cancerimagingarchive.net	help@cancerimagingarchive.net
Assay Portal	https://assays.cancer.gov/	cancer.proteomics@mail.nih.gov
Antibody Portal	https://antibodies.cancer.gov/	cancer.proteomics@mail.nih.gov

Acknowledgements

NCI
OCCPR
Henry Rodriguez
Xu Zhang
Mehdi Mesri
Ana Robles
Eunkyung An
Tara Hiltke
Jasmin Bavarva

CPTAC DCC
Karen Ketchum
Yin Lu
Shuang Cai

Leidos
Sudha Venkatachari
John Otridge

CRDC
Emi Casas-Silva
CRDC Stakeholders
CBIIT Leadership

PDC
Michael Holck
Alexander Pilozzi
Deepak Singhal
Development Team

SME
Michael MacCoss
Nathan Edwards
Paul Rudnick

CPTAC Consortium Data Donors