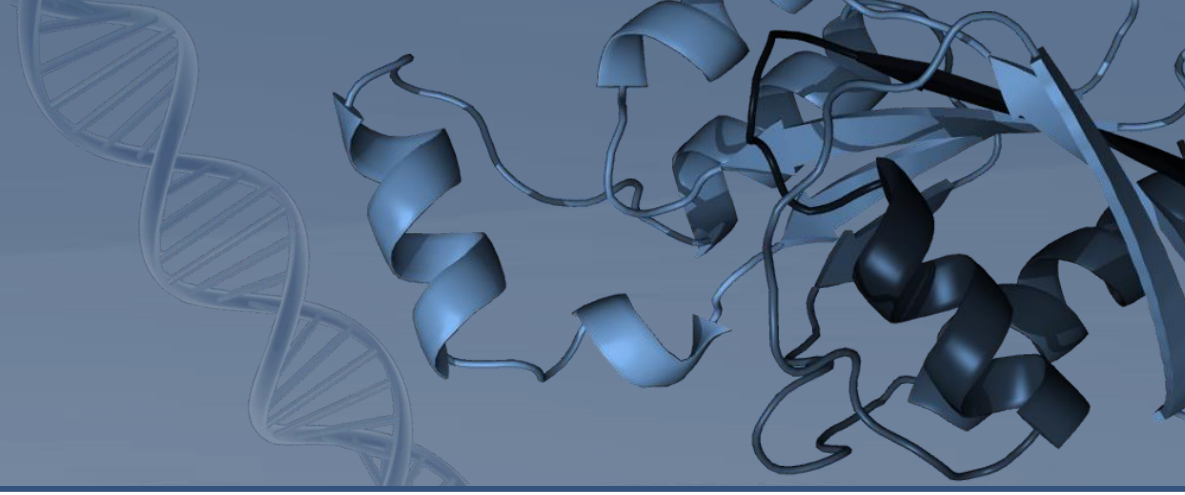
OFFICE OF CANCER CLINICAL PROTEOMICS RESEARCH



Proteogenomic Approaches

Chris Kinsinger, PhD

Program Director

Office of Cancer Clinical Proteomics Research

Bring Your Own Bioinformatics

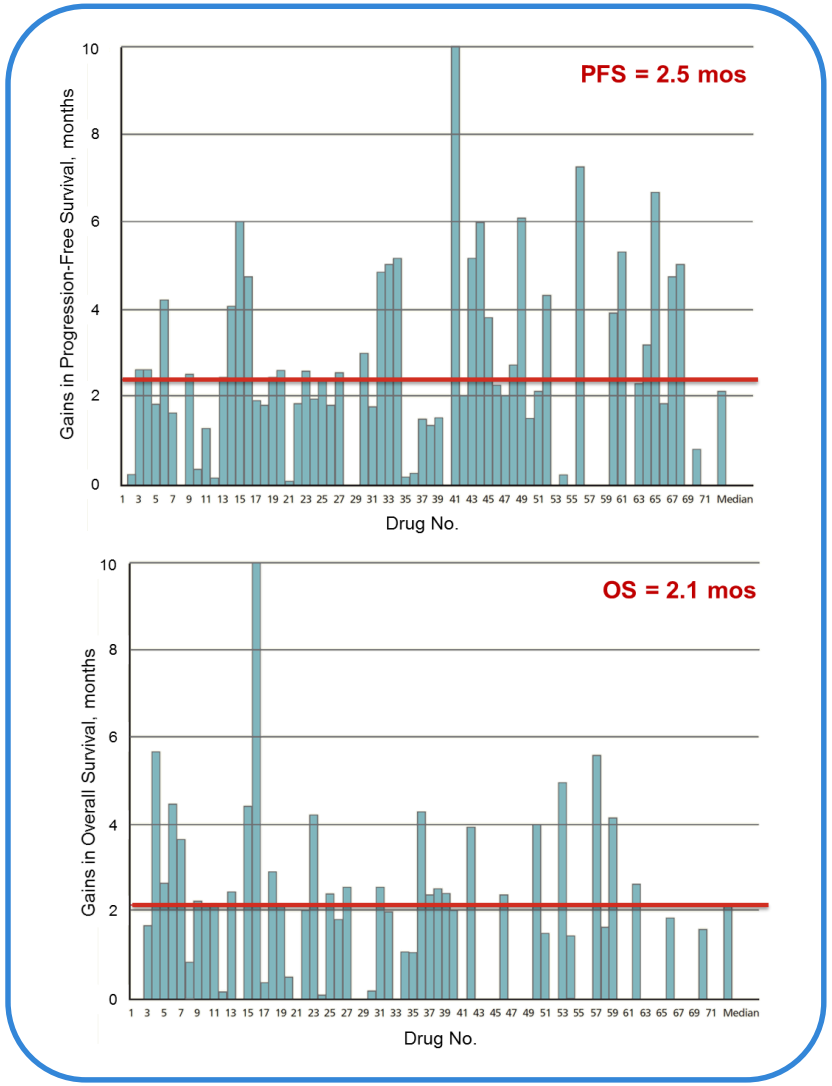
November 2019

NIH NATIONAL CANCER INSTITUTE

Outline

- Proteogenomic approaches
  - Sequence-centric
  - Proteogenomic relationships
  - Integrative models
- Accessing the data
  - Cancer research data commons
  - Proteomic data commons
  - Genomic data commons
  - The Cancer Imaging Archive

Drugs approved by FDA for advanced cancer



Gains in overall survival for the 71 drugs approved by FDA from 2002 to 2014 for advanced cancer

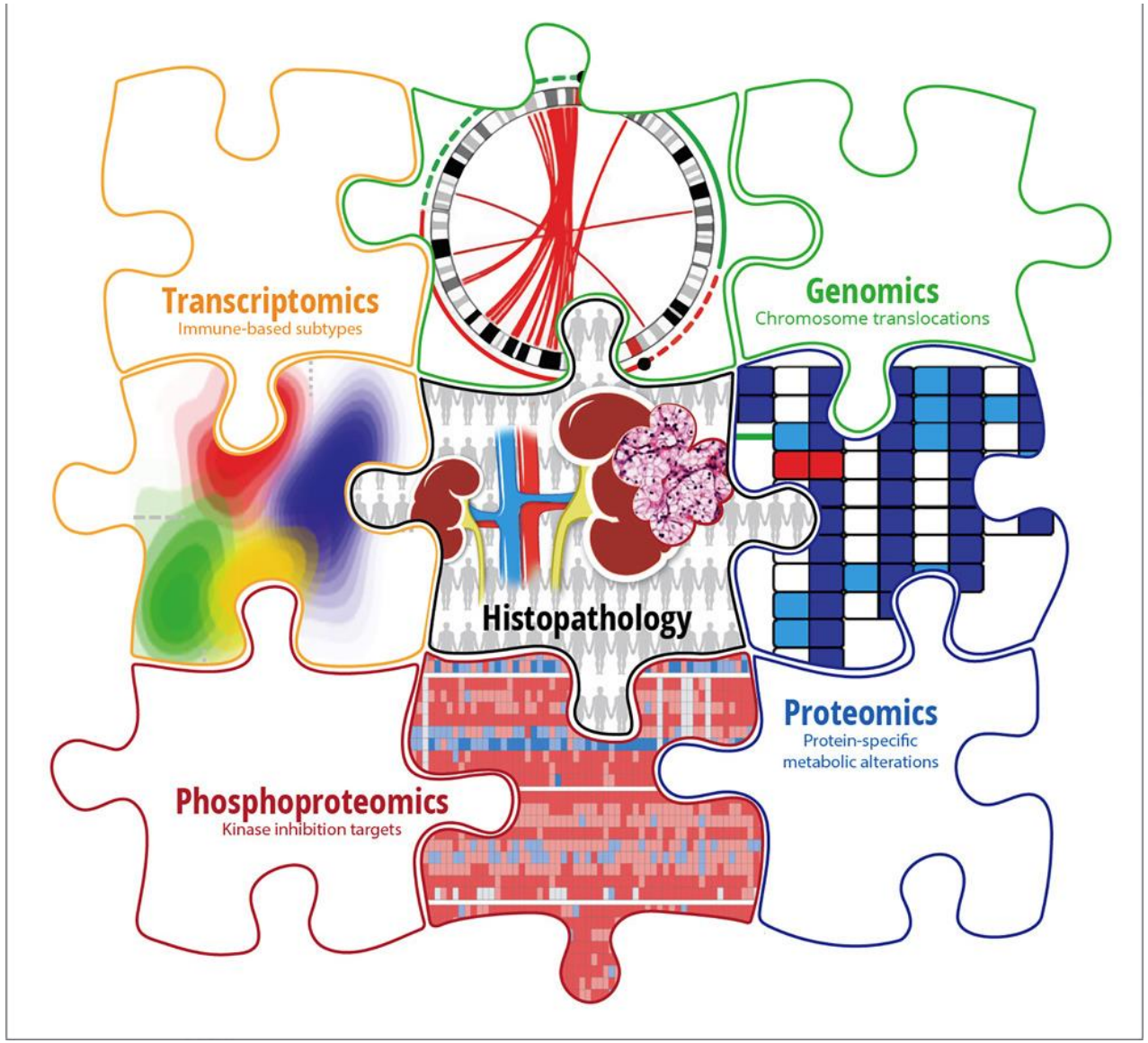
PFS: 2.5 mos

OS: 2.1 mos

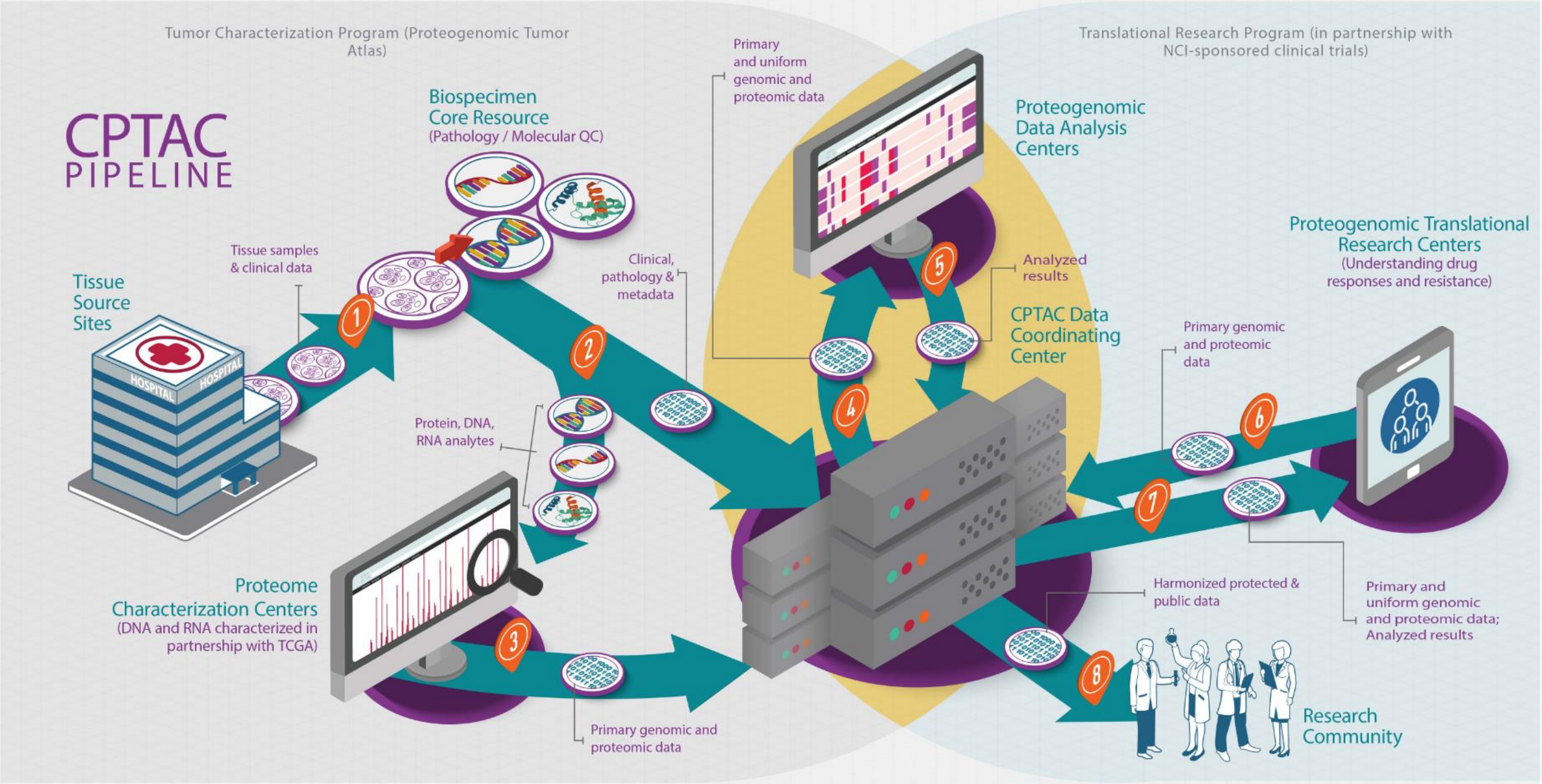
Solid tumors

Fojo et al. *JAMA Otolaryngol Head Neck Surg.* 2014;140:1225-1236.

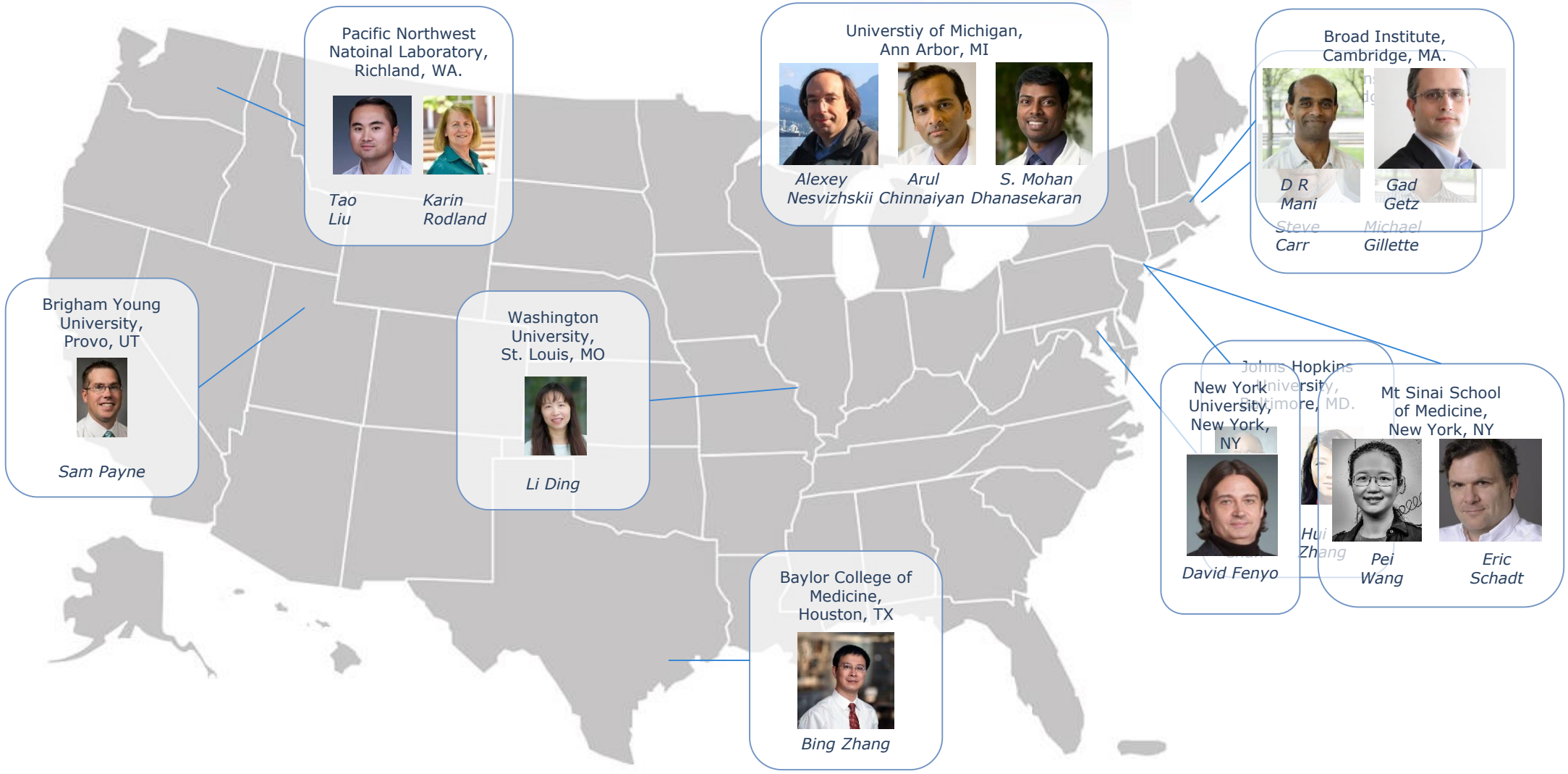
Proteogenomics



Clark, DJ et al. *Cell*, **2019**, *179*, 964-983


Tumor Characterization Program (Proteogenomic Tumor Atlas)
Translational Research Program (in partnership with NCI-sponsored clinical trials)
CPTAC PIPELINE
Biospecimen Core Resource
(Pathology / Molecular QC)
Primary and uniform genomic and proteomic data
Proteogenomic Data Analysis Centers
Proteogenomic Translational Research Centers
(Understanding drug responses and resistance)
Tissue samples & clinical data
Tissue Source Sites
HOSPITAL
HOSPITAL
Clinical, pathology & metadata
Analyzed results
CPTAC Data Coordinating Center
Primary genomic and proteomic data
Protein, DNA, RNA analytes
Proteome Characterization Centers
(DNA and RNA characterized in partnership with TCGA)
Harmonized protected & public data
Primary and uniform genomic and proteomic data; Analyzed results
Primary genomic and proteomic data
Research Community
1
2
3
4
5
6
7
8

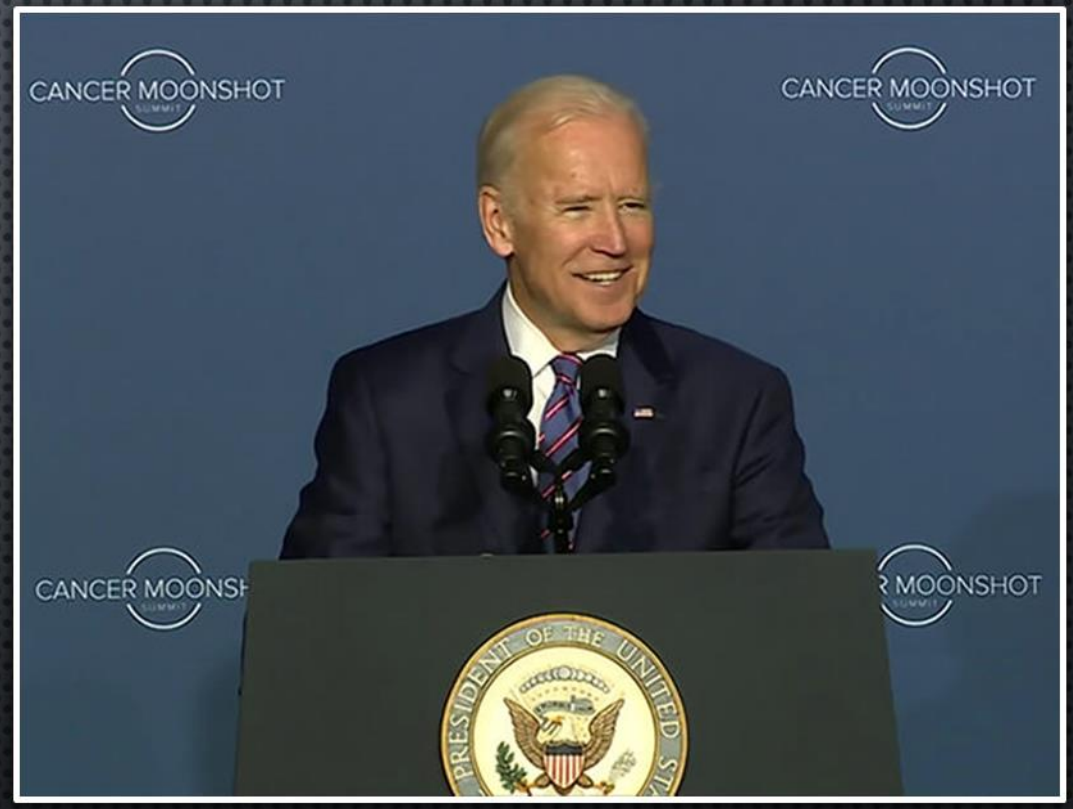
The players


Pacific Northwest Natoinal Laboratory, Richland, WA.
Tao Liu
Karin Rodland
Universtiy of Michigan, Ann Arbor, MI
Alexey Nesvizhskii
Arul Chinnaiyan
S. Mohan Dhanasekaran
Broad Institute, Cambridge, MA.
D R Mani
Gad Getz
Steve Carr
Michael Gillette
Brigham Young University, Provo, UT
Sam Payne
Washington University, St. Louis, MO
Li Ding
Baylor College of Medicine, Houston, TX
Bing Zhang
New York University, New York, NY
David Fenyo
Johns Hopkins University, Baltimore, MD.
Hui Zhang
Mt Sinai School of Medicine, New York, NY
Pei Wang
Eric Schadt

White House Cancer Moonshot

GOAL: Make 10 yrs of progress against cancer in 5 yrs.

CANCER MOONSHOT



Joe Biden
Vice President of the United States

APOLLO Clinical Network

(**A**pplied **P**roteogenomics **O**rganizational **L** **L**earning and **O**utcomes)

The White House
Office of the Press Secretary

For Immediate Release

June 28, 2016

FACT SHEET: At Cancer Moonshot Summit, Vice President Biden Announces New Actions to Accelerate Progress Toward Ending Cancer As We Know It

WASHINGTON, D.C. – Today, the Cancer Moonshot is hosting a summit at Howard University, in Washington, D.C. as part of a national day of action that also includes more than 270 events in

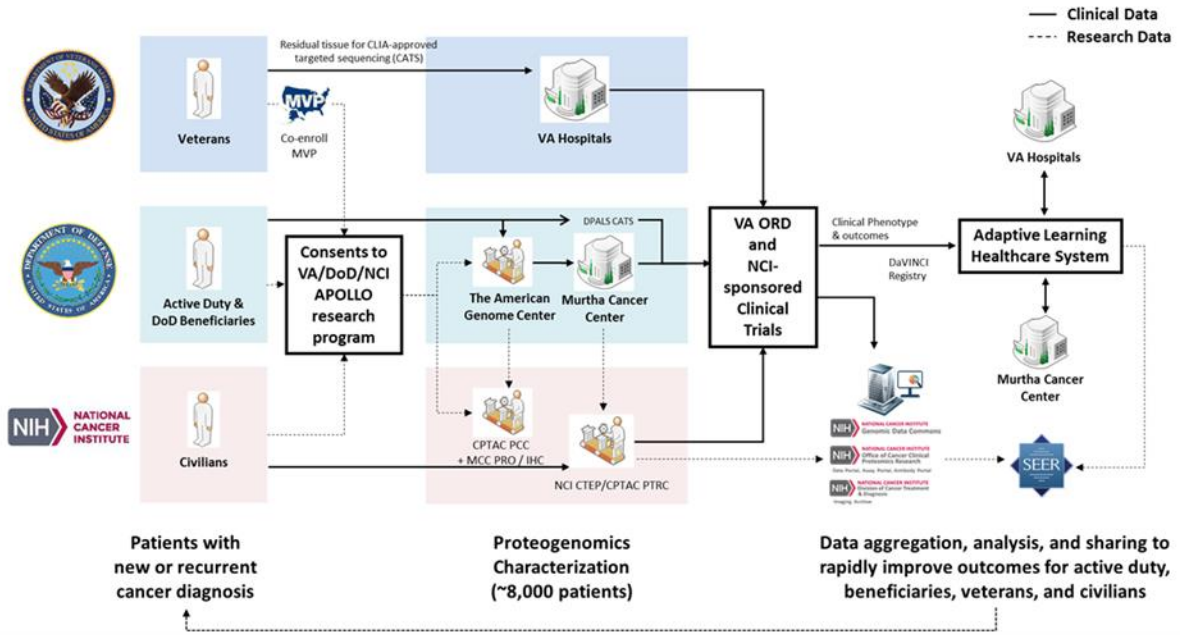
APOLLO Inaugural Leadership Meeting
29 August 2016


Col. Craig Shriver (DoD)


Jennifer Lee (VA)

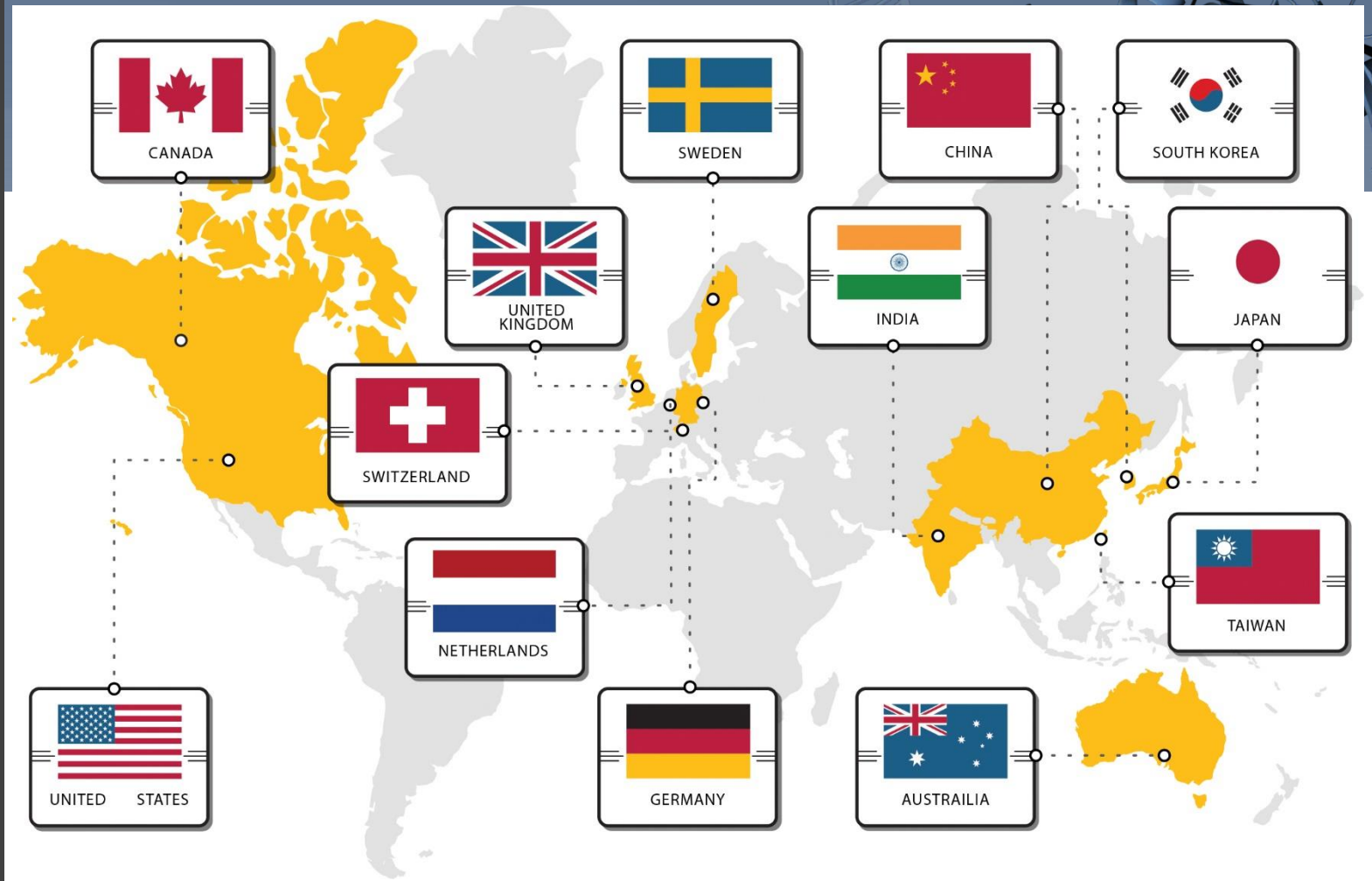

Henry Rodriguez (NCI)

- **Tri-lateral Collaboration:** NCI, Department of Defense, Veterans Administration
- Nation's first healthcare system where cancer patients routinely screened for genomic abnormalities and proteomic information to match tumor types to targeted therapies



13 countries from **33** institutions

13 cancer types selected

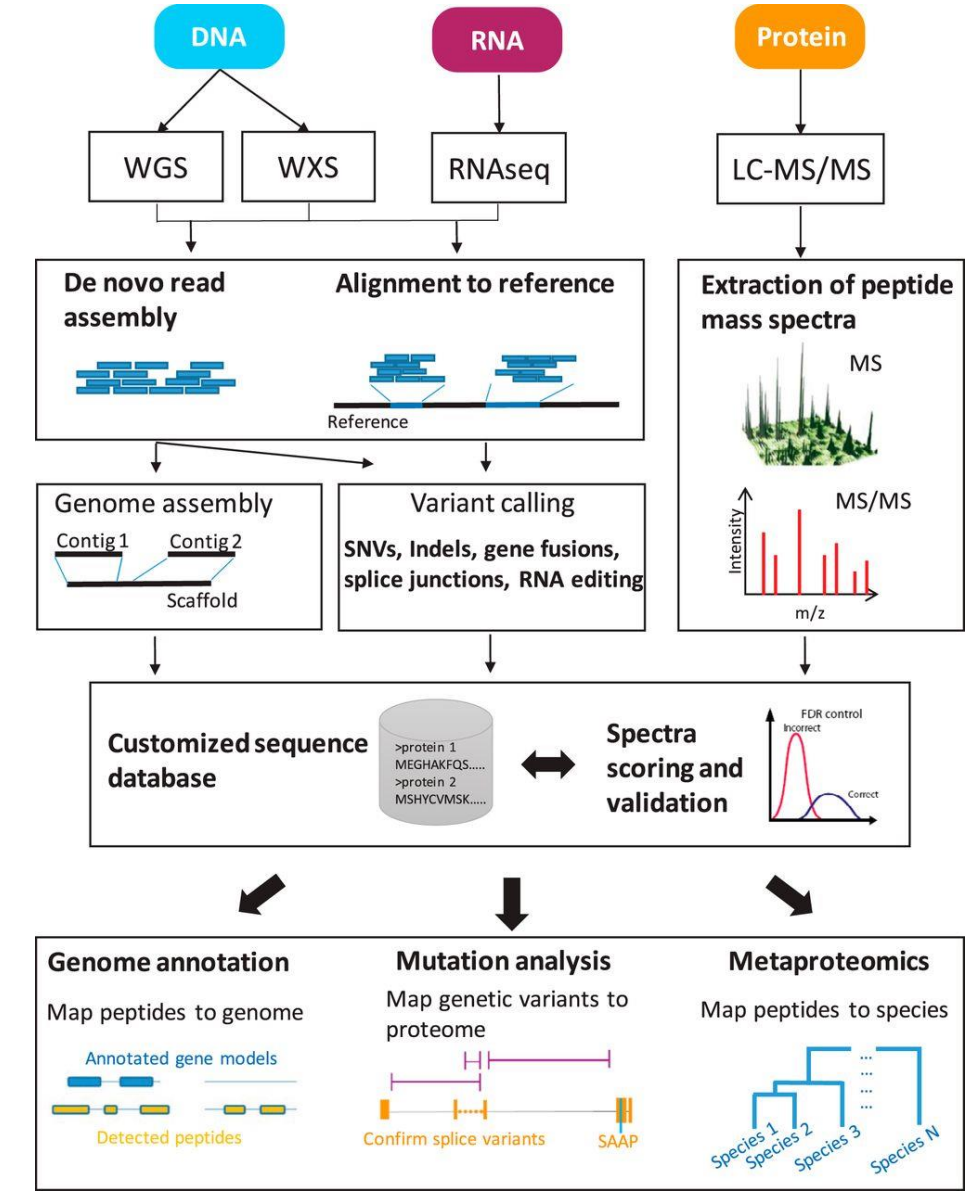


ICPC

INTERNATIONAL CANCER PROTEOGENOME CONSORTIUM

Sequence-centric proteogenomics

- QUILTS
- FDR caution

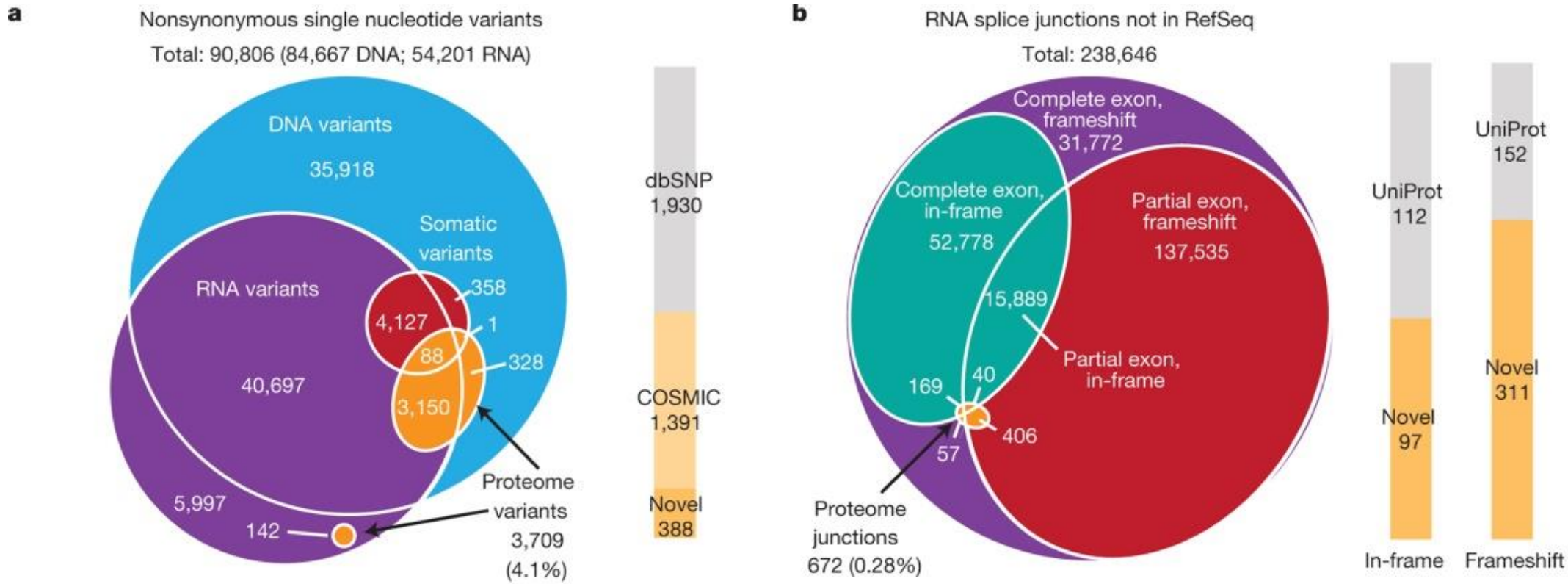


Ruggles, et al. *MCP* 16: 10.1074/mcp.MR117.000024, 959, 2017.

QUILTS – Ruggles et al. *MCP* **2017**, 15.3 p. 1060

FDR cautions - Nesvizhskii, AI. “Proteogenomics: concepts, applications, and computational strategies.” *Nature Methods*. **2014**, 11, 1114.

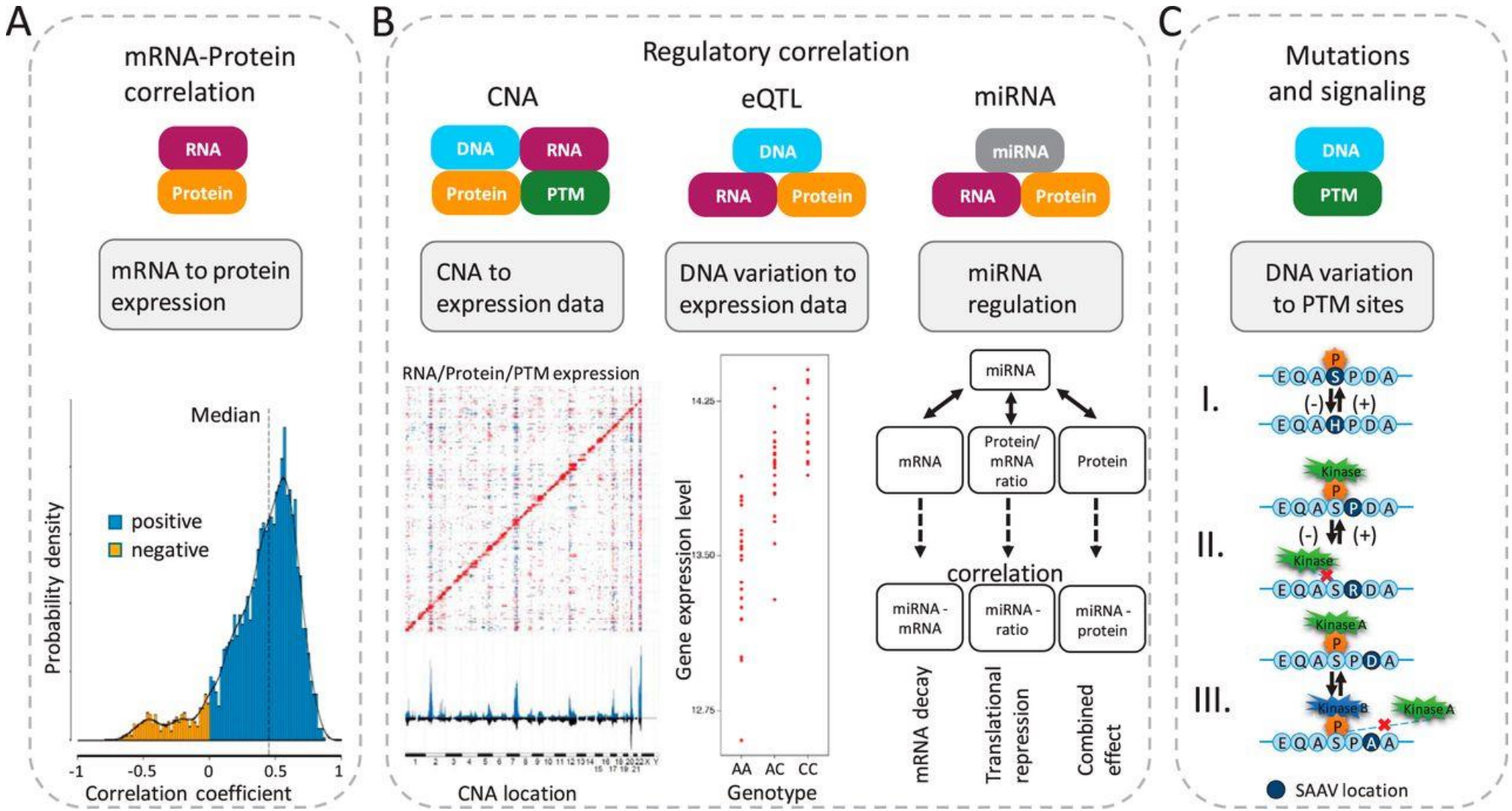
Proteogenomic mutation detection in breast cancer



- Most single amino acid variants previously reported
- Few splice junctions detected, but many are novel

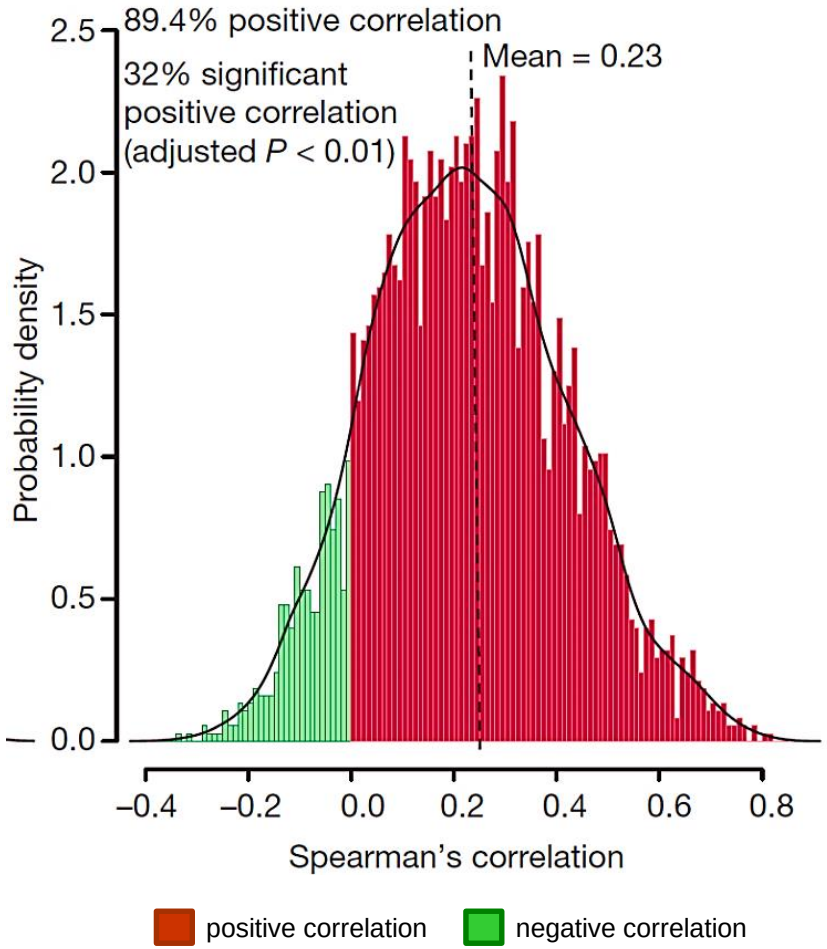
Mertins, et al. “Proteogenomics connects somatic mutations to signaling in breast cancer *Nature* **2016,** doi:10.1038/nature18003

Analysis of proteogenomic relationships



Gene expression and protein abundance have fairly low correlation

mRNA and protein abundance correlation for individual genes across all colon tumors (samples)



Mean Correlation:

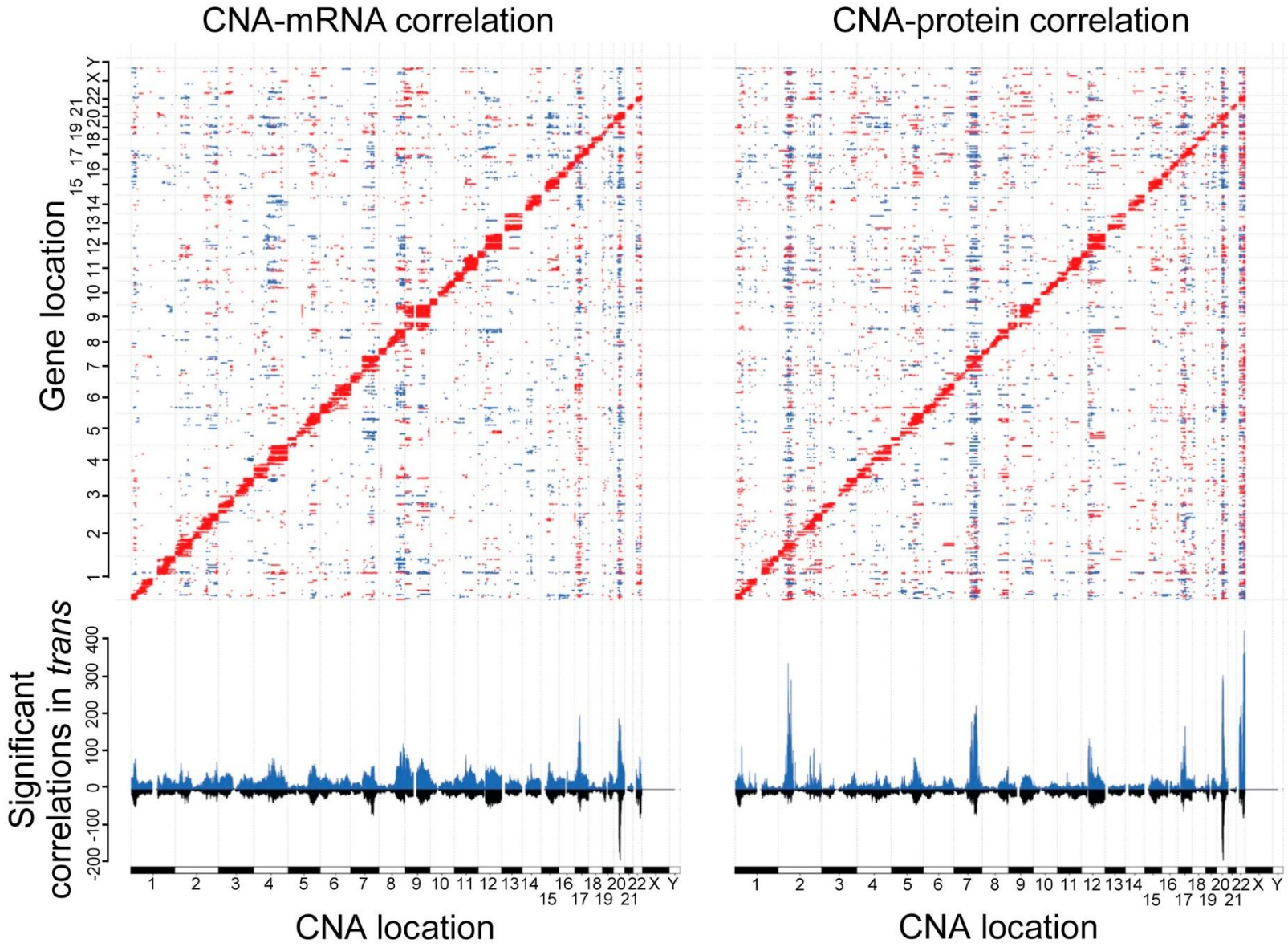
- within 0.47
- across 0.23

- ***Similar correlations have been shown for breast, ovarian, and gastric cancers***

nature
2014 Jul 20; doi: 10.1038 / ***Nature*** 13438

Cis and trans effects in ovarian cancer

- **174 ovarian HGSC tumors**
  - Selection criteria:
    - Overall Survival (OS)
    - Homologous Recombination Deficiency status (HRD)
- **5 proteomic subtypes**
  (4 transcriptomic subtypes)
  - mRNA subtypes translate at protein level
  - New “stromal’ subtype emerged



Zhang et al. *Cell* **2016**, 166, 755-765.

Ovarian Cancer

(pathway activation correlates with overall survival)

NIH NATIONAL CANCER INSTITUTE

Network Data Exchange

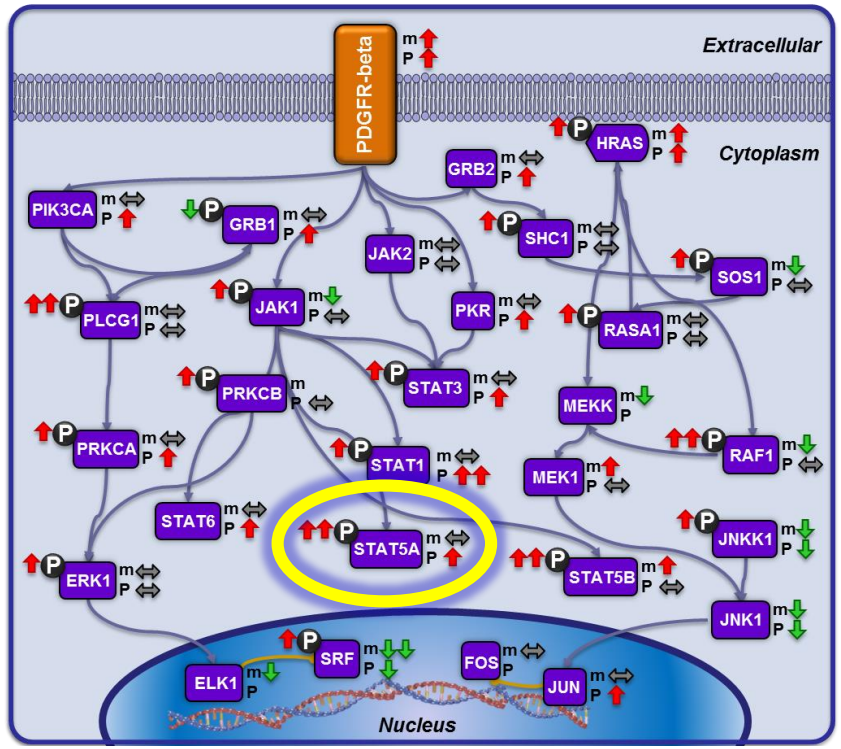
[NCI Pathway Interaction Database]

(214 signaling pathways)

- Significantly upregulated pathways with short OS
  - Protein data (p<0.05)
  - Phosphorylation data (p<0.0001)
  - mRNA data (p<0.05)

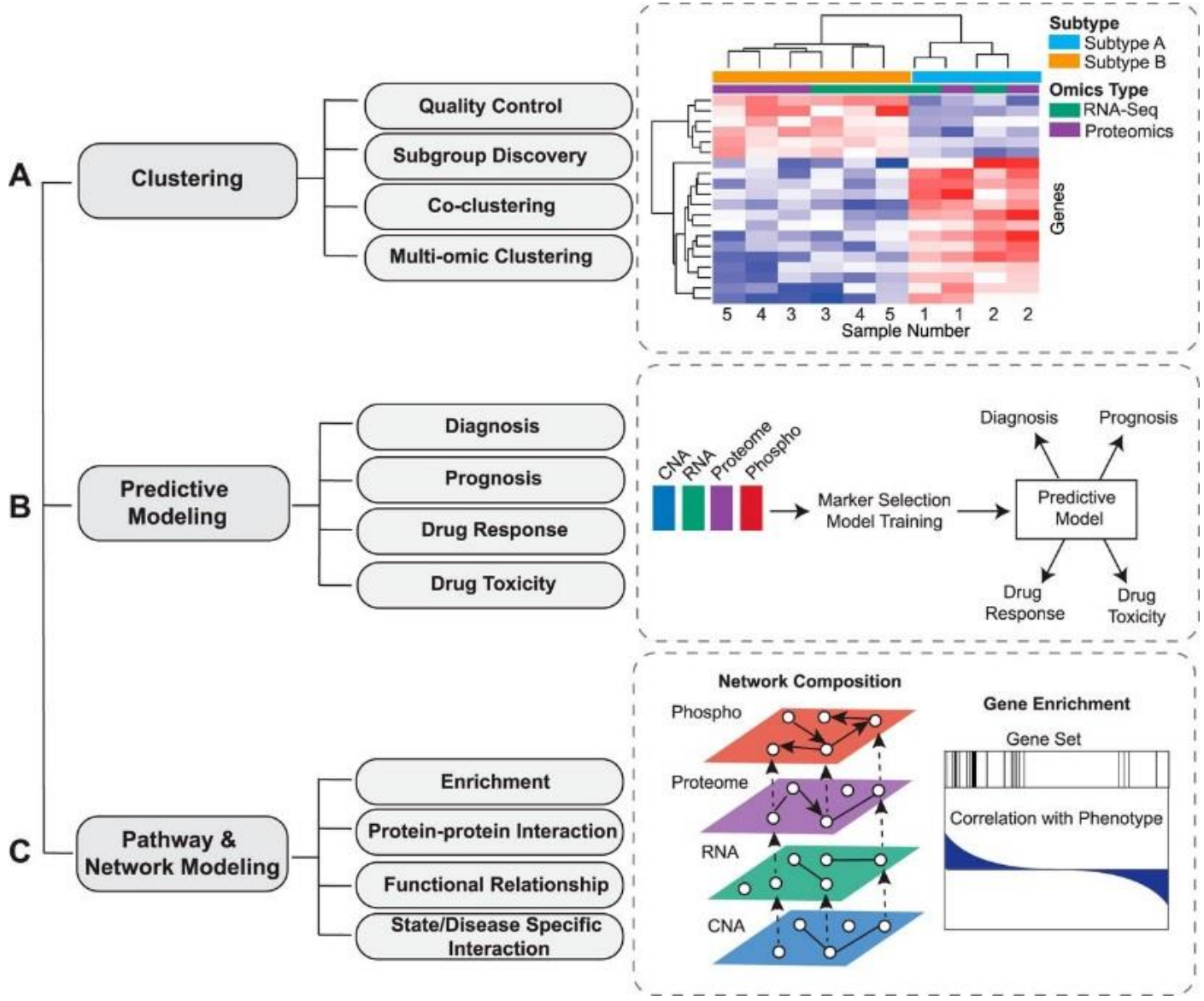
- *Combining comprehensive proteomic, phosphoproteomic and transcriptomic analysis better elucidated the proteogenomic complexity of pathway activation not obtainable at the subtype level.*

*PDGFR pathway upregulation in TCGA **tumors** with short OS*



Cell ***Cell*** 166(3):755-65 (28 Jul 2016)

Integrative modeling of proteogenomic data



Colorectal Cancer: global protein abundance
(new proteome subtypes identified)

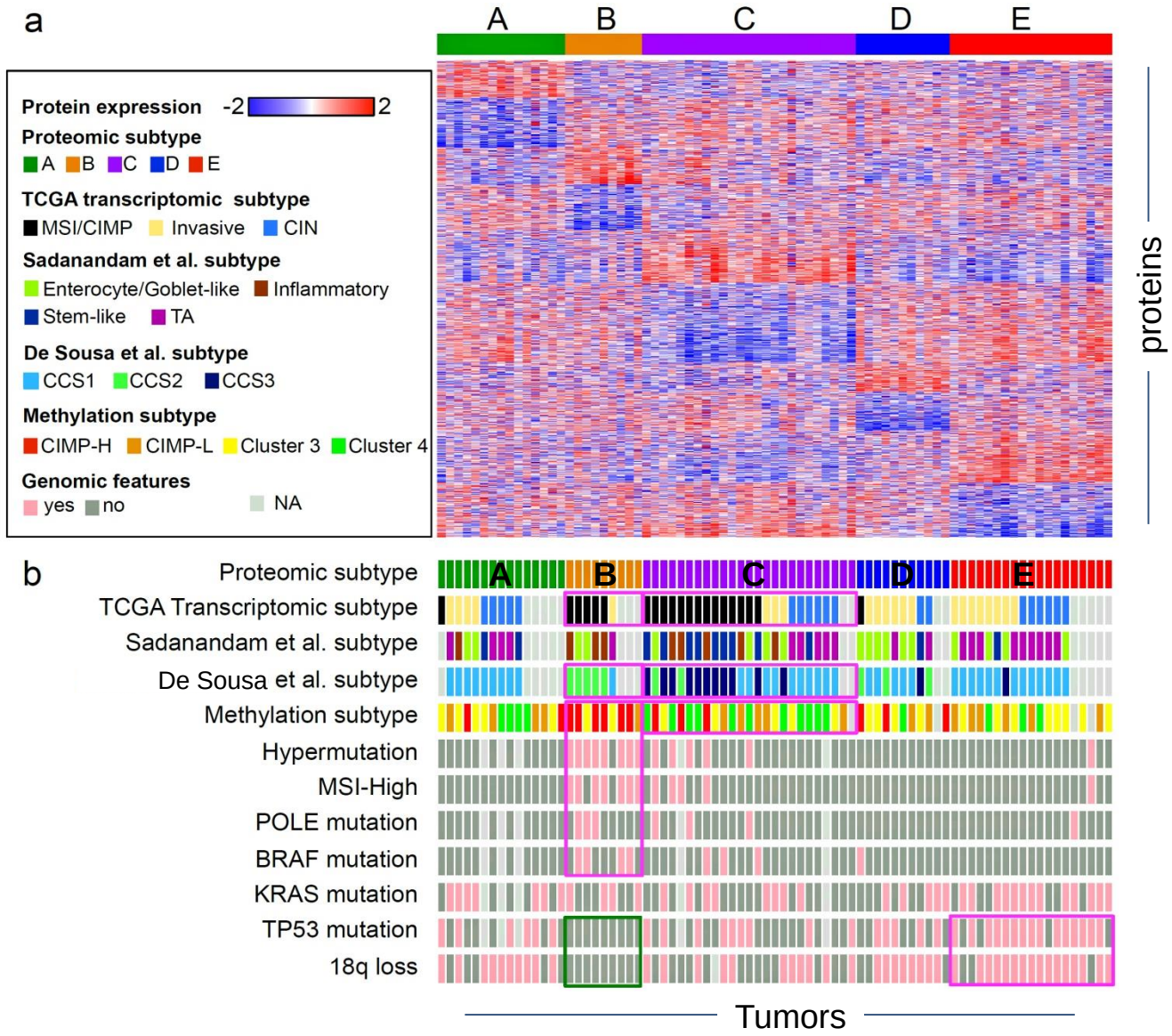
Transcriptome Subtypes

- MSI/CIMP
- Invasive
- CIN

Proteome Subtypes

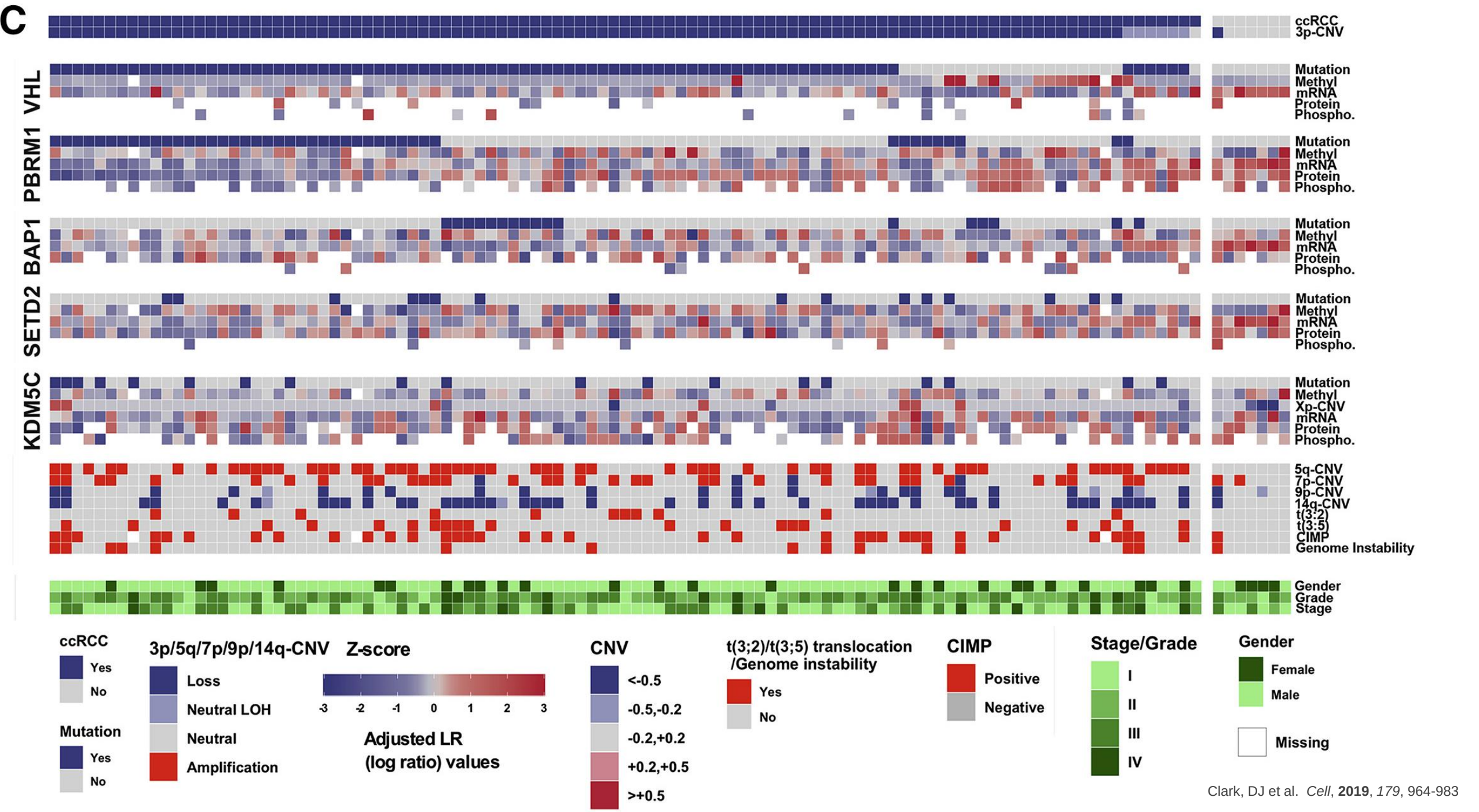
A B C D E

- **MSI/CIMP transcriptome subtype split into two proteome subtypes**
- **Subtype C displayed protein network features characteristic of EMT, associated with rapid metastasis and overall poor survival**



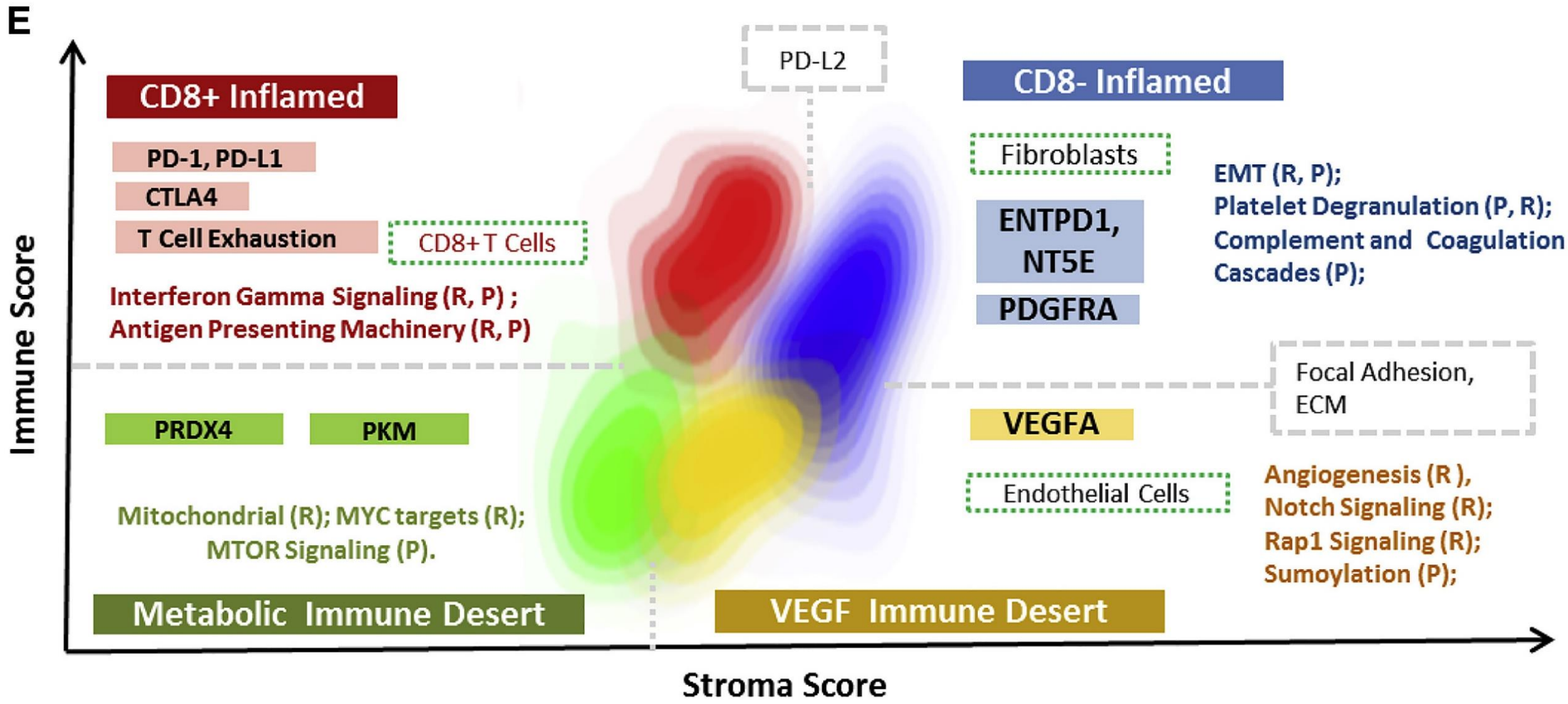
nature

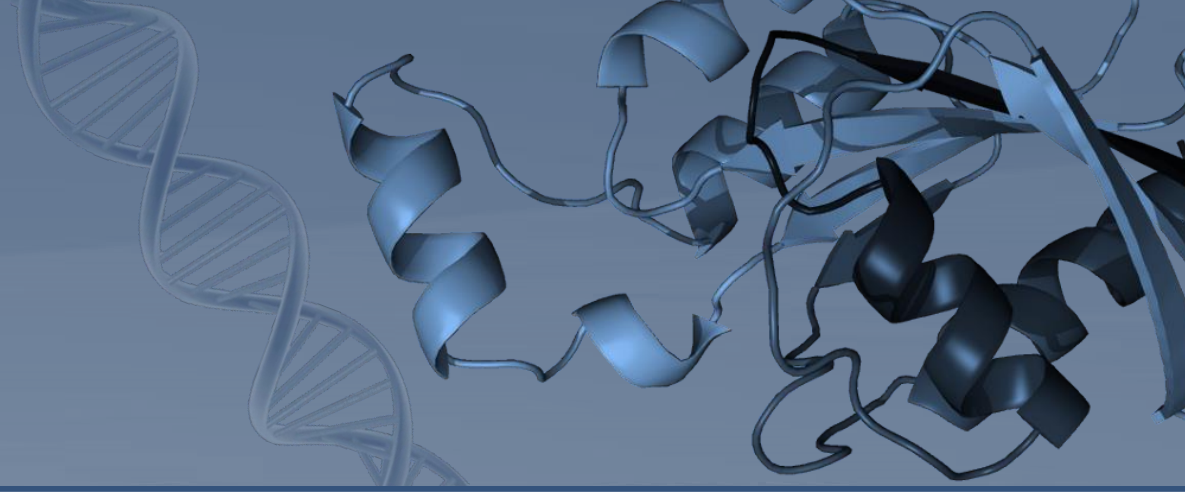
2014 Jul 20; doi: 10.1038 / *Nature* 13438



Clark, DJ et al. *Cell*, **2019**, *179*, 964-983

Immune landscape of kidney cancer proteome

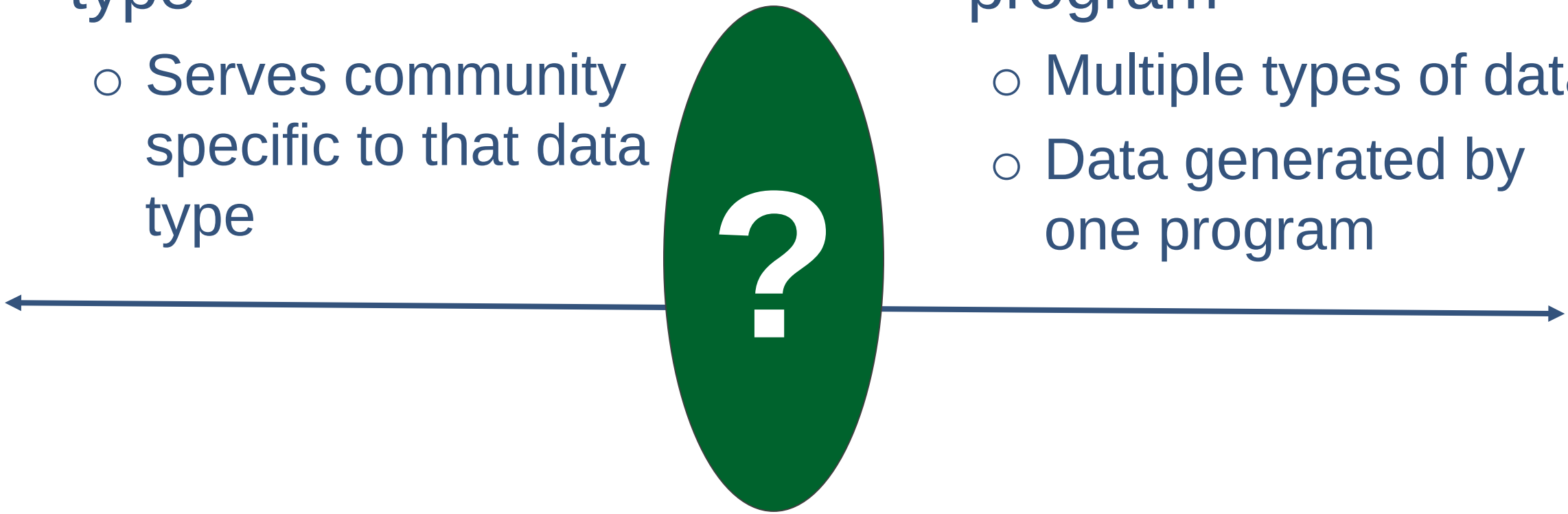




Accessing the data

Two types of data repositories

- Repository by data type
  - Serves community specific to that data type
- Repository by program
  - Multiple types of data
  - Data generated by one program



The Beau Biden Cancer Moonshot℠

Overarching goals

- Accelerate progress in cancer, including prevention & screening
  - From cutting edge basic research to wider uptake of standard of care
- Encourage greater cooperation and collaboration
  - Within and between academia, government, and private sector
- Enhance data sharing

Blue Ribbon Panel – October, 2016

Recommendations include:

- Build a National Cancer Data Ecosystem
  - Enhanced cloud-computing platforms
  - Services that link disparate information, including clinical, image, and molecular data
  - Essential underlying data science infrastructure, standards, methods, and portals for the Cancer Data Ecosystem

NCI Cancer Research Data Commons (CRDC)

Data Sources:

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

Canine Immuno-oncology studies

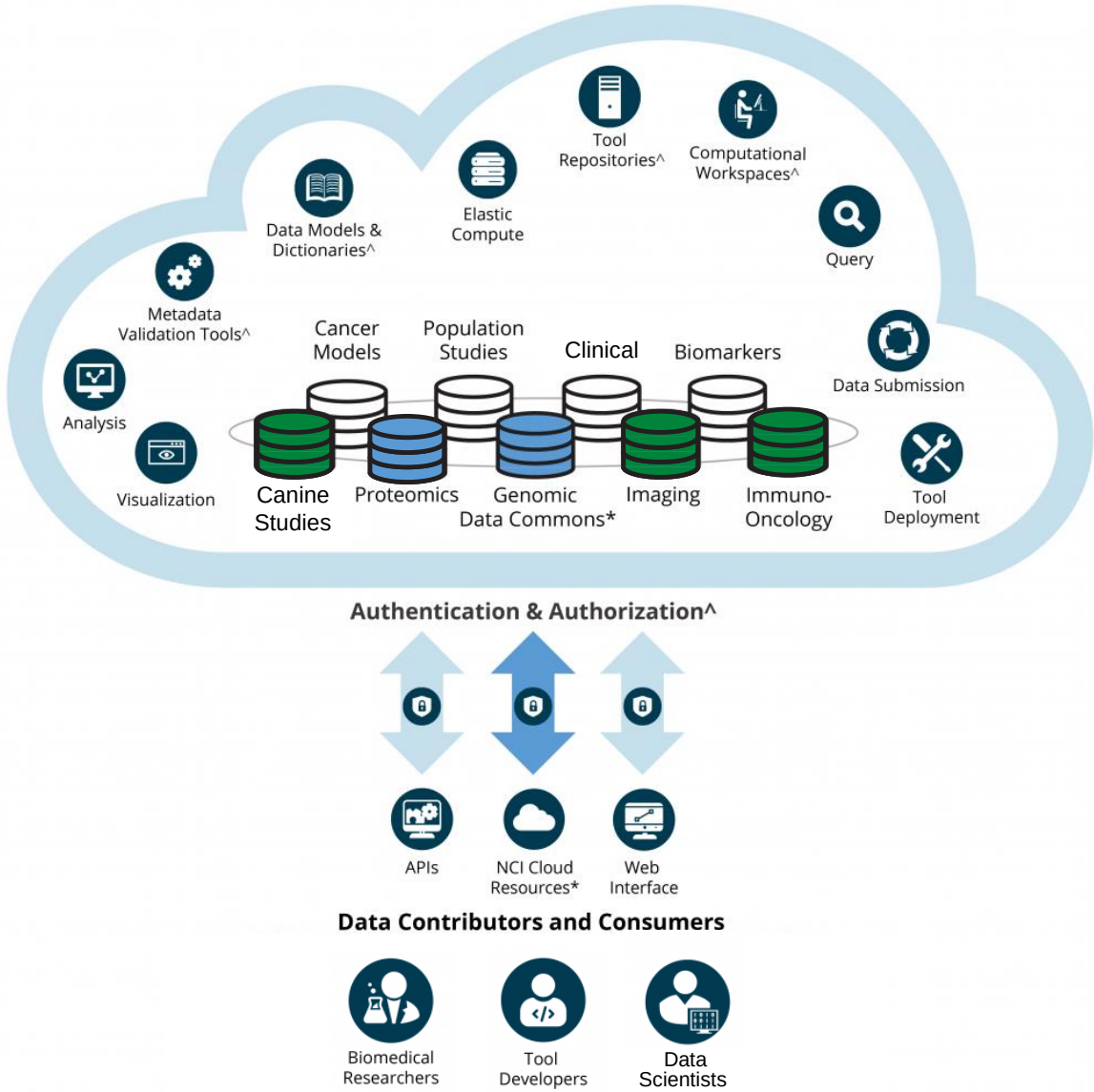
CPTAC
*Clinical Proteomics Tumor Analysis Consortium**

ICPC INTERNATIONAL CANCER PROTEOGENOME CONSORTIUM

TCIA
*The Cancer Imaging Archive**

APOLLO Network
A NCI-DoD-VA Proteogenomic Translational Initiative
T1 T2 T3 T4

BLUE RIBBON PANEL



Components:

- Data Nodes
- Data Commons Framework
- Data Aggregators

- Cloud Resources

- APIs
- Elastic compute resources
- Portals
- Workspaces

- Analytic Tools
- Tool repositories

Data Commons Framework Services

Data Commons Core Framework Services

- Authentication and authorization
- Digital ID and associated metadata services

Extensible, Interoperable Data Models

- Data models
- Data model based APIs

University of Chicago

- https://gen3.org

NCI Cancer Genomics Cloud Pilots

Cloud Pilots provide:

- Access to large omic data sets without need to download
- Access to popular pipelines and visualization tools
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses

Data

- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Compute

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

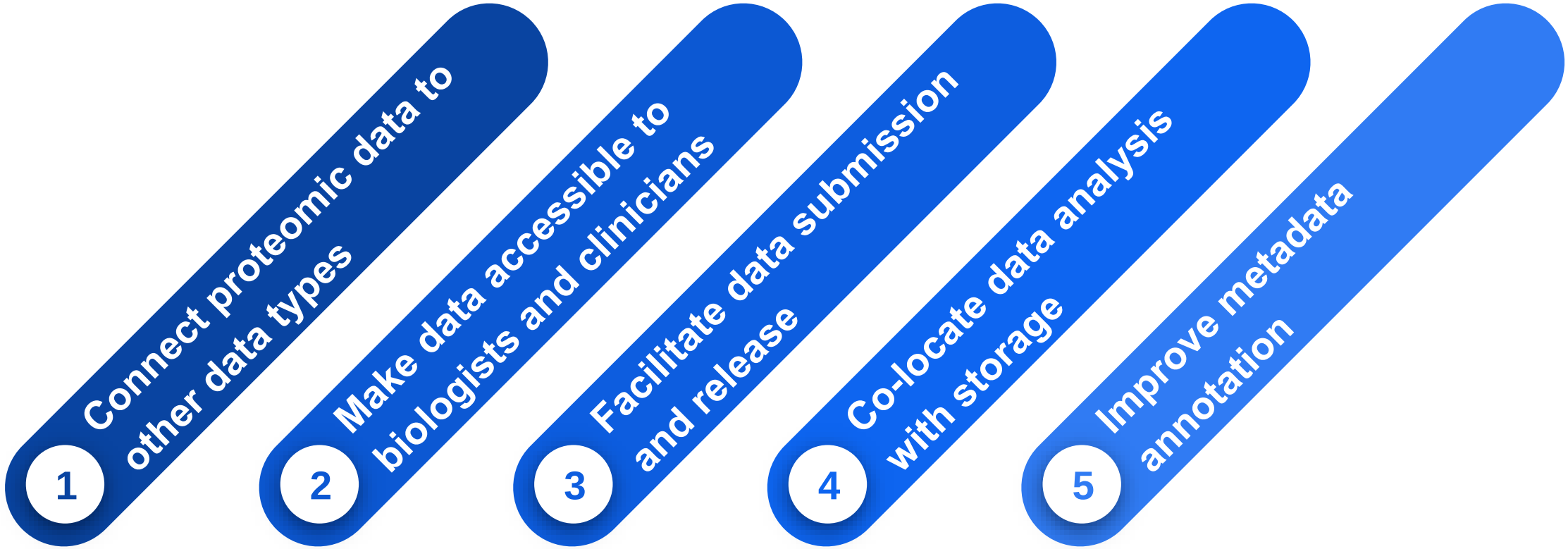
Security

- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

#NCICloud

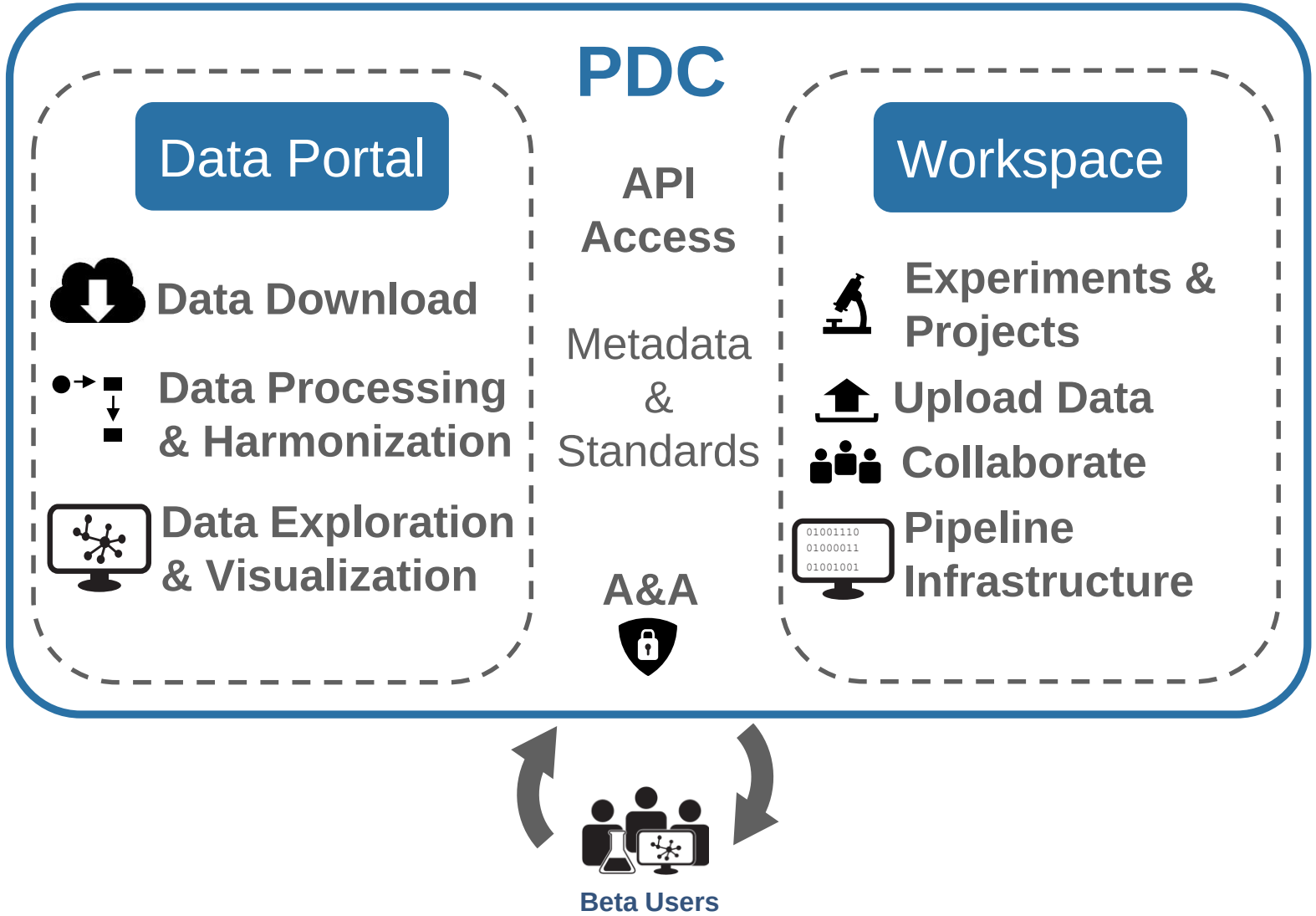
Proteomic Data Commons: High Level Goals



Proteome Data Commons – democratize access to cancer-related proteomic datasets

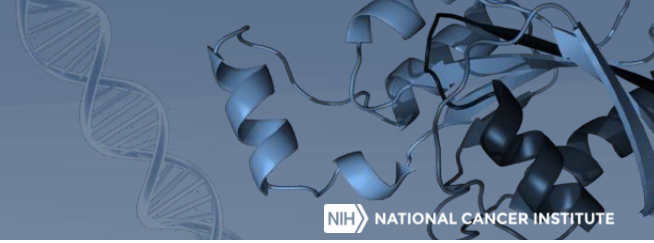
NIH NATIONAL CANCER INSTITUTE

https://pdc.esacinc.com/pdc/pdc



Courtesy of Izumi HInkson

PDC fast facts



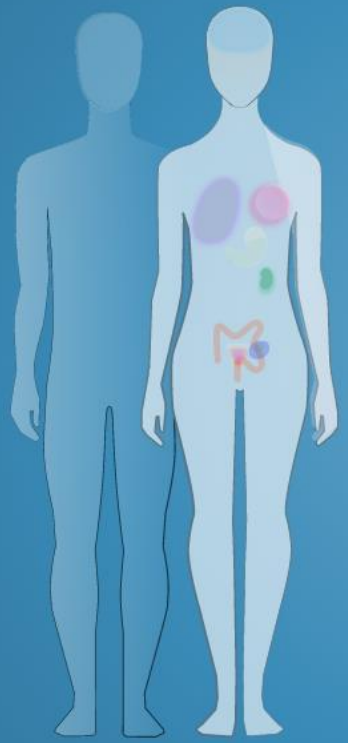
- Mass spectra
- TMT, label-free, and DIA data
- Data from CPTAC, CBTTC
- In sync with GDC, KidsFirst, Seven Bridges, and TCIA

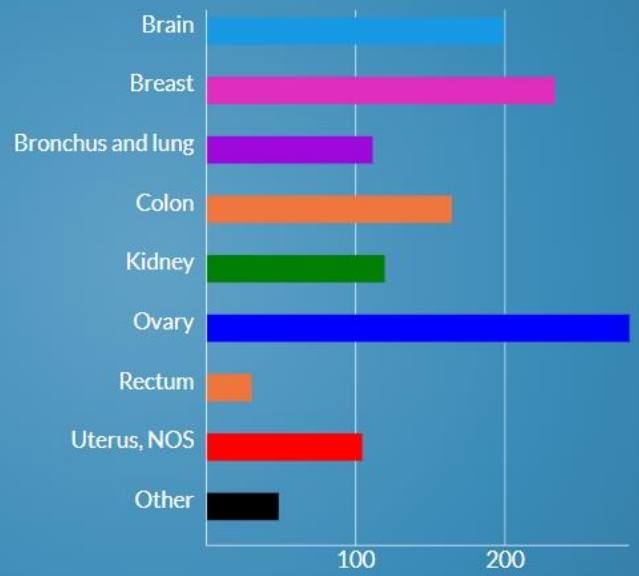
PDC - https://pdc.esacinc.com/pdc/pdc

NIH NATIONAL CANCER INSTITUTE

3	9 TB	42,959	64,720,831	991,578	13,713
Programs	Data volume	Data files	Spectra	Peptides	Proteins

PRIMARY SITES





DISEASE TYPES

Breast Invasive Carcinoma	233
Chromophobe Renal Cell Carcinoma	1
Clear Cell Renal Cell Carcinoma	116
Colon Adenocarcinoma	164
Lung Adenocarcinoma	111
Other	48
Ovarian Serous Cystadenocarcinoma	283
Papillary Renal Cell Carcinoma	2
Pediatric/AYA Brain Tumors	219
Rectum Adenocarcinoma	30
Uterine Corpus Endometrial Carcinoma	104

RECENT RELEASES

- Pediatric Brain Tumor Atlas - CBTTC program Pediatric/AYA brain tumor dataset released, November 2019
- CPTAC3 program Lung Adenocarcinoma (LUAD) dataset released, September 2019
- CRDC cloud resources piloted integration with Proteomic data commons for multiomic data analysis, August 2019
- MVP 0.5 Removed the requirement to register and login to the PDC Data Browser, June 2019

NEWS

- 18th Human Proteome Organization World Congress, September 15-19, 2019, Adelaide, Australia
- 68th ASMS Conference on Mass Spec and Allied Topic June 2-6, 2019 Georgia World Congress Center, Atlanta, GA
- CPTAC Releases UCEC, ccRCC Discovery Data and Other Study Datasets

Feedback

GENERAL
BIOSPECIMEN
CLINICAL
FILES
GENES

FILTERS

Ethnicity

- [ ] Hispanic or Latino
- [ ] Not Hispanic or Latino
- [ ] Not Reported
- [ ] Unknown

Race

- [ ] American Indian or Alaska Native
- [ ] Asian
- [ ] Black or African American
- [ ] Native Hawaiian or Other Pacific Islander
- [ ] Not Reported
- [ ] Unknown
- [ ] White

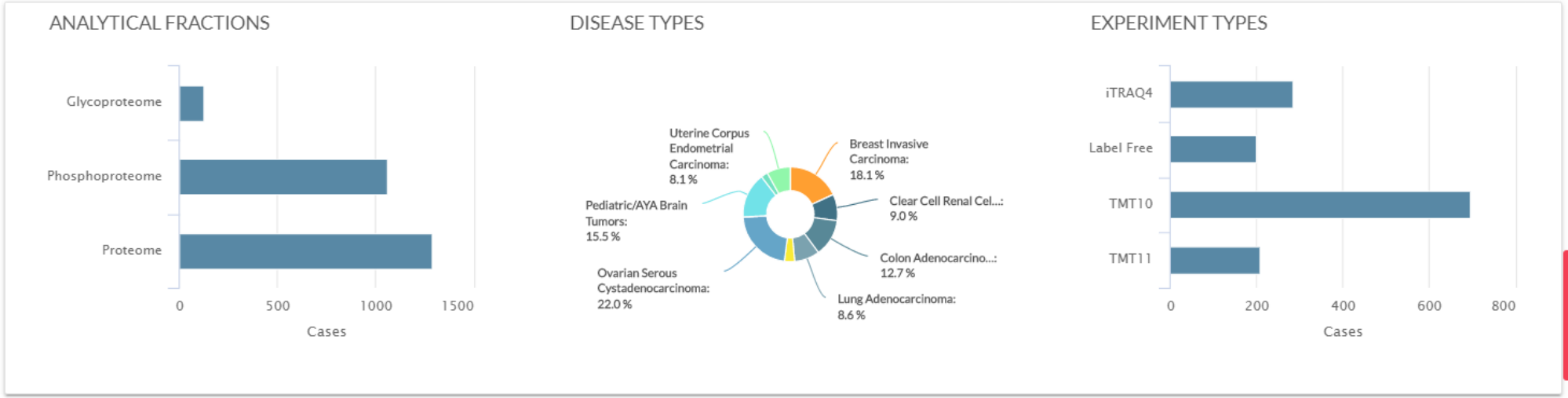
Gender

- [ ] Female
- [ ] Male
- [ ] Not Reported

Tumor Grade

« Start searching by selecting a facet

Charts



Studies (24) | Biospecimens (1887) | Clinical (1311) | Files (42959) | Genes (13917)

Total records: 1311

Download All Manifests

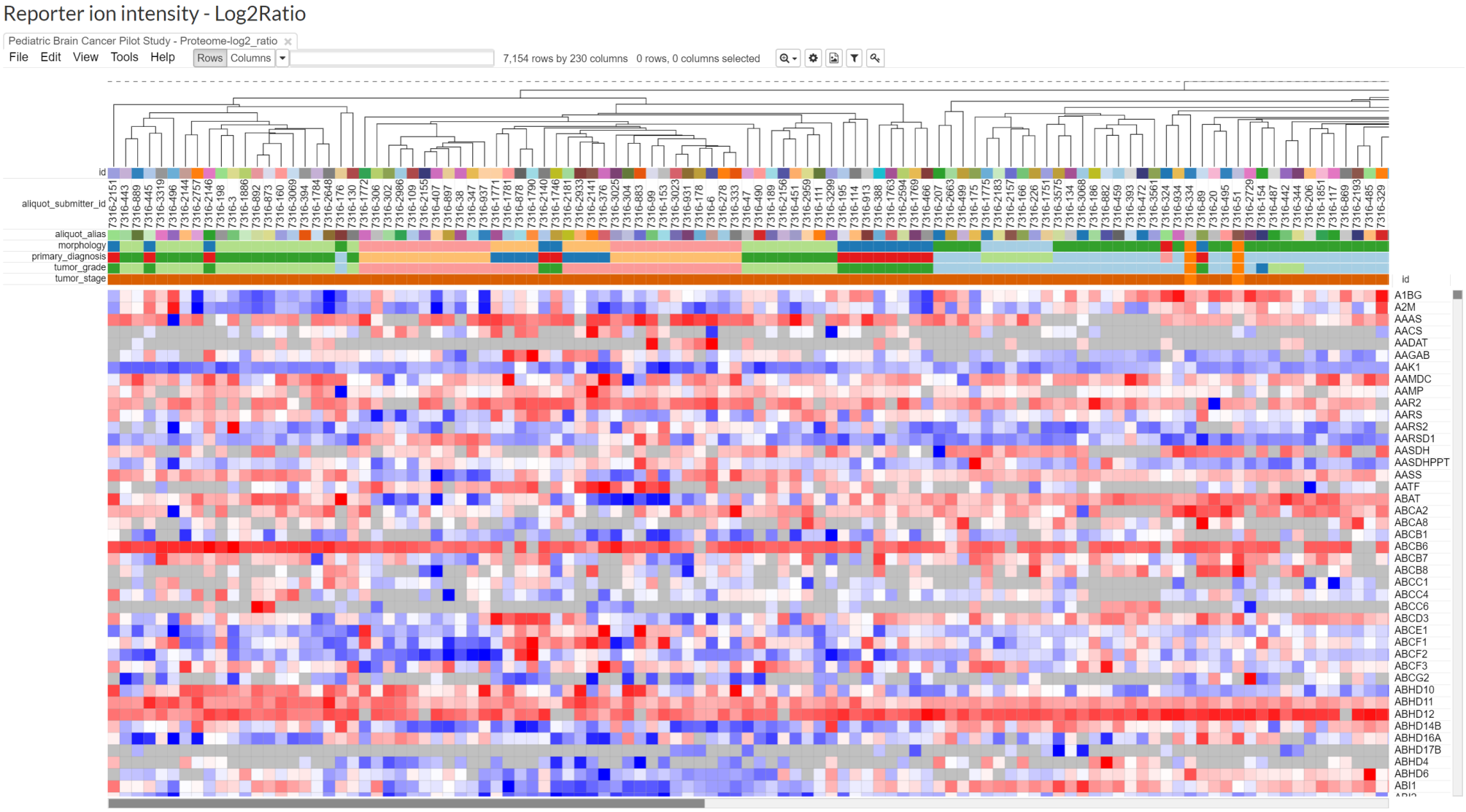
Download Clinical Manifest

1 2 3 4 5 10

Cases Submitter ID	External Data Resource	Ethnicity	Gender	Race	Morphology	Primary Diagnosis	Site of Resection or Biopsy	Tissue or Organ of Origin	Tumor Grade	Tumor Stage
C16974	Kids First	Not Hispanic or Latino	Male	White	9505/1	Ganglioglioma, NOS	Temporal Lobe	Temporal Lobe	G1	Unknown
C94956	Kids First	Not Hispanic or Latino	Male	White	9380/3	Glioma, malignant	Temporal Lobe	Temporal Lobe	High Grade	Unknown
C94956	Kids First	Not Hispanic or Latino	Male	White	9380/3	Glioma, malignant	Temporal Lobe	Temporal Lobe	High Grade	Unknown

Feedback

Customizable heat maps of peptides



Visualizations based on Morpheus, https://software.broadinstitute.org/morpheus

User flow to compute PDC data on Cloud Resources

NATIONAL CANCER INSTITUTE

NATIONAL CANCER INSTITUTE
Proteomic Data Commons

1. User starts on PDC portal to identify cohort of files
2. User downloads **files manifest** of selected cohort
   a. Note - there are multiple manifest options on the PDC, only the "files manifest" will work

https://pdc.esacinc.com/pdc/

CANCER GENOMICS CLOUD

1. User moves to CGC, creates a project
   a. Files → Add files → Import from PDC
2. User prompted to upload the manifest from the PDC
3. PDC files copied into user's project
   a. FileID → file URL via Fence

http://www.cancergenomicscloud.org/

Source: Manisha Ray, SBG CGC

PDC: PepQuery

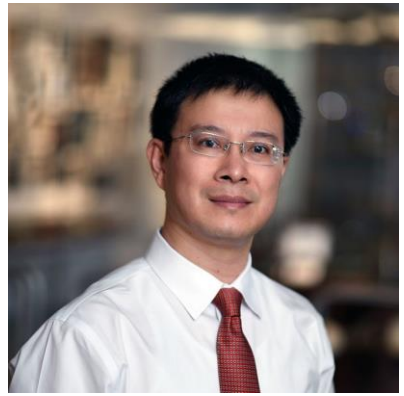
- Search data for specific variant peptide

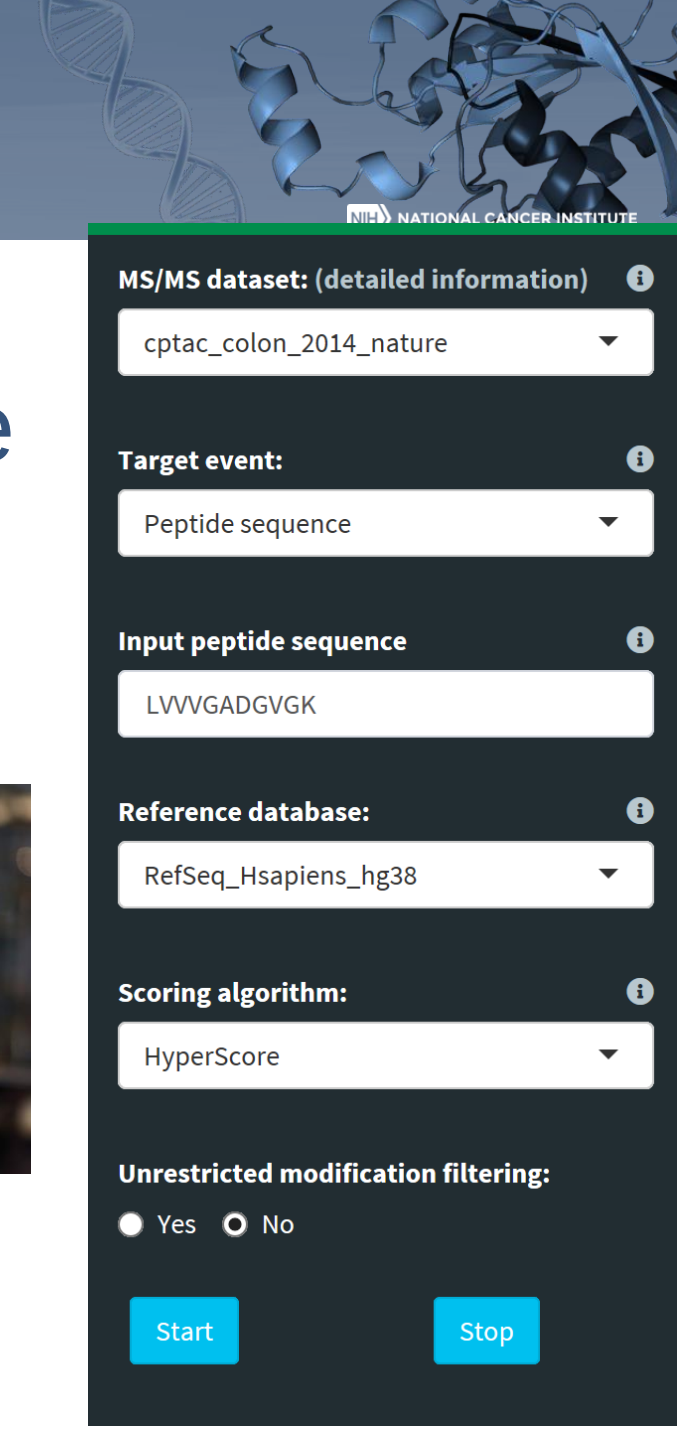
Resource

PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations

Bo Wen,[1,2] Xiaojing Wang,[1,2] and Bing Zhang[1,2]

[1]*Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [2]*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA*





Gen. Res. 2019, 29:485–493

PDC: PepQuery

Identification overview

MS/MS searching parameters:

	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

Identification result

The identification result is presented in the below table:

Show 10 entries

Search:

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

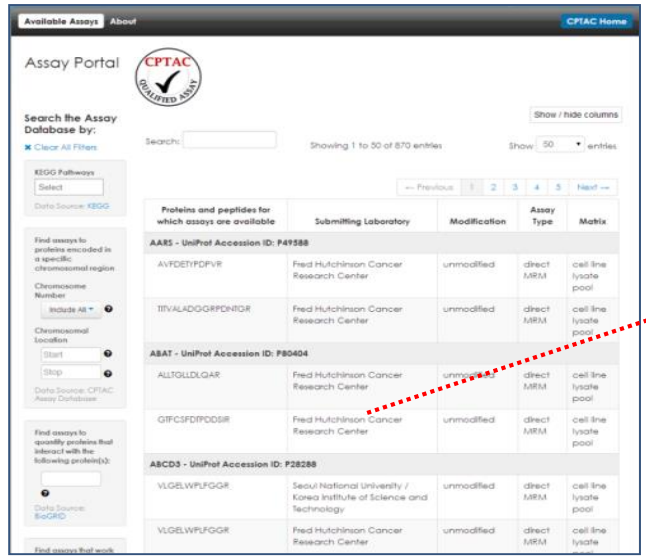
Antibodies and targeted assays

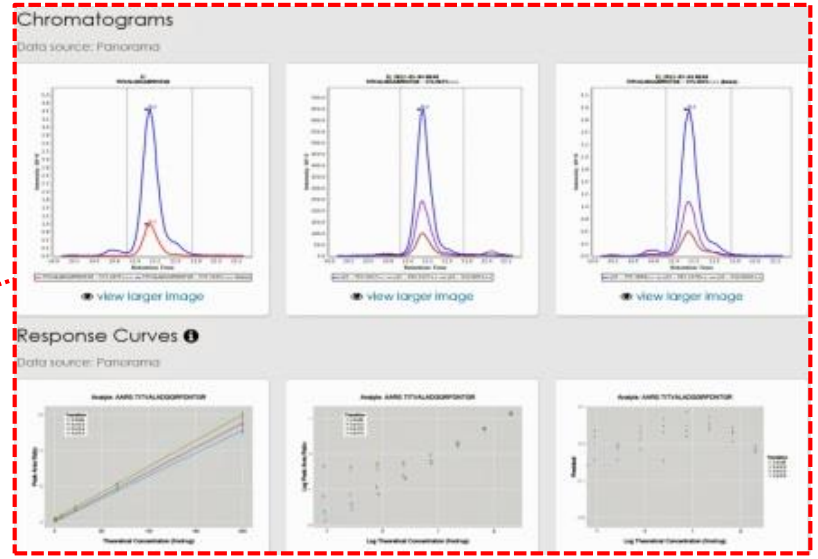
NIH NATIONAL CANCER INSTITUTE

CPTAC Assay Portal

2375 “fit-for-purpose” targeted assays
(6,584 users/month)

assays.cancer.gov

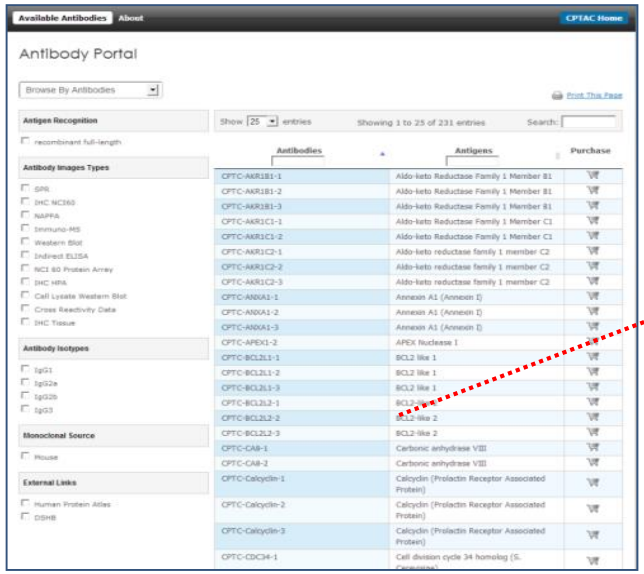




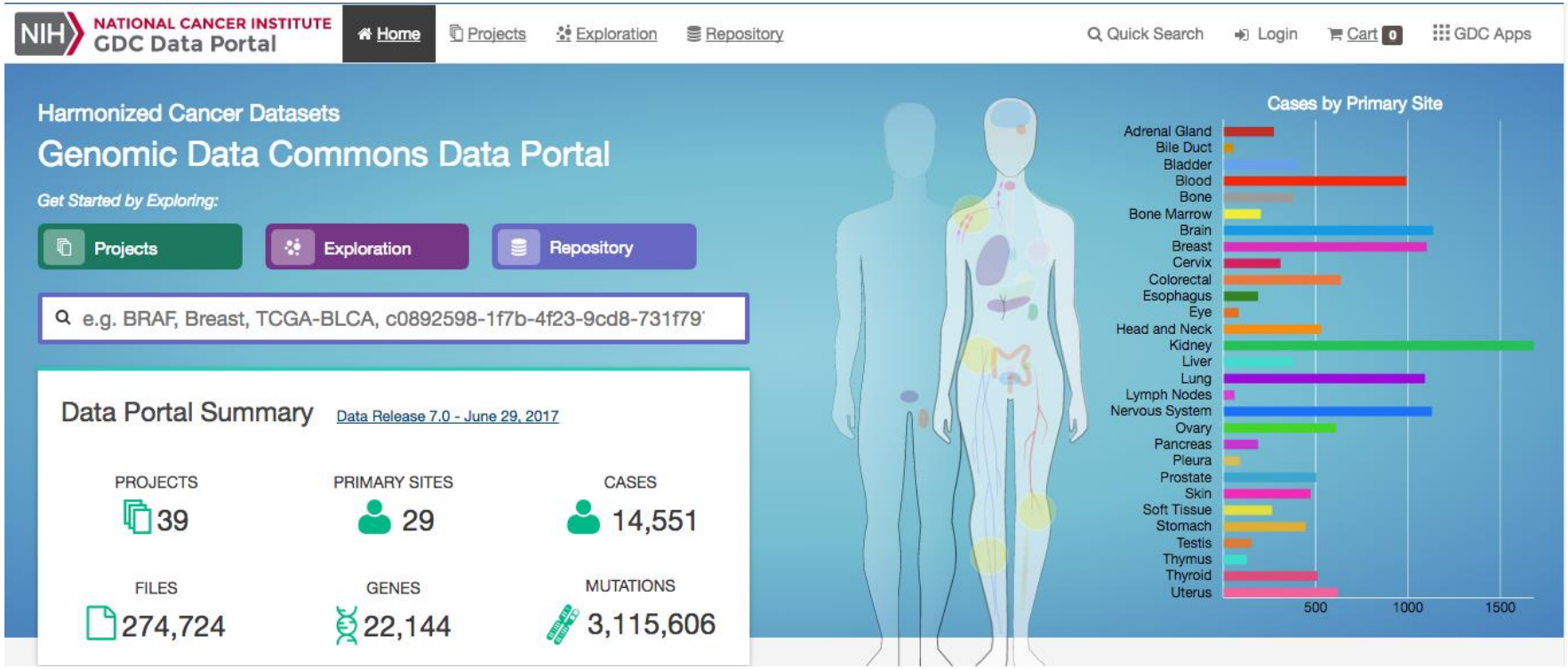
NCI Antibody Portal

583 mAbs available
(2,653 units distributed)

antibodies.cancer.gov



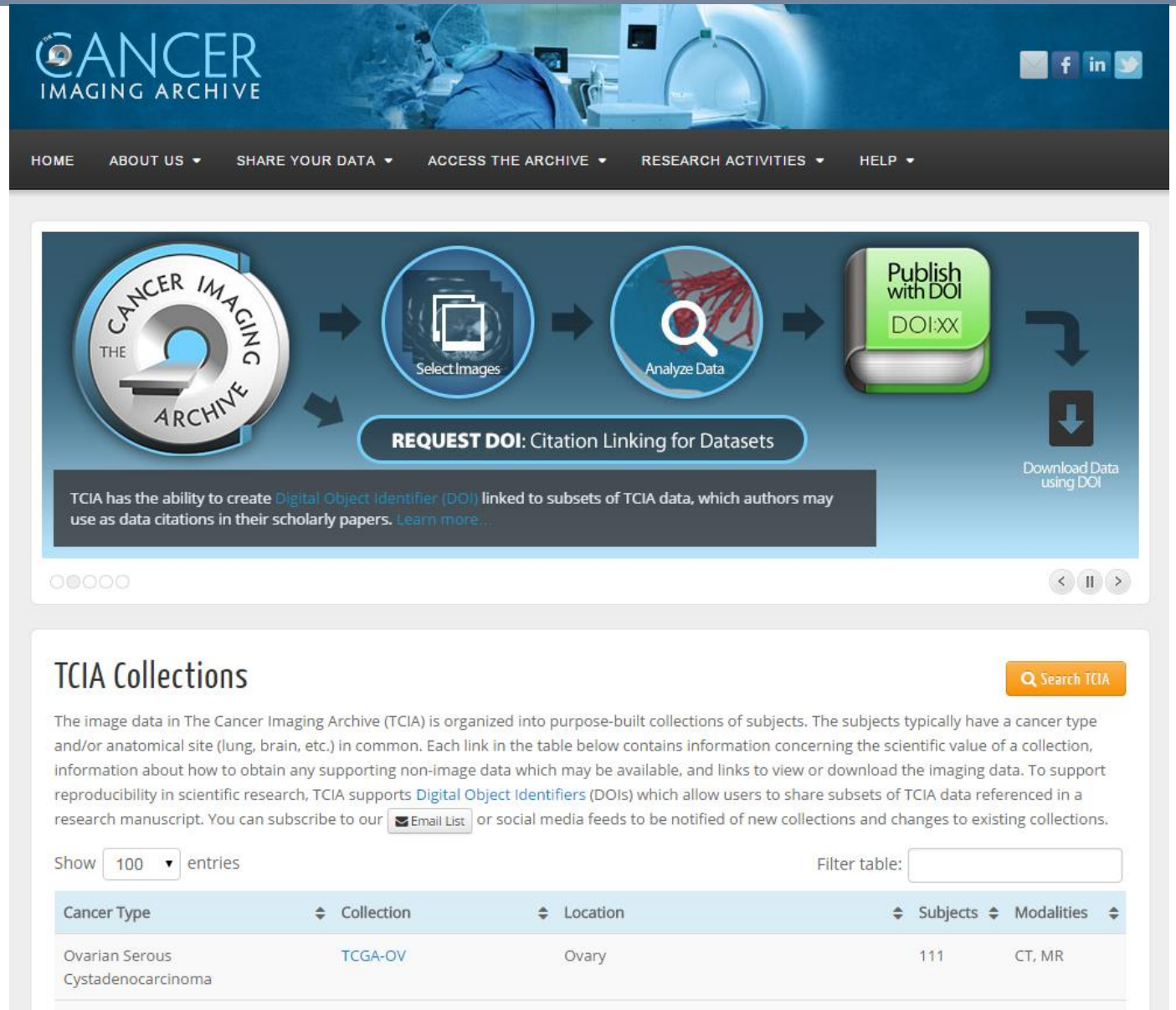
NCI Genomic Data Commons (GDC)



- Launched in 2016 with over 4 PB of data.
- Joint project with OICR.
- Used by 1000 -2000+ users per day.
- Based upon an open source software stack that can be used to build other data commons.

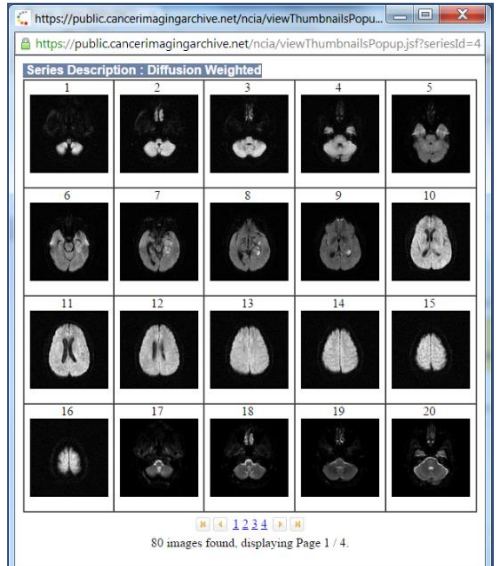
*See: NCI Genomic Data Commons: Grossman, Robert L., et al. "Toward a shared vision for cancer genomic data." New England Journal of Medicine 375.12 (2016): 1109-1112.

The Cancer Imaging Atlas (TCIA)



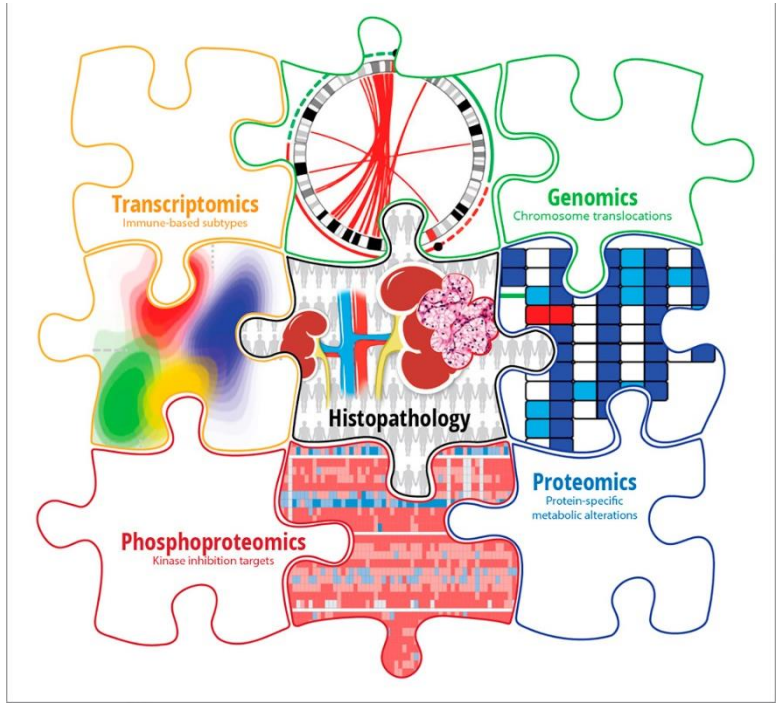
http://cancerimagingarchive.net

- ~40,000 total subjects in the archive
- ~32 M images available
  - Quantitative Imaging Network
  - TCGA - CPTAC
  - ECOG-ACRIN and RTOG



Summary

- **Invitation**
  - Submit data to **https://pdc.esacinc.com**
  - Provide feedback **nci.pdc.help@esacinc.com**



Thank You

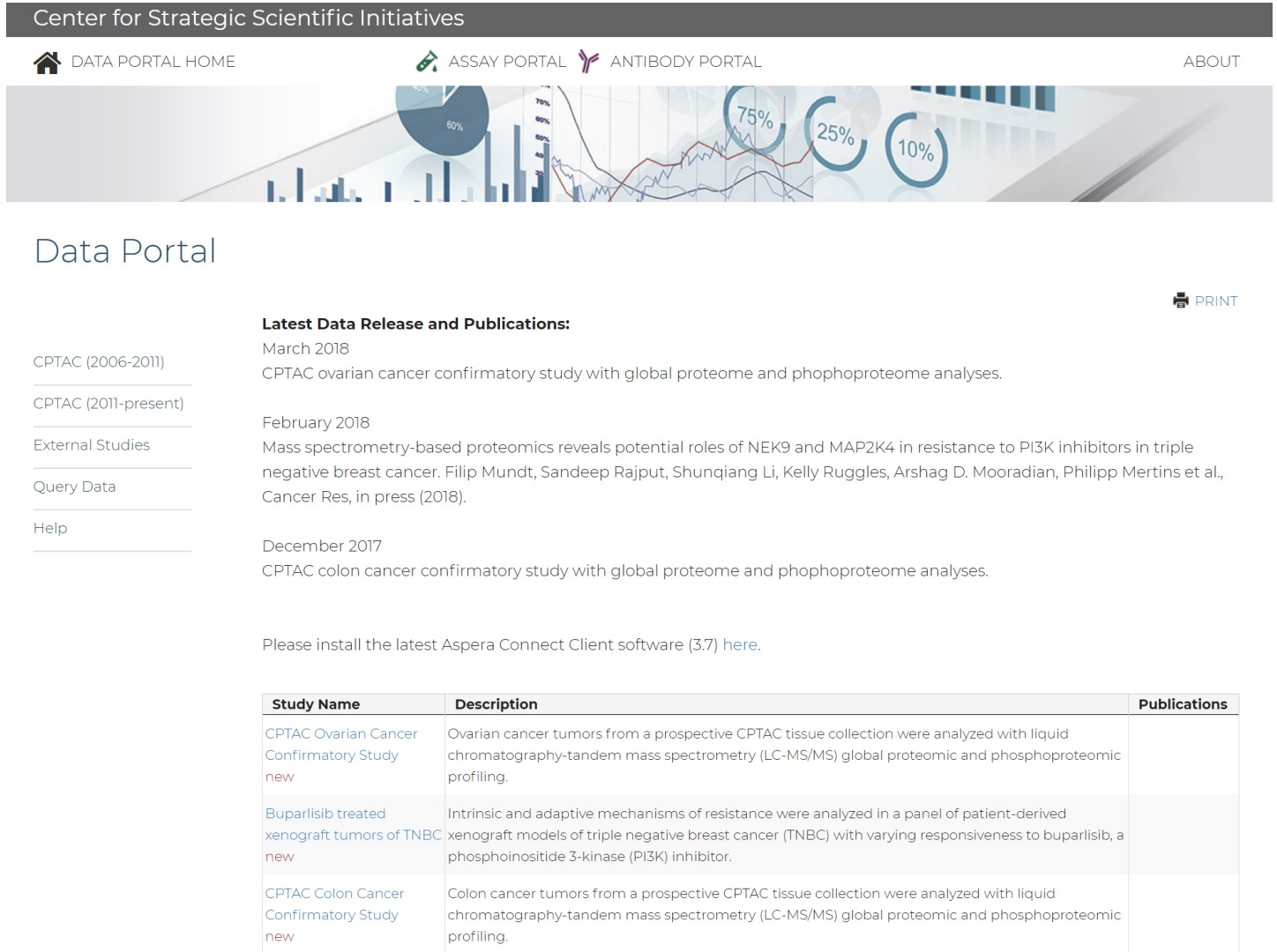
Backup slides

Where to get the data

- CPTAC data portal

- https://cptac-data-portal.georgetown.edu/cptacPublic/

- https://proteomics.cancer.gov and select “Data Portal”



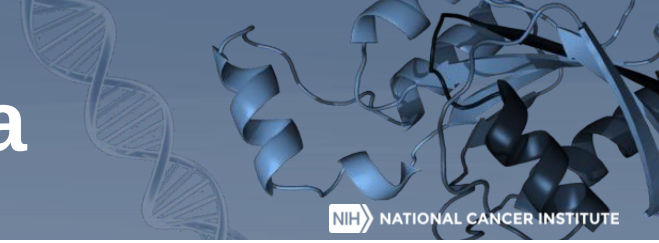
Available datasets

Tumor type	TCGA residual		CPTAC confirmatory	
	Tumor	Normal	Tumor	Normal
Colorectal	90	0	104	96
Breast	105	0	100	
Ovarian	174	0	100	25

Genomic and imaging data also available for these samples at Genome Data Commons and The Cancer Imaging Archive

https://gdc.cancer.gov/ http://www.cancerimagingarchive.net/

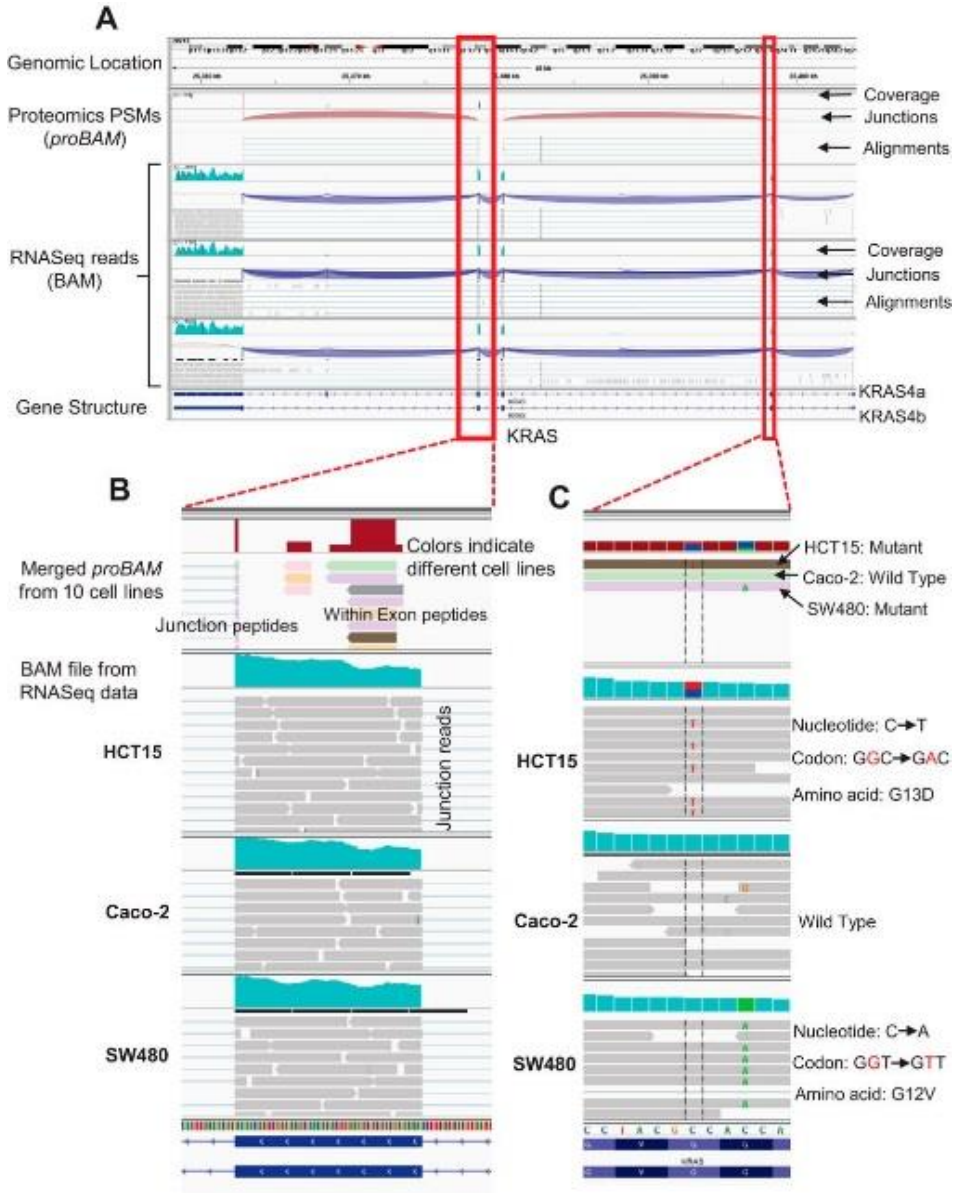
2 approaches to manage proteogenomic data



- Current approach
  - Local server storage
  - Users download data
- Future approach
  - Cloud storage
  - Users bring tools to cloud for data access and computation

t

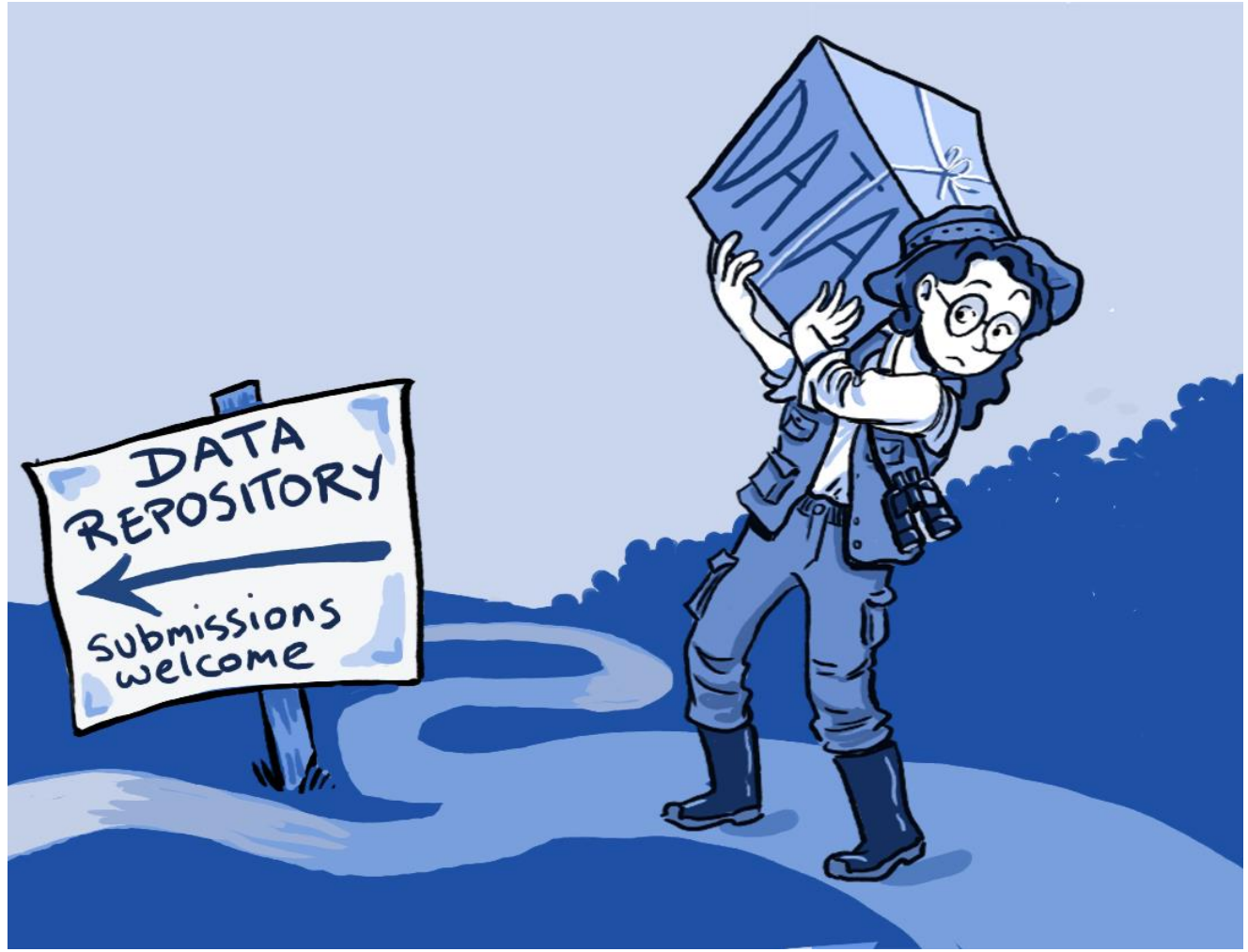
Data sharing and visualization



Data sharing and visualization

What problem will the PDC solve?

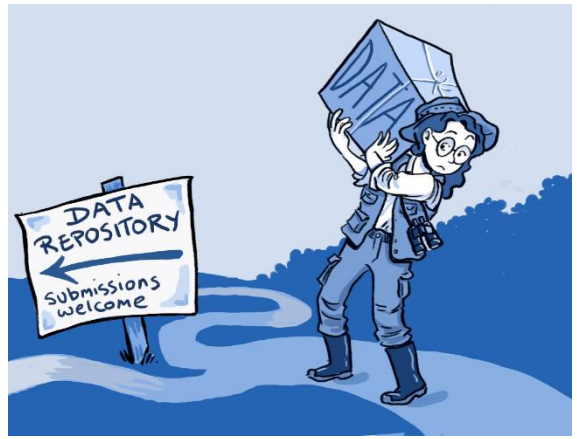
- Simplify access to cancer research data



What problem will the PDC solve?

- Simplify access to cancer research data

Proteomic data



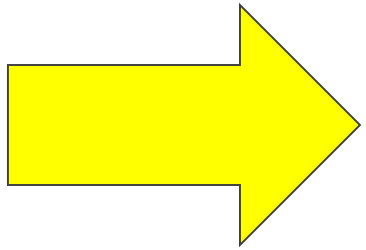
Genomic data

Imaging data

Clinical data

Big Data V's applied to Proteogenomics data

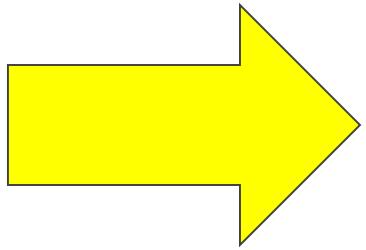
- Big Data V's
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Valence



- Proteogenomics data
  - P – TB; G - PB
  - Highly multidimensional
  - Low in research space
  - Analytical, clinical quality
  - Gene-protein, protein-protein, subject-sample

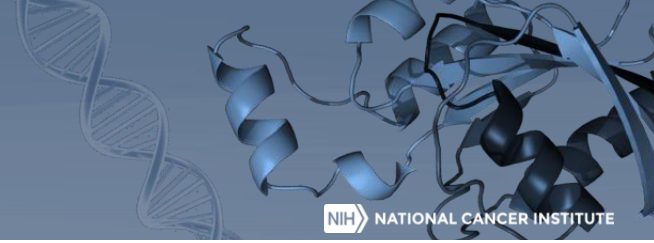
Big Data V's applied to Proteogenomics data

- Big Data V's
  - Volume
  - **Variety**
  - Velocity
  - **Veracity**
  - Valence



- Proteogenomics data
  - P – TB; G - PB
  - **Highly multidimensional**
  - Low in research space
  - **Analytical, clinical quality**
  - Gene-protein, protein-protein, subject-sample

FAIR principles of data sharing



- Findable
- Accessible
- Interoperable
- Reusable

Data Biosphere == Data Ecosystem

- Modular – composed of functional components with well-specified interfaces
- Community-driven – created by many groups to foster a diversity of ideas
- Open – developed under open-source licenses that enable extensibility and reuse
- Standards-based – consistent with standards developed by coalitions such as GA4GH

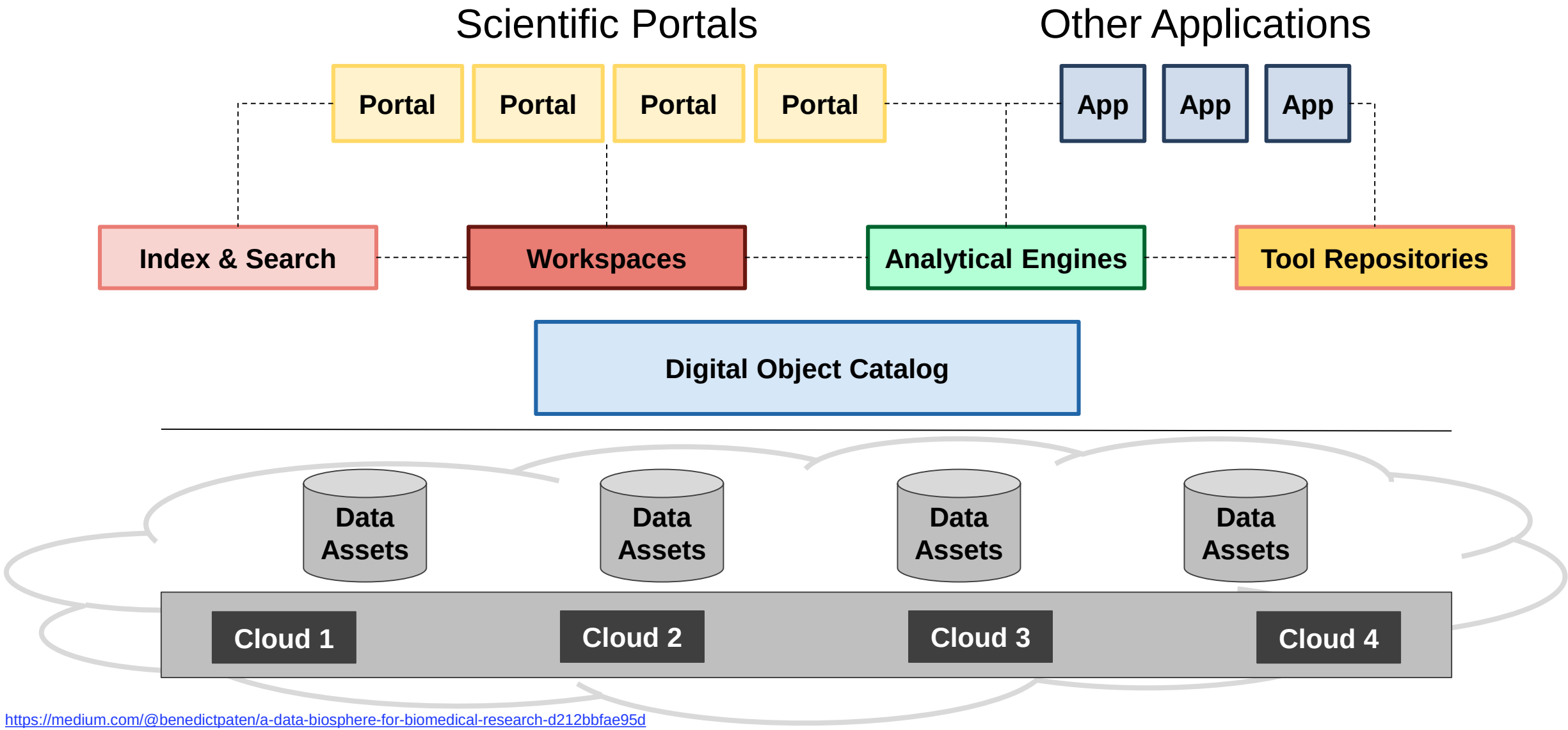
https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d

Potential functions of a data commons

- Data repository for submission, processing, and downloading
- System that applies a common set of bioinformatics pipelines to submitted data
- System that reanalyzes submitted data when new bioinformatics pipelines are developed
- System that allows users to build their own paplications and systems that interoperate the commons through the API

Wilson et al. *Cancer Res.* **2017**, *77*, e15-e18.

Data environment of modular components



https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d

What is a data commons?

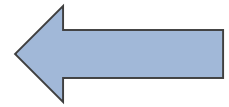
- A data commons **colocates** data, storage, and computing infrastructure with commonly used services, tools, and apps for analyzing and sharing data to create an **interoperable** resource for the research community

How is it different from a repository?

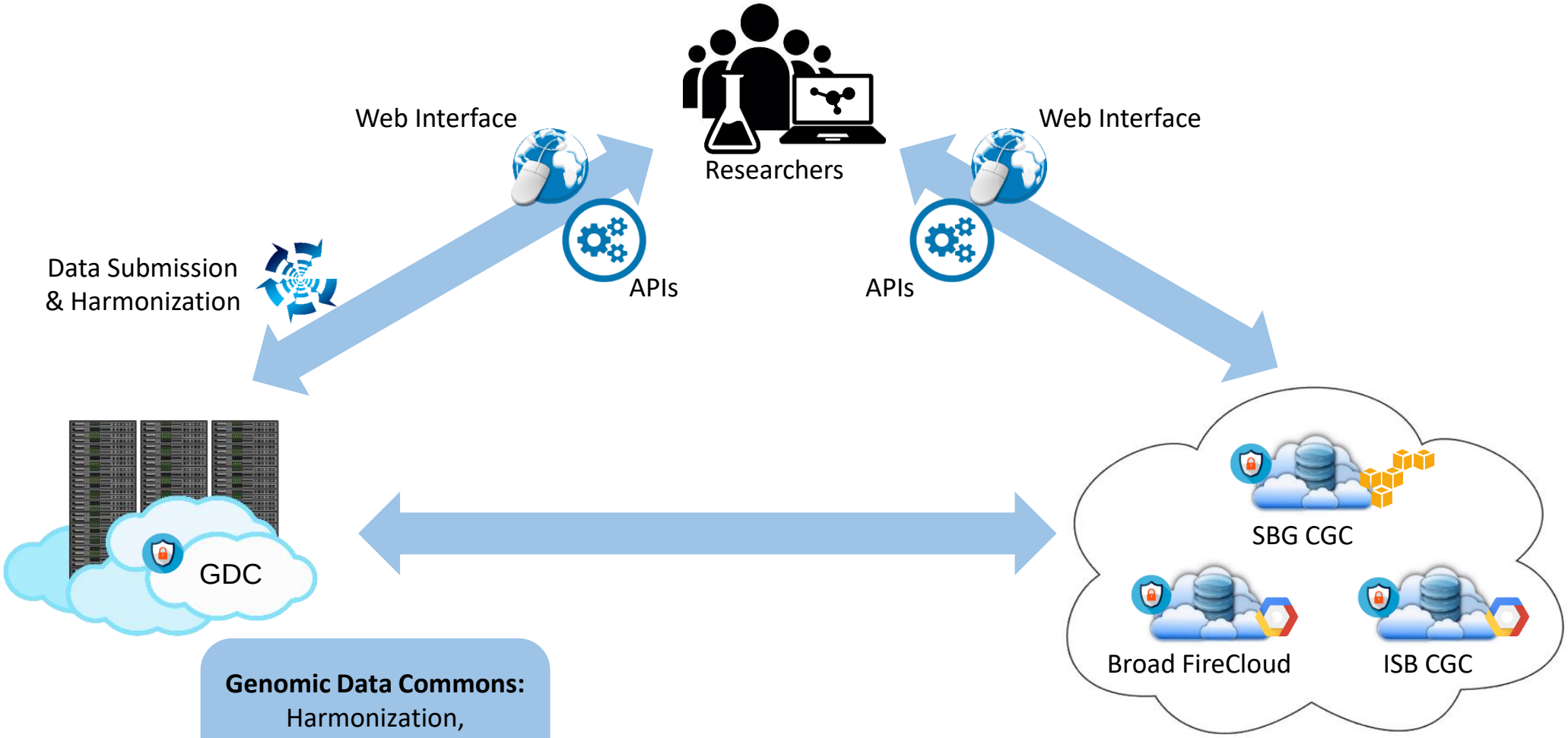
- Data repository

Data  Users

- Data commons

Data & Compute  Users

GDC and NCI Cloud Pilots/Framework Today

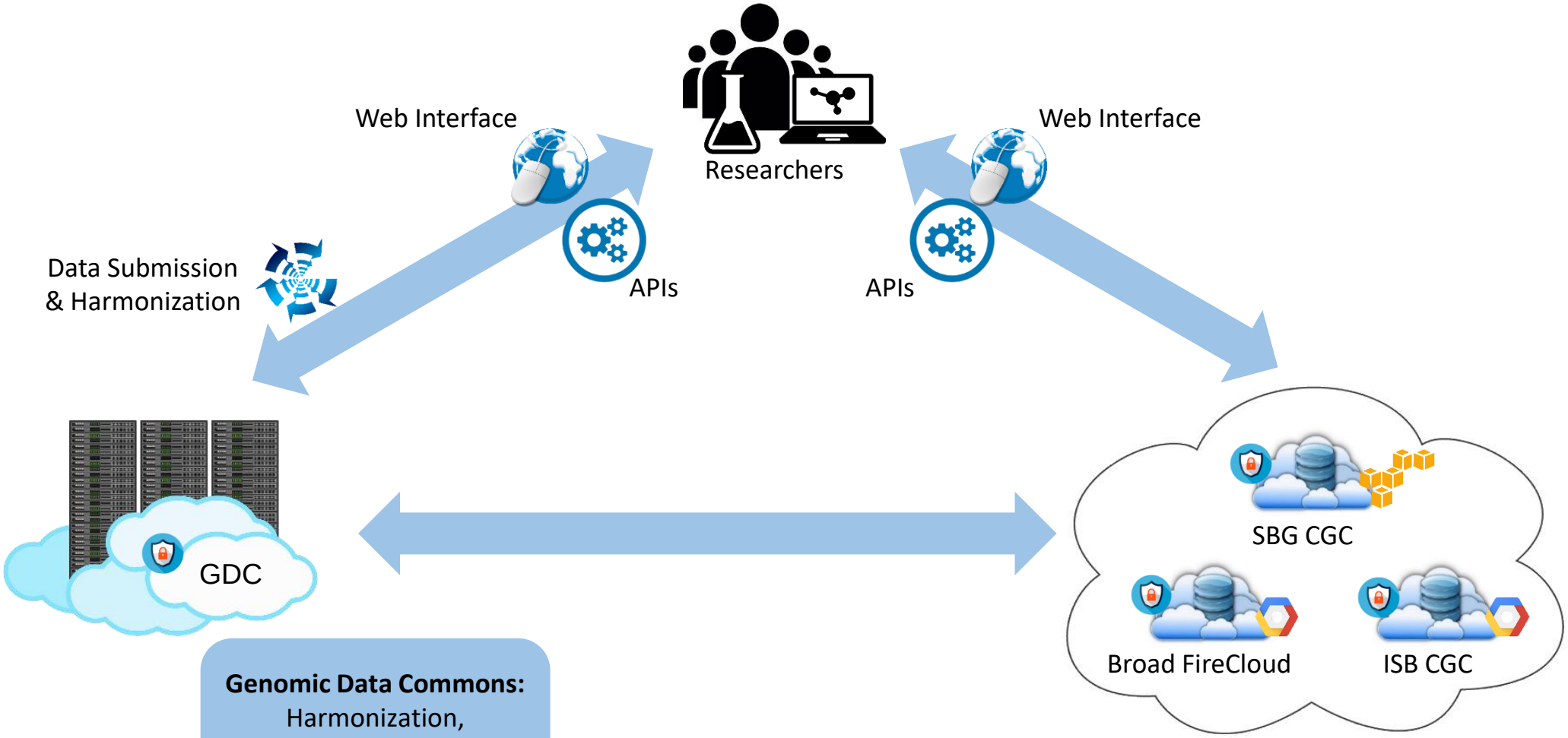


Genomic Data Commons:
Harmonization,
Visualization,
& Download

Genomics Cloud Pilots:
Visualization, Compute,
Pipelines, Workspaces

Authentication
& Authorization thru
eRA Commons & dbGaP

GDC and NCI Cloud ~~Pilots~~ Resources/Framework Today



NCI Cancer Research Data Commons (NCRDC)

Genomic node: GDC

Proteomic node: PDC

Imaging node: IDC

API's

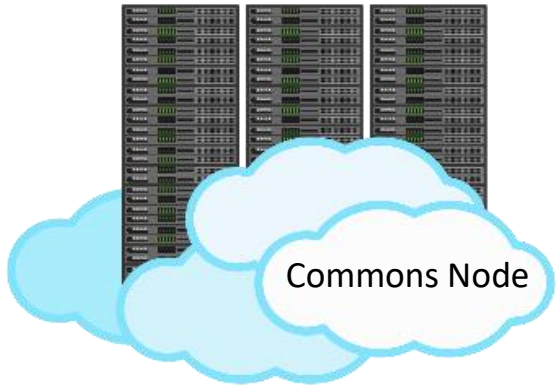
Framework – Core Services

Authentication
Data Model
Workflows/workspaces
Cross cloud services
Common architecture

Key Components of a Commons Node

Data Acquisition and Curation

Mechanisms for the community to contribute and store data; ensuring the quality, consistency, and usability of the data



Innovation

Analysis resources

New visualization and analytic tools; new data types; mechanisms for researchers to share data, results, and tools

DockStore

Support / Help / Operations

Support for the community's needs to use and maximize impact of each Commons Node

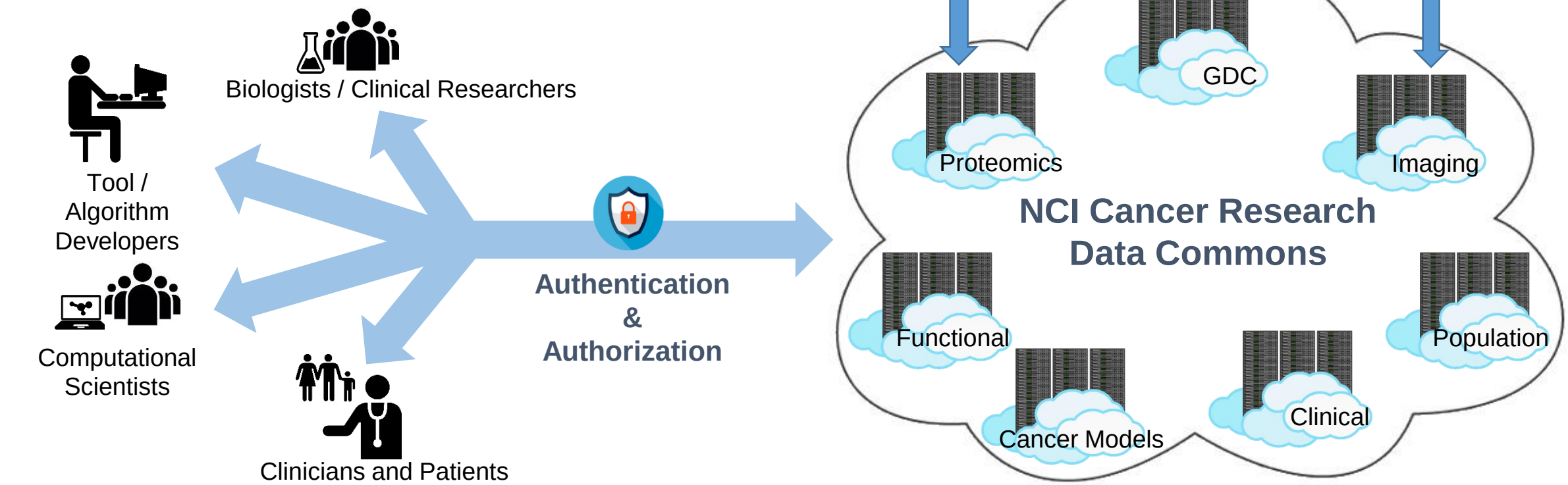
Infrastructure

Physical hardware, software, security policies and protocols, cloud computing

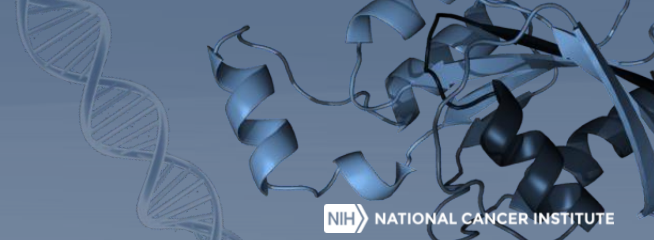
The NCI Cancer Research Data Commons
A virtual, expandable infrastructure

- Standardized data submission and Q/C
- Controlled vocabularies
- Harmonization by subject matter experts

- Secure data access through API or web UI
- Query across data domains
- Analytics, elastic compute, visualization

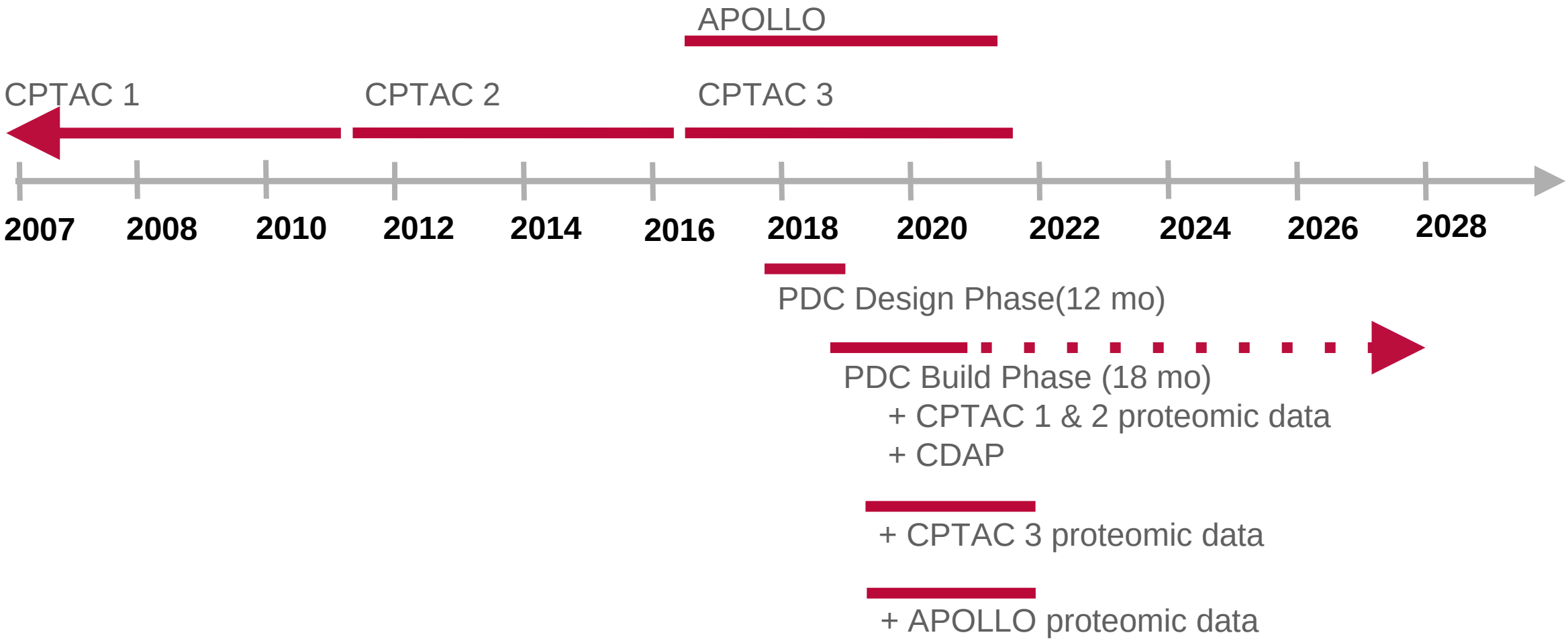


What other problem will the PDC solve?



- Increase access and usage to researchers beyond proteomics domain

PDC Timeline



What will the PDC do?

- Current award is to design and build a **prototype**
  - Gather requirements
  - Ensure interoperability with CRDC framework and other nodes
  - Design architecture, data model, analysis and visualization environments

Who is the PDC?

- ESAC
  - PI Anand Basu
  - Karen Ketchum, Rajesh Thangudu

Acknowledgments

NCI CBIIT

Izumi Hinkson
Tanja Davidsen
Tony Kerlavage
Juli Klemm
Allen Dearry

NCI CSSI

Jerry Lee
Sean Hanlon
Henry Rodriguez
Steve Cole

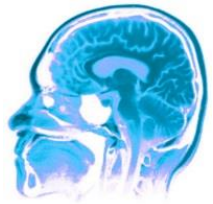
NCI OA

Andy May
Tim Crilley

Outside perspectives

Nuno Bandeira
Jake Jaffe

Opportunity for Big Data



Imaging

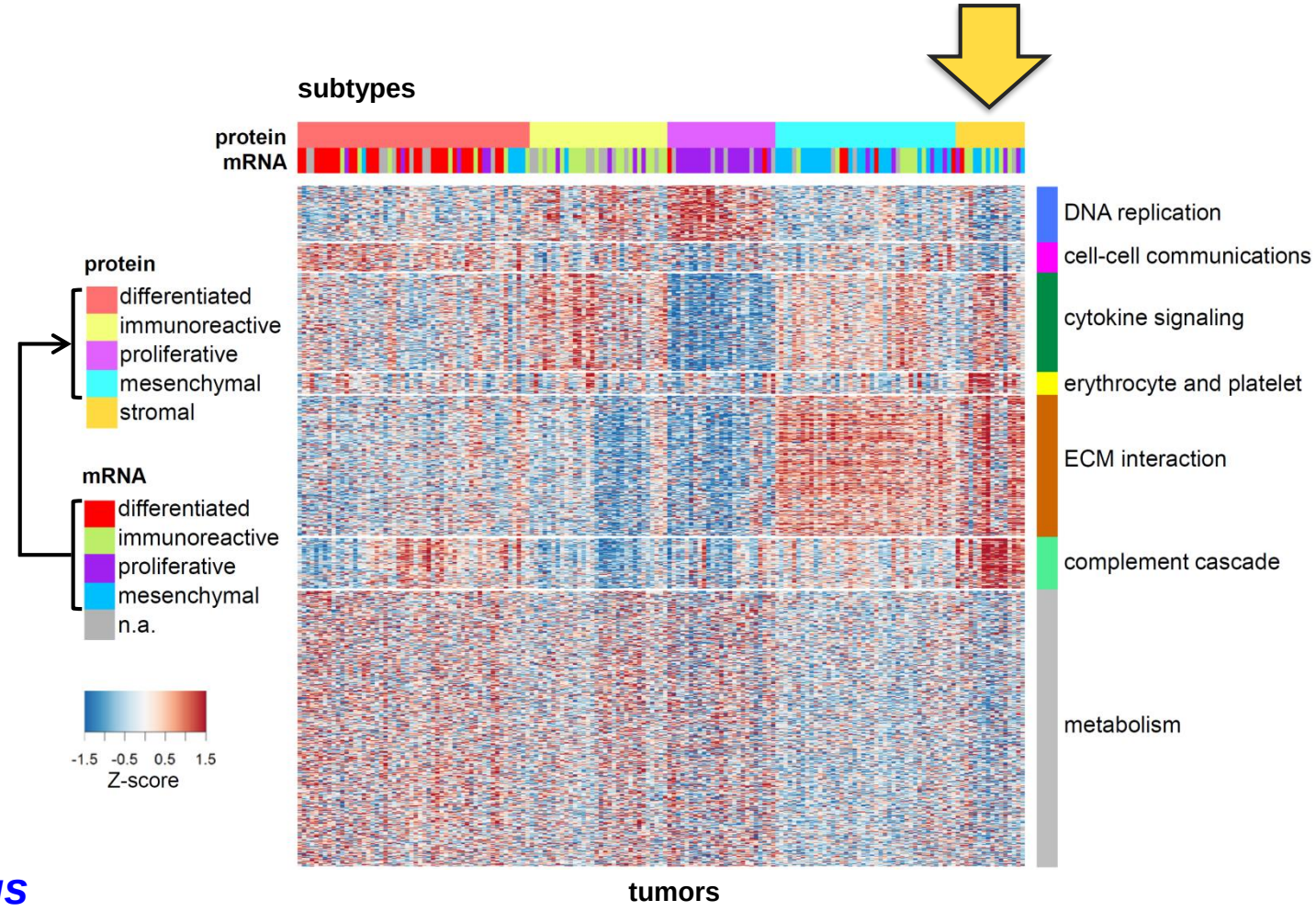
Genomics

Proteomics

Improved Prevention, Diagnosis, Treatment

Ovarian Cancer
(PROTEIN ABUNDANCE - new proteome subtype identified)

- **174 ovarian HGSC tumors**
  - Selection criteria:
    - Overall Survival (OS)
    - Homologous Recombination Deficiency status (HRD)
- **5 proteomic subtypes**
  (4 transcriptomic subtypes)
  - mRNA subtypes translate at protein level
  - New “stromal’ subtype emerged
- ***While interesting observations, no strong separation of OS and HRD status***



Cell *Cell* 166(3):755-65 (28 Jul 2016)

Clinical Proteomic Tumor Analysis Consortium (CPTAC)

NIH NATIONAL CANCER INSTITUTE

Proteogenomics builds on TCGA

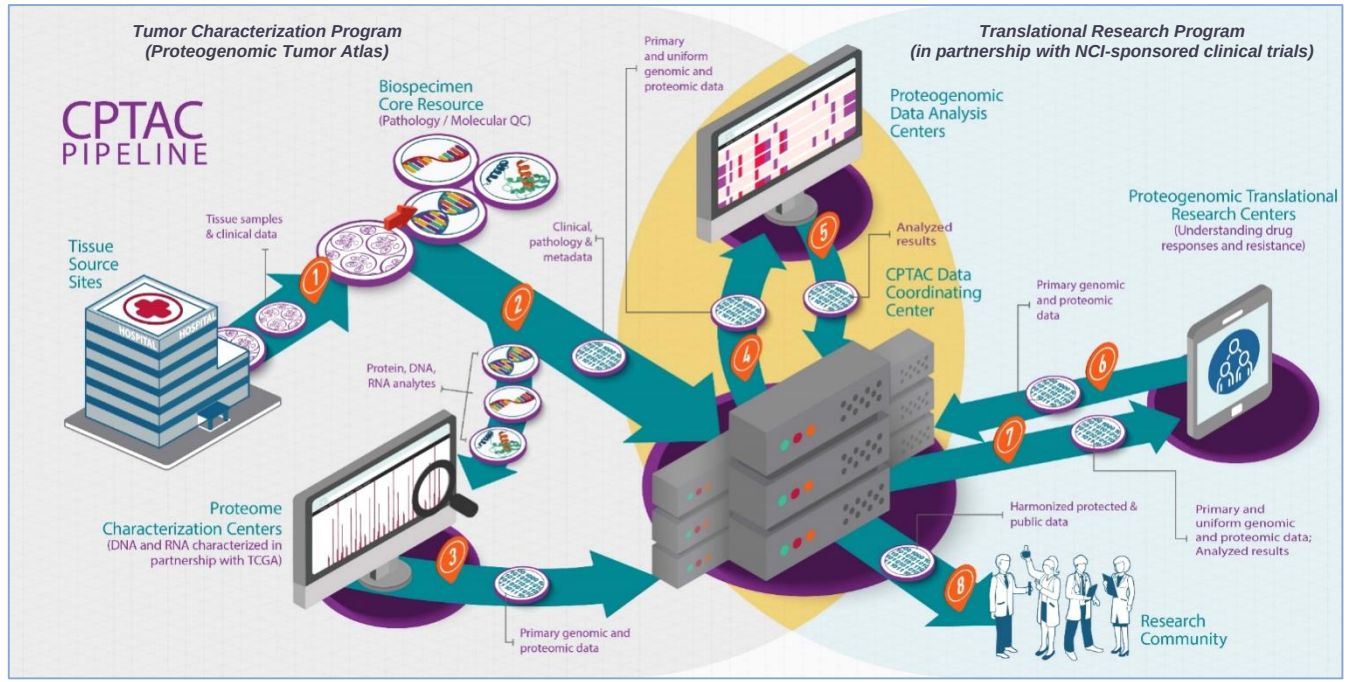
- **Tumor Characterization Program**
  (treatment naïve – tumor and NAT):

Data Generation	***Data Analysis***
JHU	*BCM*
PNNL	*UMich*
Broad	*NYU/WashU/BYU*
	MtSinai
	Broad

- **Translational Research Program**
  (pre-clinical and clinical trials):

  OHSU/PNNL
  BCM/Broad
  FHCRC/UAB

CPTAC builds on genomics



Clinical Proteomic Tumor Analysis Consortium (CPTAC)

NIH NATIONAL CANCER INSTITUTE

Proteogenomics builds on TCGA

- **Tumor Characterization Program**
  (treatment naïve – tumor and NAT):
  characterize proteins and genes to better understand the molecular basis of cancer

- **Translational Research Program**
  (pre-clinical and clinical trials):
  understand [predict] drug response and resistance to therapies in context of a clinical trial

CPTAC builds on genomics