

Cancer Research Data Commons (CRDC) Ecosystem Overview

Tanja Davidsen, Ph.D.

October 17, 2024

CRDC Website: datacommons.cancer.gov

Resources: <https://datacommons.cancer.gov/resources>

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Search

About ▾ Explore ▾ Analyze ▾ Submit ▾ Publications ▾ News ▾ Support ▾

Connecting Data to Accelerate Cancer Research

The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

[Watch CRDC Video](#)

9.4PB+ Data

134K+ Subjects / Participants

354 Studies

2K+ Public Tools and Workflows

82.3K+ Unique Users / Year

Explore Data

Overview

Many large NCI-funded research projects share their rich multi-modal data with Research Data Commons (CRDC). There are a number of ways users can explore CRDC's Data Commons portals, the Cancer Data Aggregator API and notebook Resources.

Among the many NCI-funded projects that share their data via the CRDC are:

- **APOLLO**: Applied Proteogenomics Organizational Learning and Outcomes Network
- **CCDI**: Childhood Cancer Data Initiative
- **CPTAC**: Clinical Proteomic Tumor Analysis Consortium
- **HTAN**: Human Tumor Atlas Network
- **TARGET**: Therapeutically Applicable Research to Generate Effective Treatments
- **TCGA**: The Cancer Genome Atlas

[Select Datasets](#) Learn more about CRDC hosted datasets.

Exploring Data using Data Commons

Each data commons provides a search interface to explore data by demographic name of a specific study, among other variables. Users can explore data from various initiatives, and can build cross-cutting "virtual" cohorts for aggregated analysis. In a cloud-based compute environment, users can build a data manifest to pull data into their local compute environment.

In addition to this general exploration, many data commons provide data visualization tools within the data portal environment.

[Data Commons](#) Learn more about each Data Commons.

Aggregated Exploration Across Data Commons

The Cancer Data Aggregator (CDA) combines descriptive information about CRDC data commons making it possible to search across multiple data commons via a single participant, sample, tissue, or disease.

While anyone can browse the CDA's indexed metadata, researchers wanting to work with actual data still need to apply for appropriate access to work with actual data (via the Cancer Data Aggregator).

[Cancer Data Aggregator](#) Learn more about the Cancer Data Aggregator.

Data Exploration through the CRDC Cloud Resources

The CRDC Cloud Resources (CR) also serve as entry points for exploring CRDC data commons, each with distinct features, provide secure workspaces and the ability to use analytical tools and workflows from their platforms.

One of the key benefits of using the CR is that users can access the data with amounts of data to a local compute environment, which can involve high download speeds.

[Cloud Resources](#) Learn more about the Cloud Resources.

CRDC Core Standards and Services

Ensuring that CRDC-hosted data meet the FAIR standards – Findable, Accessible, Reusable – data must be organized, stored, and searchable based on common core data standards and services related to data tracking and secure access across the CRDC data ecosystem.

[CRDC Standards and Services](#) Learn more about CRDC standards and services.

Select Datasets

The NCI CRDC provides access to a variety of open, registered, and controlled datasets from NCI programs and key external cancer programs. Many of these datasets are accessible through the [CRDC Cloud Resources \(CR\)](#), where users can also bring their own data to a secure cloud workspace for comparative analysis.

SELECT DATASETS		
DATASET NAME	DESCRIPTION	ACCESSIBLE THROUGH*
Applied Proteogenomics Organizational Learning and Outcomes (APOLLO)	A collaboration between NCI, the Department of Defense (DoD), and the Department of Veterans Affairs (VA), that incorporates proteomic data with patient care, with a focus on the activity and expression of the proteins that the genome encodes.	GDC, PDC, SB*
Cancer Genome Characterization Initiatives (CGCI)	An initiative examining genomes, exomes, and transcriptomes of various types of adult and pediatric cancers.	GDC, SB*, ISB*
Children's Brain Tumor Tissue Consortium (CBTTC)	A collaborative research consortia focused on identifying therapies for children with brain tumors.	PDC, SB*, ISB*
Childhood Cancer Data Initiative (CCDI)	A consortium of children's hospitals, clinics, or networks that make their clinical care and research data accessible.	CG, SB*
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	A national effort to accelerate the understanding of the molecular basis of cancer through the application of proteogenomics (large-scale proteome and genome analysis).	GDC, PDC, FC*, ISB*, SB*, IC*
Comparative molecular life history of spontaneous canine and human gliomas (GLIOMAD1)	A collaborative effort to characterize the genomic and transcriptomic landscape of canine glioma to enable cross-species comparative genomic analysis of sporadic glioma.	ICDC, SB*
Foundation Medicine (FM)	Foundation Medicine Inc., a molecular information company, makes accessible sequencing data from thousands of adult patients, in an effort to match patients with personalized treatment plans.	GDC, SB*, ISB*
Genetics and Epidemiology of Colorectal Cancer Consortium (GEOCC)	A research collaboration to detect colorectal cancer susceptibility loci using genome-wide sequencing.	CDS, SB*
Human Cancer Model Initiative (HCM)	An international consortium that is generating novel, next-generation, and tumor-derived culture models complete with genomic and clinical data.	GDC, SB*, ISB*

CRDC Insights: Quarterly Newsletter

<https://datacommons.cancer.gov/crdc-insights>

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

CRDC Insights Quarterly: March 2023

NIH Data Management and Sharing Policy: CRDC's Role

The new policy is in effect, and applies to new grant applications, competitive renewals, or competitive revisions. In brief:

- Data sharing now pertains to all researchers with no budget minimum.
- Applications, renewals, or revisions must include a data management and sharing plan.
- Data must be shared at time of publication or by the end of the performance period, whichever is sooner.

The Cancer Research Data Commons (CRDC) is home to a collection of data commons and cloud resources that host datasets from NCI-funded research, and make those datasets accessible to the research community. [Learn more about submitting and accessing data, and using CRDC tools for your research.](#) [\(link to data landing page\)](#)

CRDC Resources in the Classroom



Faculty members and data scientists with Purdue University and Velsara (Seven Bridges) teamed up to produce a four-part online workshop that introduces the Cancer Genomics Cloud. The series also provides hands-on lessons in bulk- and single-cell RNA-seq analysis using datasets provided by Purdue researchers. [Read the full story.](#)

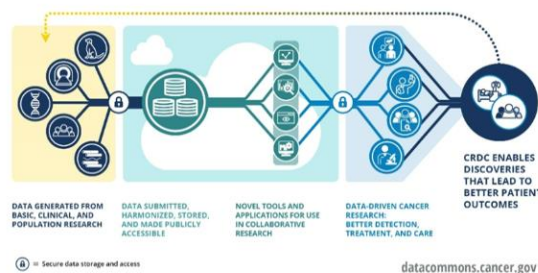
HTAN: Methods Workshop at AACR 2023 Annual Meeting

The Human Tumor Atlas Network (HTAN) is working closely with CRDC to ensure long-term legacy and reuse of HTAN data, and to share data through NCI's Cloud Resources. This methods workshop will demonstrate how to access, query, use data within the cloud environment, and visualize HTAN data derived from a variety of assay types. Read more about this workshop on the [AACR Annual Meeting website](#).

Announcement: Funding Opportunities

The Office of XYZ has released a RFP/grant solicitation regarding data interoperability. Find more information on their Interoperability Initiative page.

CRDC: Empowering the Scientific Community to Make New Discoveries



A new infographic illustrates how CRDC supports the work of cancer researchers. This is available for use in presentations. Contact our general email box below.

In the News



NCI Director Monica Bertagnolli was recently interviewed by National Public Radio (NPR) about the work of the NCI and its impact on patients and families. She also discussed her own cancer diagnosis and her commitment to participating in research trials. Listen here through the NPR website.

About the Cancer Research Data Commons

The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

Subscribe to this newsletter here: [LINK TO SUBSCRIBE](#) button on our news page.

Quick Links

Data Releases: Updated March 2023

A round-up of new datasets available through our data commons and cloud resources.

Datasets, Access, and Submission

Aggregated information across CRDC data commons and cloud resources about filing a data submission request, and how to access data currently housed in a CRDC data commons or on a cloud resource.

Getting Started

Aggregated listings of user manuals, tutorials, and virtual office hours, across CRDC data commons and cloud resources.



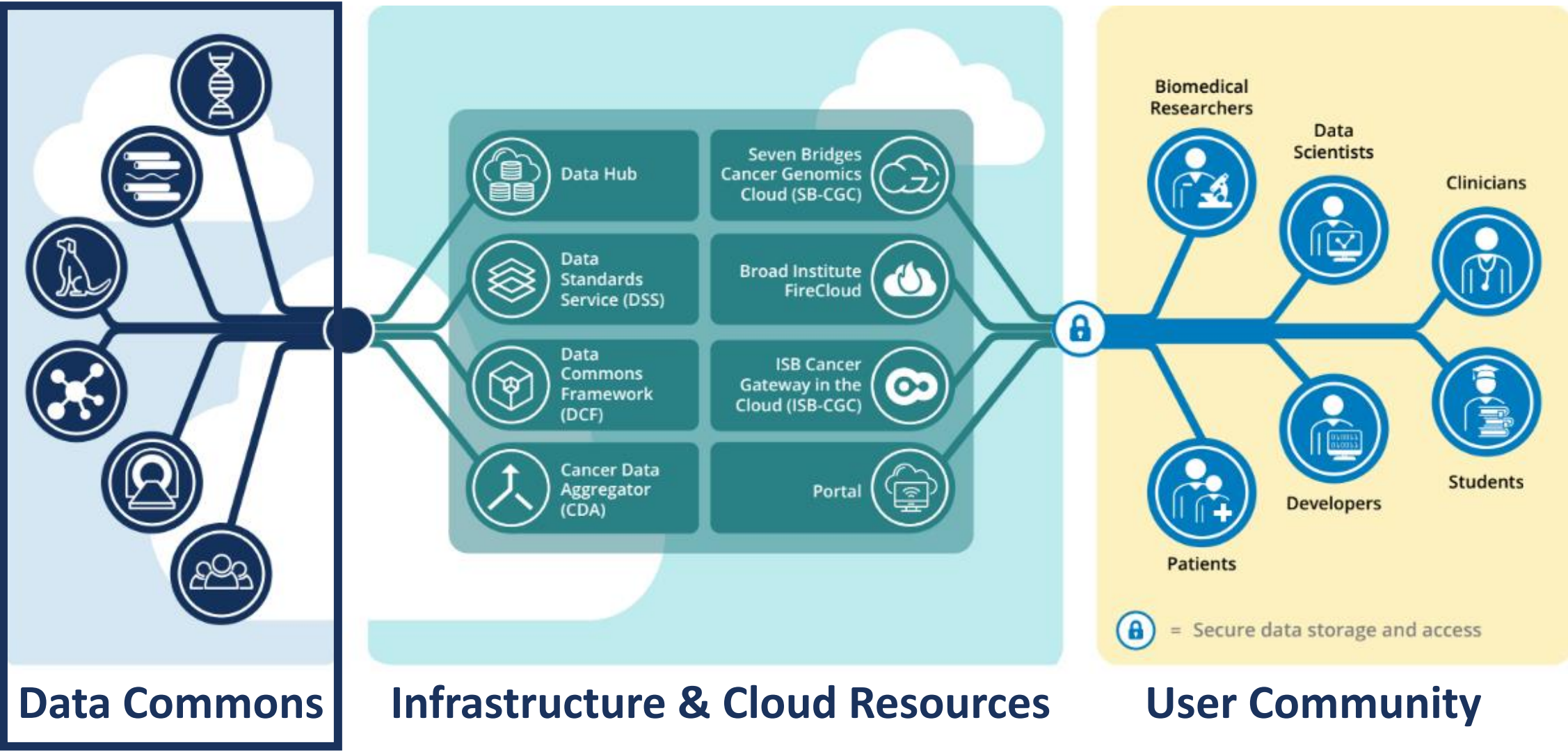
Contact us: NCICRDC@mail.nih.gov

And subscribe to this newsletter on our CRDC Insights page.

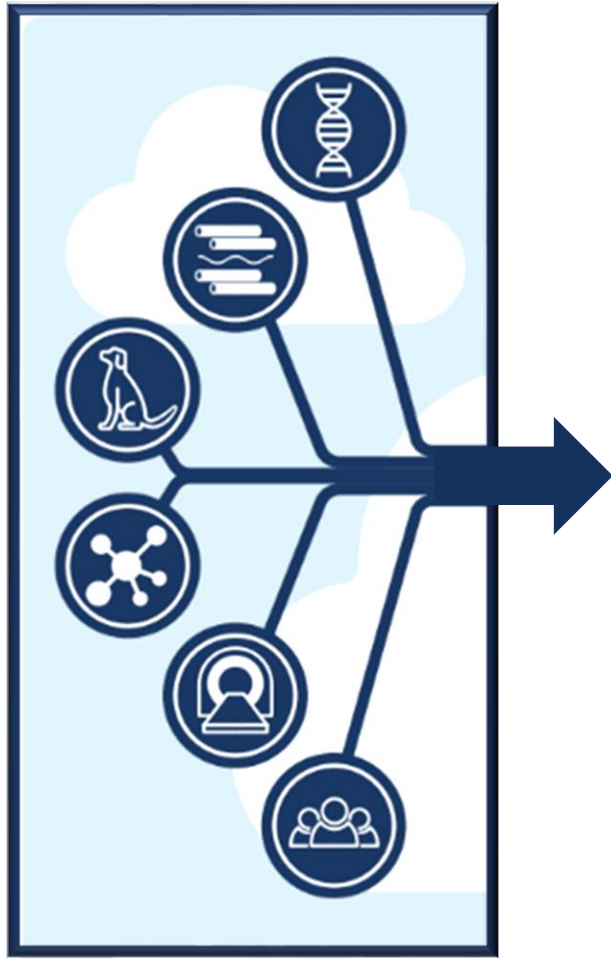


CRDC Ecosystem: Data Commons

CRDC Ecosystem: Data Commons



CRDC Ecosystem: Data Commons



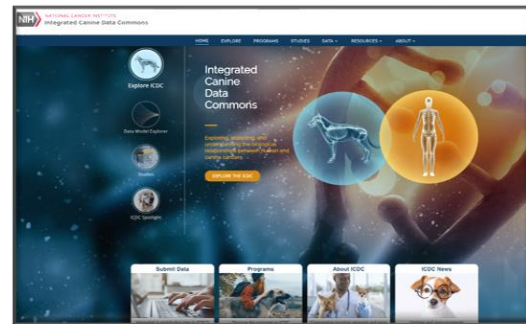
Genomic Data Commons (GDC)



Proteomic Data Commons (PDC)



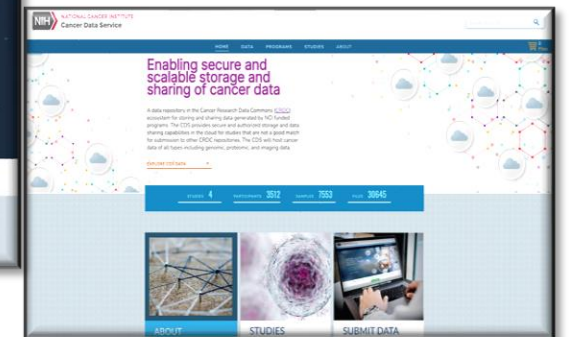
Imaging Data Commons (IDC)



Integrated Canine DC (ICDC)

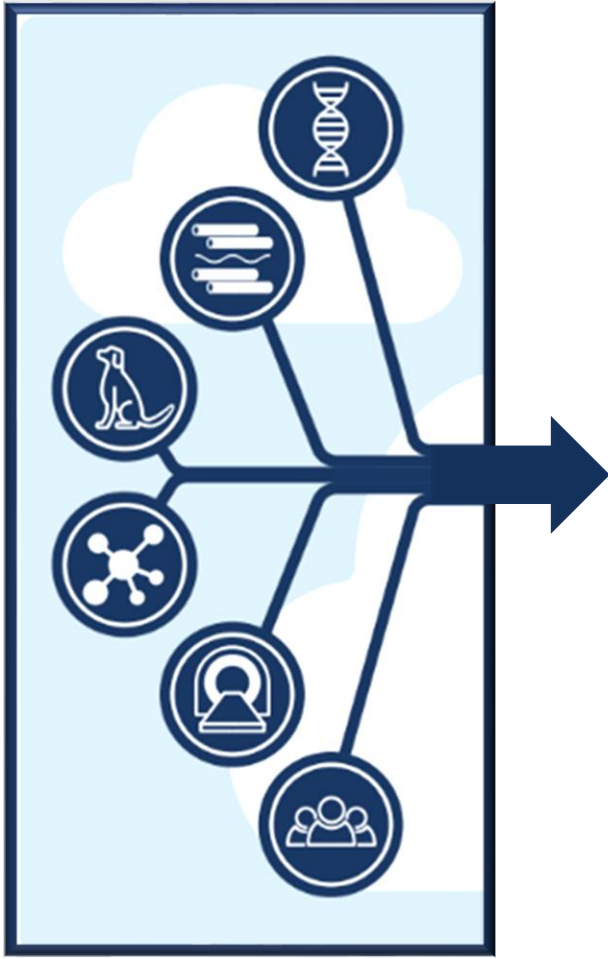


Clinical & Translational DC (CTDC)



Cancer Data Service (CDS)

CRDC Ecosystem: Data Commons



Coming Soon...
**Population Science Data
Commons currently in
development and testing**


Genomic Data Commons (GDC)

<https://portal.gdc.cancer.gov>

NIH NATIONAL CANCER INSTITUTE

Genomic Data Commons Data Portal

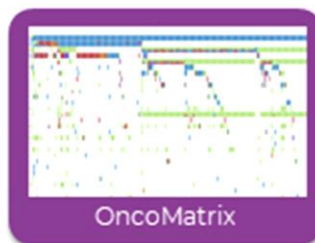
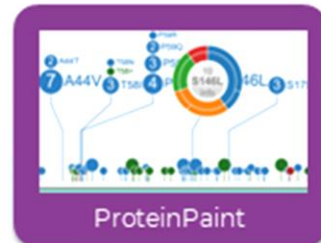
Harmonized Genomic and Clinical Data
Web-Based Analysis and Visualization Tools
API Data Access Options



Cases by Major Primary Site

Primary Site	1000s of Cases
Adrenal Gland	0.1
Bile Duct	0.1
Bladder	0.2
Bone	0.1
Bone Marrow and Blood	8.0
Brain	0.2
Breast	4.0
Cervix	0.2
Colorectal	3.0
Esophagus	0.2
Eye	0.1
Head and Neck	0.2
Kidney	2.0
Liver	0.2
Lung	5.0
Lymph Nodes	0.2
Nervous System	0.2
Ovary	1.0
Pancreas	0.2
Pleura	0.1
Prostate	0.2
Skin	0.2
Soft Tissue	0.1
Stomach	0.2
Testis	0.1
Thymus	0.1
Thyroid	0.2
Uterus	0.2

Introducing GDC 2.0: Analyze your custom cohort with new web-based tools



Explore all the data and tools at portal.gdc.cancer.gov

Genomic Data Commons (GDC)

<https://portal.gdc.cancer.gov>

- **Data:** Filter, query, visualize, analyze & download
- **Harmonized** to same genome standard and variant calling pipeline
- **New:** WGS Variant Callers, Copy Number Variation, new single cell RNA-Seq data, new whole slide images
- **Programs:** TCGA, TARGET, CPTAC, MATCH, more
- **GDC 2.0 Launched June 2024:** Empowers researchers with a cohort-centric approach
 - Custom Cohort Creation & Data Download
 - Advanced Analysis Tools
 - SDK for Tool Integration



Users: >90K+ unique users per month

Citations: 2,500+



Downloads: >4PBs data downloaded per month



Global Access:

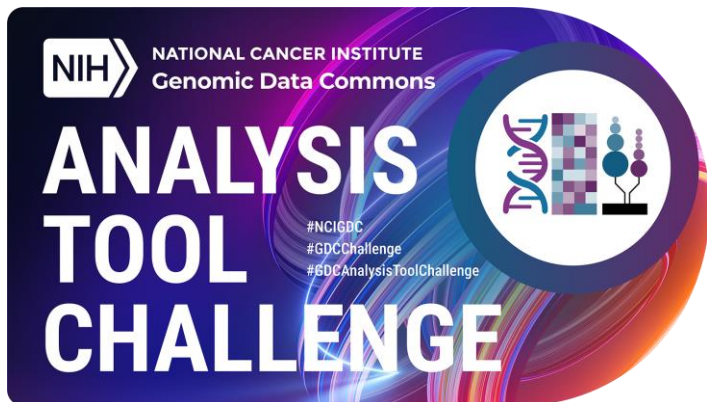
Users from >90 different countries

GDC Analysis Tool Challenge

Join the GDC Analysis Tool Challenge and help shape the future of cancer research by integrating a novel analysis tool to explore GDC data

- **Requirements:**

- **Scientific Need & Innovation:** Integrate a tool that addresses a scientific need and goes beyond existing GDC Analysis Tools
- **GDC Integration:** Utilize GDC data and leverage the GDC SDK for tool integration
- **Open-source:** Integrated tools must support open-source distribution

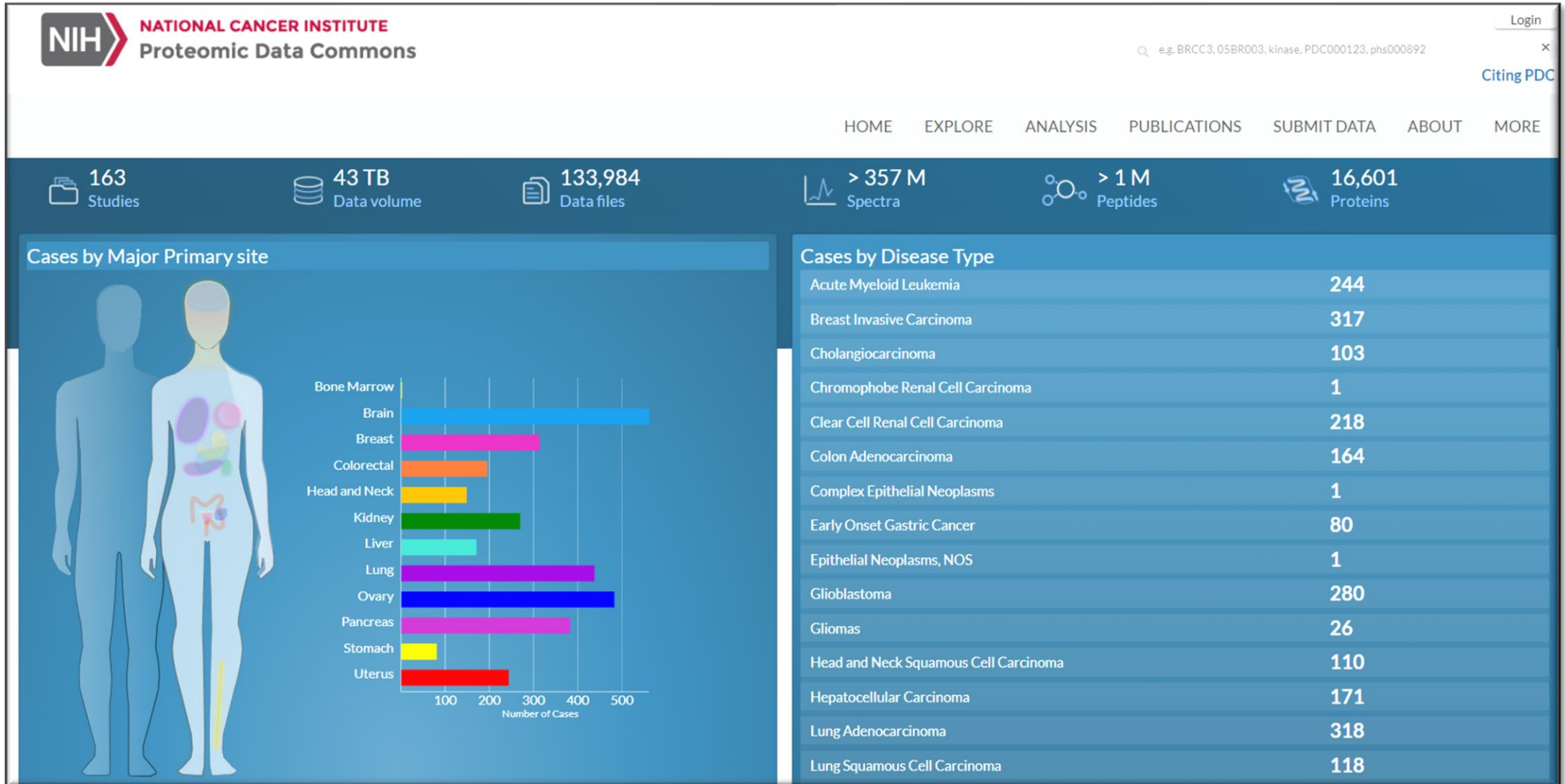


Register today!
Closes October 31, 2024

<https://www.challenge.gov/?challenge=nci-gdc-analysis-tool-challenge>

Proteomic Data Commons (PDC)

<https://pdc.cancer.gov>

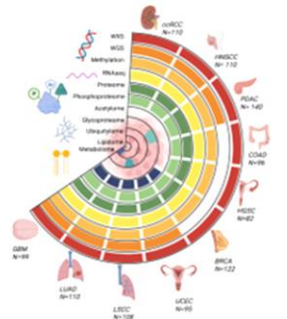


Proteomic Data Commons (PDC)

<https://pdc.cancer.gov>

- **Data:** Filter, query, visualize & download
- **Data Harmonization:** for uniform analysis of all PDC data
- **Programs:** CPTAC, ICPC, APOLLO, TCGA
- **New:** Now supports metabolomics & lipidomics data
- **PDC By The Numbers:**
 - 163 Studies
 - +130K Data Files
 - 43 TB Data
 - >1PB Download
- Data Analysis in the cloud facilitated by the NCI Cloud Resources

CPTAC's Pan-Cancer Multi-omic Papers & Data



Annual Clinical Data Updates

New Clinical Data

Access 1,000 new clinical outcome data files from Clinical Proteomic Tumor Atlas Consortium (CPTAC) through #NCICommons.



Cancer Research Communications

@CRC_AACR

Recently published:

NCI's Proteomic Data Commons: A Cloud-Based Proteomics Repository Empowering Comprehensive Cancer Analysis Through Cross-Referencing with Genomic & Imaging Data, by Ratna R. Thangudu et al.

<https://doi.org/10.1158%2F2767-9764.CRC-24-0243>

Imaging Data Commons (IDC)

<https://portal.imaging.datacommons.cancer.gov/>

NIH NATIONAL CANCER INSTITUTE
Imaging Data Commons

Explore Images Collections Getting Started User Forum News About Help Sign In

Explore Images

Cases by Major Primary Site

Primary Site	Cases (Approximate)
Adrenal Gland	150
Bile Duct	50
Bladder	100
Blood	100
Brain	150
Breast	1,000
Cervix	150
Chest	1,500
Colorectal	100
Esophagus	50
Head and Neck	100
Kidney	100
Liver	100
Lung	1,000
Ovary	100
Pancreas	100
Prostate	100
Skin	100
Stomach	100
Testis	100
Thymus	100
Thyroid	100
Uterus	100

Data Portal Summary
Data Release 19.0 August 20, 2024

145 Collections

67,307 Cases

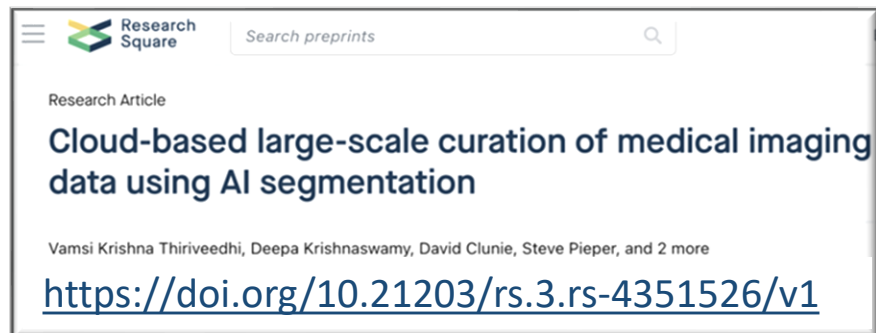
78.78 TB Data Volume

937,495 Image Series

Imaging Data Commons (IDC)

<https://portal.imaging.datacommons.cancer.gov/>

- **Share, Analyze, Visualize:** de-identified, harmonized, multi-modal, open access imaging data
- **Programs:** CCDI, CMB, HTAN, GTEx, TCGA, CPTAC, more
- **Data:** Images and annotations
 - ~80TB data, ~1M images
 - DICOM standard
 - Interoperable & Searchable
- **Quantitative features:**
 - Lower barriers for analysis
 - Enable cross-omics studies
- **Continuous Data Enrichment:**
 - Expert and AI-based curation
 - Collaborations, benchmarking, comparison studies
- **Cloud-based computing at scale:**
 - Manual annotation not practical
 - Unparalleled ability to scale AI-based automatic curation
 - Enabled by Cloud Resources



Integrated Canine Data Commons (ICDC)

<https://caninecommons.cancer.gov/>

The screenshot shows the homepage of the Integrated Canine Data Commons (ICDC). At the top left is the NIH logo and the text "NATIONAL CANCER INSTITUTE Integrated Canine Data Commons". At the top right is a search bar labeled "SEARCH THE ICDC". Below the header is a navigation menu with links for HOME, EXPLORE, PROGRAMS, STUDIES, DATA, RESOURCES, and ABOUT. On the far right of the navigation bar is a "MY FILES" icon with a zero notification count. The main content area features a large background image of a dog's paw and a human hand. On the left side, there are three circular icons: a dog (labeled "Explore the ICDC"), a funnel (labeled "Data Model Navigator"), and a server rack (labeled "Studies"). In the center, the title "Integrated Canine Data Commons" is displayed in large white text. Below the title is a short paragraph: "Exploring, analyzing, and understanding the biological relationships between human and canine cancers." At the bottom center is a prominent orange button labeled "EXPLORE THE ICDC". On the right side of the main content area, there are two large circular graphics: a blue circle containing a white wireframe of a dog's skeleton, and a yellow circle containing a white wireframe of a human skeleton.

Integrated Canine Data Commons (ICDC)

<https://caninecommons.cancer.gov/>

- **Canine Clinical Trial Data:**
 - PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT)
 - Comparative Oncology Program
- **Recent Accomplishments:**
 - "Leading the pack: Best practices in comparative canine cancer genomics to inform human oncology"
 - Direct export of genomic files to Seven Bridges Cancer Genomics Cloud (SB-CGC)
- **Future Directions:**
 - Database of annotated biospecimen locations, biobank contact information, and data from other sources like veterinary foundations
 - Longitudinal data
 - Include CDEs in Data Model
 - New studies: PRECINCT01, COTC021, OSA02, OSA04
 - **Data Submission:** via CRDC Centralized Data Submission Portal



Clinical & Translational Data Commons (CTDC)

<https://clinical.datacommons.cancer.gov/>

The screenshot shows the homepage of the Clinical & Translational Data Commons (CTDC). The top navigation bar includes links for "Explore", "Studies", "Data", "For Developers", "About", and "Request Access". On the right side of the navigation bar, there is a "LOGIN" button and a shopping cart icon with "0 FILES". The main content area features a large background image of a doctor in a white coat and stethoscope, holding a tablet and talking to a patient. Overlaid on the right side of this image is a circular menu with the text "CLINICAL & TRANSLATIONAL DATA COMMONS" in the center. Surrounding this central text are five icons and labels: "PARTICIPANTS" (with a group of people icon), "DIAGNOSES" (with a stethoscope icon), "TARGETED THERAPIES" (with a cross-in-a-circle icon), "FILES" (with a document icon), and "PARTICIPANTS" (with a group of people icon).

Explore Studies Data ▾ For Developers ▾ About ▾ Request Access

LOGIN  0 FILES

CLINICAL & TRANSLATIONAL DATA COMMONS

PARTICIPANTS

DIAGNOSES

TARGETED THERAPIES

FILES

FUELING DISCOVERY: HARNESSING THE POWER OF DATA FROM CANCER STUDIES

Clinical & Translational Data Commons (CTDC)

<https://clinical.datacommons.cancer.gov/>

- **Launched:** September 2024
- **Datatypes:** clinical, genomic, molecular, PRO, pharmacological and more
- **Access Levels:** Hosting open, registered, controlled access
- **Accepting Data:** from NCI-funded clinical studies (not just trials) including immunoncology data
- **First Dataset:** Cancer Moonshot Biobank
- **Future Datasets:** to include
 - Molecular Analysis for Therapy Choice (MATCH) Trial
 - Cancer Immune Monitoring and Analysis Centers (CIMAC)
- **Data Submission:** via CRDC Centralized Data Submission Portal



Cancer Data Service (CDS)

<https://dataservice.datacommons.cancer.gov/>

The screenshot shows the top section of the Cancer Data Service website. At the top left is the NIH logo and the text "NATIONAL CANCER INSTITUTE Cancer Research Data Commons". To the right is a search bar labeled "SEARCH CDS". Below this is a dark blue navigation bar with links for "HOME", "DATA", "PROGRAMS", "STUDIES", and "ABOUT", along with a shopping cart icon showing "0 Files". The main content area features a large purple heading: "Enabling secure, scalable storage and sharing of cancer data". Below this is a paragraph of text describing the CDS Portal as a data repository in the Cancer Research Data Commons (CRDC) ecosystem. An orange button labeled "EXPLORE CDS PORTAL" with a right-pointing arrow is positioned below the text. At the bottom, a blue bar displays four statistics: "STUDIES 30", "PARTICIPANTS 49892", "SAMPLES 80230", and "FILES 329972". The background of the main content area is decorated with a network diagram of nodes and lines, and several circular icons containing cloud symbols.

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

SEARCH CDS

HOME DATA PROGRAMS STUDIES ABOUT 0 Files

Enabling secure, scalable storage and sharing of cancer data

A data repository in the Cancer Research Data Commons (CRDC) ecosystem for storing and sharing data generated by NCI funded programs. The CDS Portal provides secure and authorized storage and data sharing capabilities in the cloud for studies that are not a good match for submission to other CRDC repositories. The CDS Portal will host cancer data of all types including genomic, proteomic, and imaging data.

EXPLORE CDS PORTAL


STUDIES **30** PARTICIPANTS **49892** SAMPLES **80230** FILES **329972**

Cancer Data Service (CDS)

<https://dataservice.datacommons.cancer.gov/>

- **Cloud-Based Repository:** secure and scalable storage for open or controlled access data
- **Datatype Agnostic:** CDS has a flexible data model for storing and sharing a variety of datatypes
 - Genomic, proteomic, imaging data, others on request
- **Since January 2022:**
 - 2.5 PB data submitted
 - 1.998 PB data released
- **Supporting compliance** with NIH data sharing policies
- **Data Discovery:** CDS Portal allows users to explore cancer research open harmonized (meta)data
- **Cloud-Based Analysis Tools:** one click manifest export to SB-CGC
 - Seamless analysis in the cloud - bring your own tools and data
- **Programs:**
 - Human Tumor Atlas Network (HTAN), Childhood Cancer Data Initiative (CCDI), CPTAC, PDXNet
- **Data Submission:** via CRDC Centralized Data Submission Portal



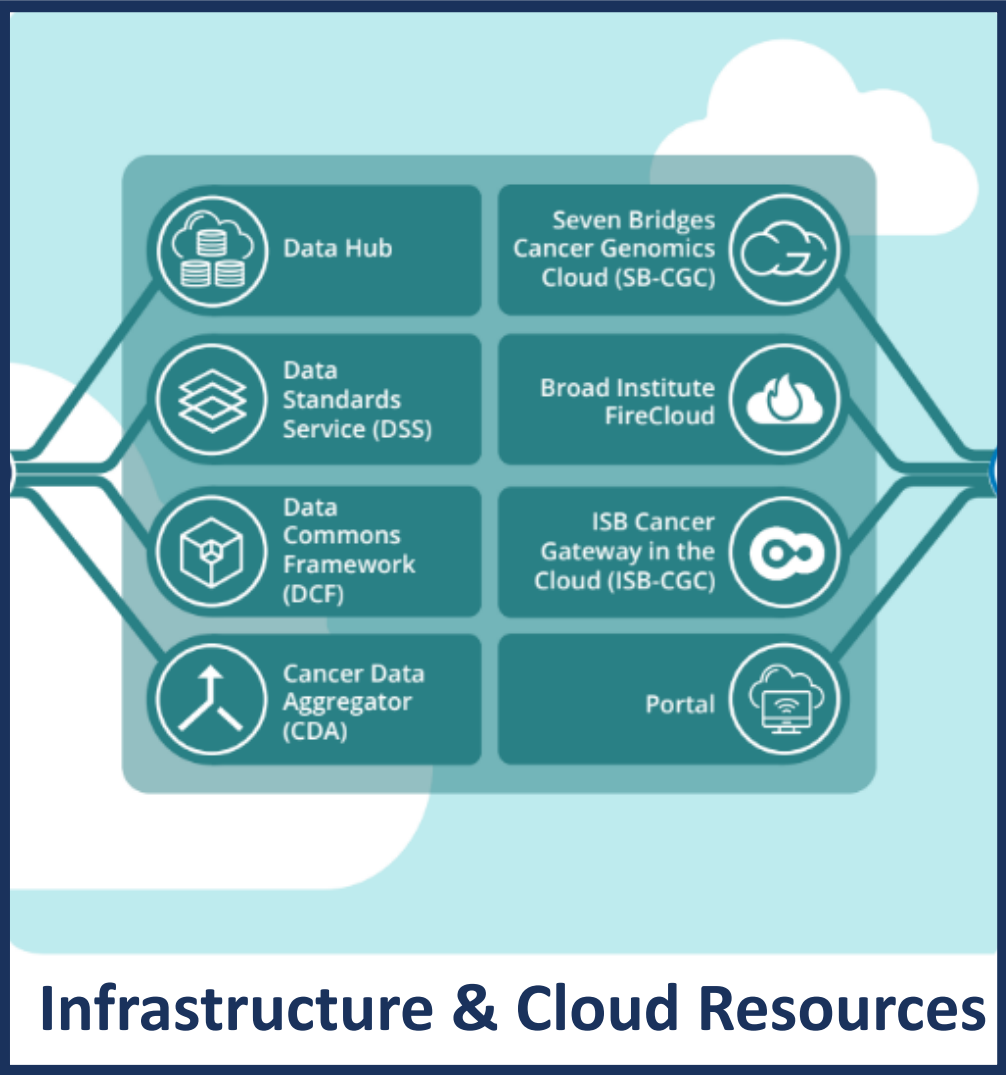


CRDC Ecosystem: Infrastructure and Cloud Resources for Interoperability

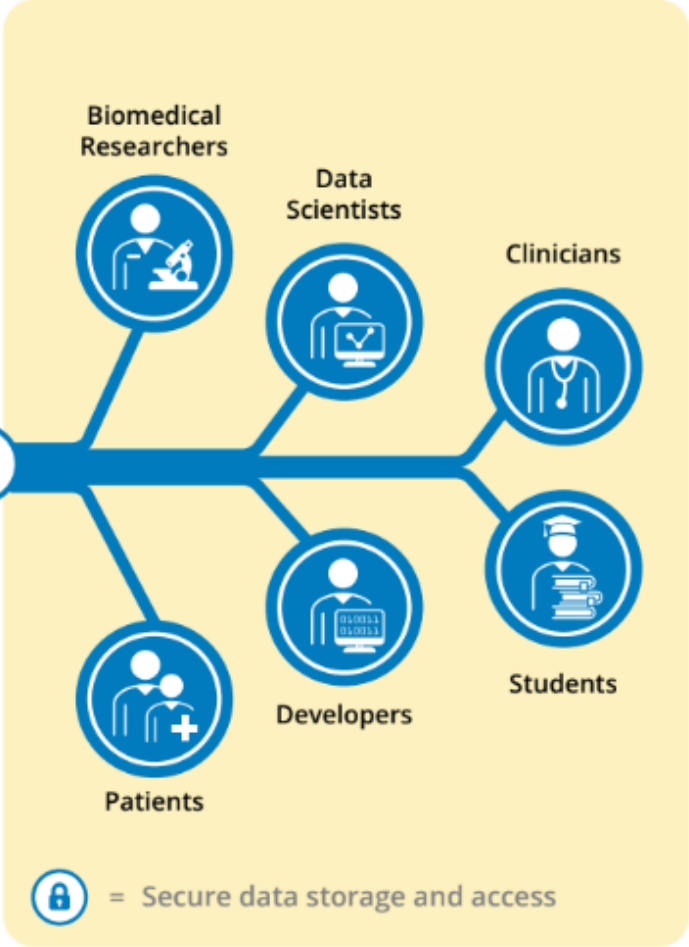
CRDC Ecosystem: Infrastructure & Cloud Resources




Data Commons



Infrastructure & Cloud Resources



User Community

 = Secure data storage and access

CRDC: Data Commons Framework

A re-usable,
expandable
framework

Core principles
and structures

Modular
components



Fence

**Centralized Authentication
& Authorization**



IndexD

Centralized Indexing



Cloud Storage


**CRDC Data in
Amazon and Google**

CRDC: NCI Cloud Resources

Democratizing access to cancer research data

- Access to **large cancer data sets** without need to download or move data
- Access to **workspaces, analysis tools, and workflows/pipelines**
- **Bring your own data and tools:** collaborative pre-publication workspaces
- Funds to get you started as a new user: **\$300-\$10,000**



ISB's Cancer Gateway
in the Cloud 

Great for command-line,
BigQuery, Specialty DBs



Broad's
FireCloud 

Great for running
production pipelines



Seven Bridges' Cancer
Genomics Cloud 

Great for non-technical user
Interface, visual displays



CRDC: NCI Cloud Resources

Institute for Systems Biology Cancer Gateway in the Cloud (ISB-CGC)

- **ISB-CGC's Cohort Builder:** combined with Cancer Data Aggregator can now generate cross-data-common cohorts
- **Google BigQuery:** Tabular CRDC data stored and browsable through BigQuery search tool
- **ISB-CGC-hosted Mitelman Database of Chromosomal Aberrations:** includes search optimizations and integrated data visualizations
- **Bioinformatics Notebooks:** ISB-CGC's growing collection includes guidance on how to use multiple datatypes

A RESOURCE OF THE NCI CANCER RESEARCH DATA COMMONS

ISB-CGC

Cancer Gateway in the Cloud

Access, Explore and Analyze Large-Scale Cancer Data Through the Google Cloud

- BigQuery Table Search**
Browse BigQuery tables of metadata and molecular cancer data from the Genomic Data Commons and other sources. Jump directly to a table to perform discovery and computation via SQL.
- Cancer Data File Browser**
Explore a comprehensive selection of cancer related data files in Google Cloud Storage Buckets, such as raw sequencing, cancer nucleotide variation, pathology or radiology images.
- Chromosomal Aberrations & Gene Fusions DB**
Browse the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer which relates cytogenetic changes, in particular gene fusions, to tumor characteristics.
- Cohort Builder / Data Explorer**
A web interface to build cohorts based on clinical demographics and molecular filters. Compare patient cohorts with various exploration tools including image viewers and the Cancer Data File Browser.
- Pipelines and Cancer Cohort API**
Learn more about how to access and analyze cancer data through programmatic interfaces including Google Cloud virtual machines and APIs.
- Notebooks**
A collection of notebooks written in R and Python, to serve as both tutorials or analysis tools for a range of users; includes reproductions of Regulome Explorer functionality.

<https://portal.isb-cgc.org/>



CRDC: NCI Cloud Resources

Broad's FireCloud, powered by Terra

- FireCloud allows data integrations with GDC, PDC, and IDC
- FireCloud is a cloud workbench - you can use CPU and GPUs to power your analysis, and scale as needed
- The FireCloud infrastructure is used to power the curation of >250,000 WGS (and growing) concurrently for the All of Us resource

The screenshot shows the FireCloud website interface. At the top left is a hamburger menu icon. To its right is the 'FireCloud' logo with a flame icon, followed by 'POWERED BY Terra' with the Terra logo. The main heading is 'Welcome to FireCloud'. Below this is a paragraph: 'FireCloud is a NCI Cloud Resource project powered by Terra for biomedical researchers to access data, run analysis tools, and collaborate.' There are three links: 'Find how-to's, documentation, video tutorials, and discussion forums', 'Already a FireCloud user? Learn what's new.', and 'Learn more about the Cancer Research Data Commons and other NCI Cloud Resources'. At the bottom, there are three white cards with blue arrows pointing right. The first card is 'View Workspaces' with the text: 'Workspaces connect your data to popular analysis tools powered by the cloud. Use Workspaces to share data, code, and results easily and securely.' The second card is 'View Examples' with the text: 'Browse our gallery of showcase Workspaces to see how science gets done.' The third card is 'Browse Data' with the text: 'Access data from a rich ecosystem of data portals.'

<https://firecloud.terra.bio/>



CRDC: NCI Cloud Resources

Seven Bridges Cancer Genomics Cloud (SB-CGC), powered by Velsera

- **Easy to Use:** point-and-click interface, >1,000 informatics workflows & tools
- **Extensive Support:** tutorials, public projects, office hours, and onboarding videos for all user types
- **Cutting-edge Research Support:** capitalizing on advances in ML & genAI
- **Interoperability/FAIR:** supporting multiple workflow languages
- **Citations:** >130 total
- **Training:** Committed to training the next generation of scientists

Public Apps Log in

Public apps for your data analysis

We offer publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

[Browse 1,044 apps](#)

CWL
RNA-seq alignment - STAR 2.5.4b
Toolkit version: STAR 2.5.4b

This workflow performs the first step of RNA-seq analysis - alignment to a reference genome and transcriptome. STAR (Spliced Transcripts Alignment to a Reference), an ultrafast RNA-seq aligner, is used in this workflow. STAR is capable of mapping full length RNA sequences and detecting de novo cano...

Alignment RNA-Seq

CWL
Whole Exome Sequencing - BWA +...
Toolkit version: GATK 4.1.0.0

This Whole Exome Sequencing (WES) workflow identifies variants from a human exome experiment by using the Broad Instit...

WES(WXS)

CWL
Whole Genome Sequencing - BWA +...
Toolkit version: GATK 4.1.0.0

This Whole Genome Sequencing (WGS) workflow identifies variants from a human whole-genome resequencing experiment by u...

WGS

CWL
[obsolete] Multi-instance Whole Geo...
Toolkit version: GATK 4.1.0.0

Configurable multi-instance workflow that processes whole genome reads in 2-3 hours. This workflow is based on several...

WGS

[Run](#) [Run](#) [Run](#)

<https://www.cancergenomicscloud.org/>

CRDC: NCI Cloud Resources

Training Tomorrow's Data Scientists



The CGC Community

CGC as a Teaching and Training platform

The National Institutes of Health (NIH) has made several four funding basic and translational research which generates ma as well as data analysis ecosystems or platforms to empower community to analyze these large datasets. Over the years, the Genomics Cloud platform has not only enabled researchers to research, but it has been used as an effective teaching platform for training the generation of national research workforce. Several universities and research institutes are effectively utilizing the CGC for their teaching and training progra

SUCCESS STORIES

1. TEACHING USING THE CGC - GEORGETOWN UNIVERSITY

We interviewed our long term collaborators Dr. Yuriy Gusev and Ms. Krithika Bhuvaneshwar from Georgetown University who have been using the CGC platform for the past 3 years to train the next generation of data scientists in their Masters in Health Inf interview:

2. TEACHING USING THE CGC - UCSD

Professor Jeremy Chien from UC Davis used CGC to teach a streamlined, hands-on course in cancer genomics to undergraduates. The platform eliminated the lengthy setup and steep learning curve usually associated with bioinformatics, allowing students to dive directly into real-world data analysis. The results? Engaged students who gained

3. TEACHING USING THE CGC - PURDUE UNIVERSITY

We partnered with Min Zhang at Purdue University to deliver a four-part series on RNAseq as part of their STAT-581 course. Each class consisted of a lecture on an RNAseq topic and hands-on training on implementing an analysis on the CGC. We used data in the Sequence Read Archive, providing a real world example for the students to follow. All students were able to accomplish the training in each class. Here is some of the written feedback we received:



Dr. Min Zhang
Professor, Indiana University

"It has been an honor and pleasure working with your group, and the participants are excited about using CGC for their research projects."



Naldyanathan Mahaganapathy, PhD
Graduate Student at Rutgers University

"Without the CGC, this would have been impossible."

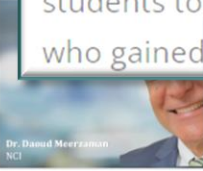
"I cannot explain the level of support I would get [from the Velsera staff]. They got me through the last moments of my PhD."



Dr. Jeremy Chien
Professor, UC Davis

"Just do it."

"...I'm really grateful that that, you know I got the chance to work with you guys in creating this course. So, if you are thinking about it, just do it."



Dr. David Meerzaman
NCI

"Over 700 analytical tools w CGC platform that can be u non-bioinformaticians. This very little to no codin

If of course you're a bioinf savvy, you have that choice to the code but most people I k want to deal with tha

CRDC: NCI Cloud Resources

Training Tomorrow's Data Scientists

CRDC Insights

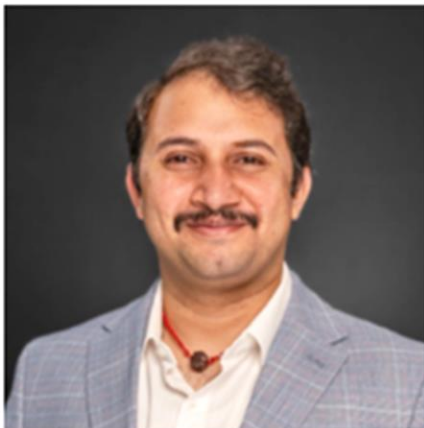
Updates from the Cancer Research Data Commons:
Empowering the Scientific Community to Make New Discoveries

Sign up for the newsletter *

Sign up

Vaidhyanathan Mahaganapathy, Ph.D., Reflects on Support Received from NCI Cloud Resources

October 14, 2024



Vaidhyanathan (Vaidhy) Mahaganapathy, Ph.D., is a Computational Biologist focused on genomics research for the Applied AI team at the [Ellison Institute of Technology \(EIT\)](#) [↗](#). He recently spoke with CRDC Insights, reflecting on the support he received during his dissertation research from the team at [Seven Bridges-Cancer Genomic Cloud \(SB-CGC\)](#) [↗](#), powered by Velsera. He also commented on the complexity of working with huge datasets and trends in cancer research data science, and offered advice to graduate students looking to start a research career.

[Dr. Mahaganapathy's dissertation](#) [↗](#) focused on detecting clonal hematopoiesis (CH), a process prevalent in older people, which has an impact on immune function and inflammation, and is related to many cancers and diseases of aging. For his dissertation, he analyzed 17,000 whole exome sequences in the TCGA dataset, housed through the CRDC, to look across thousands of subjects for mutations

consistent with CH.

Other Features from CRDC Insights

[CRDC 2024 Fall Symposium - Oct 16 & 17 - Registration is Open](#)

[The CRDC's New Data Submission Portal](#)

[Recommendations from NCI CRDC AI Data Readiness Challenge Winners](#)

[Vaidhyanathan Mahaganapathy, Ph.D., Reflects on Support Received from NCI Cloud Resources](#)

[The National Cancer Institute \(NCI\) Medical Image De-Identification Benchmark](#)

CRDC: Internal Interoperability Projects

Challenge: Access comprehensive datasets like TCGA/CPTAC from multiple commons for integrative analysis

- Use common standards

Data Standards Services (DSS)

- Semantic harmonization and shared data model across CRDC
- Leverage existing standards

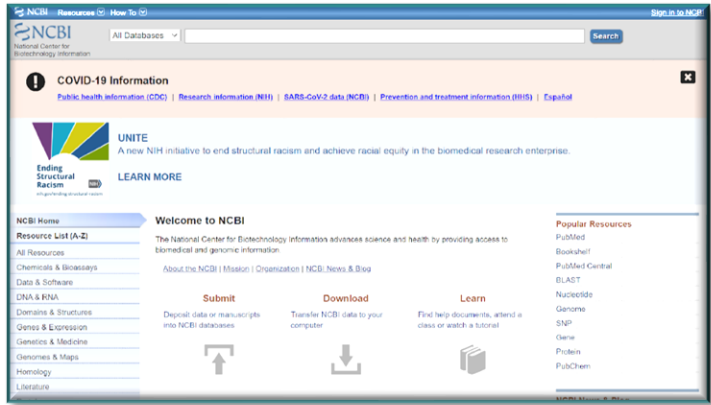
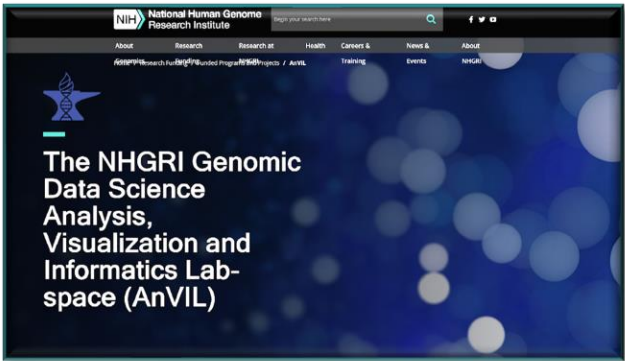
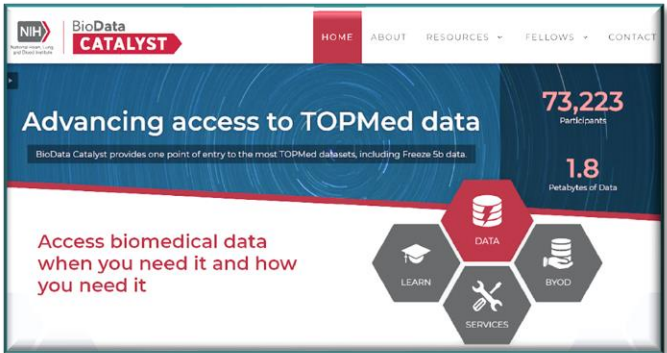
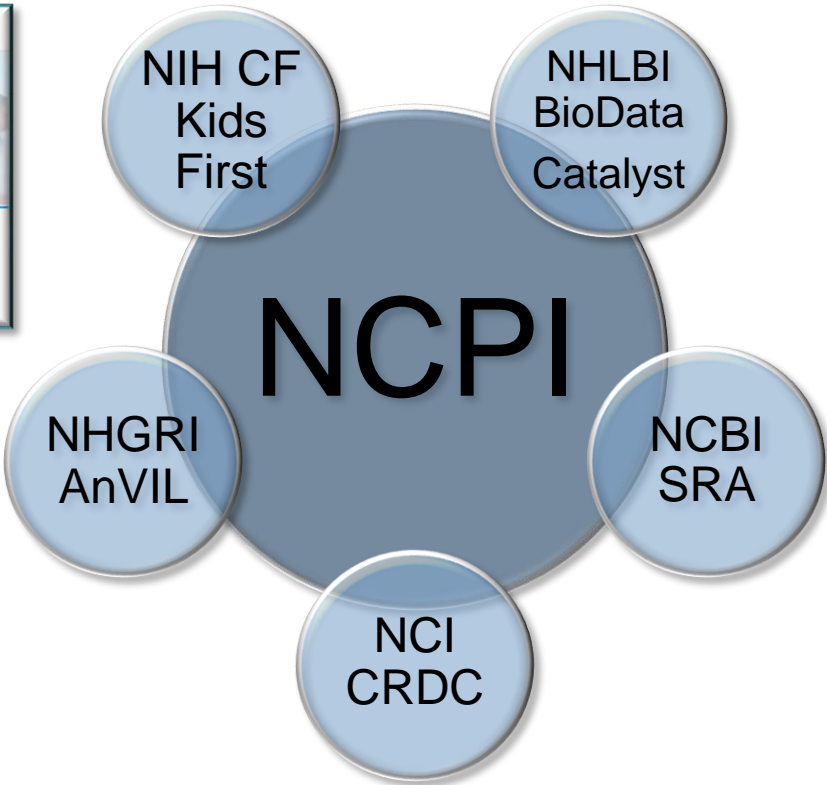
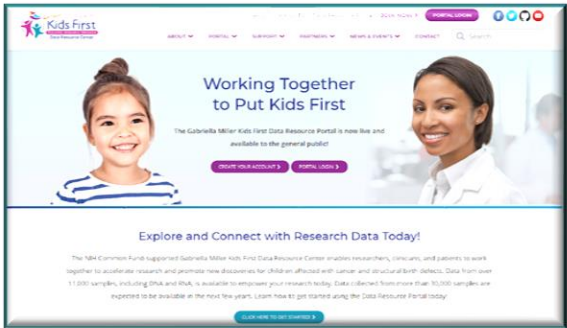
Cancer Data Aggregator (CDA)

- API layer for query/aggregation of data cross-CRDC
- Database for biospecimen, clinical, and phenotypic metadata from CRDC datasets



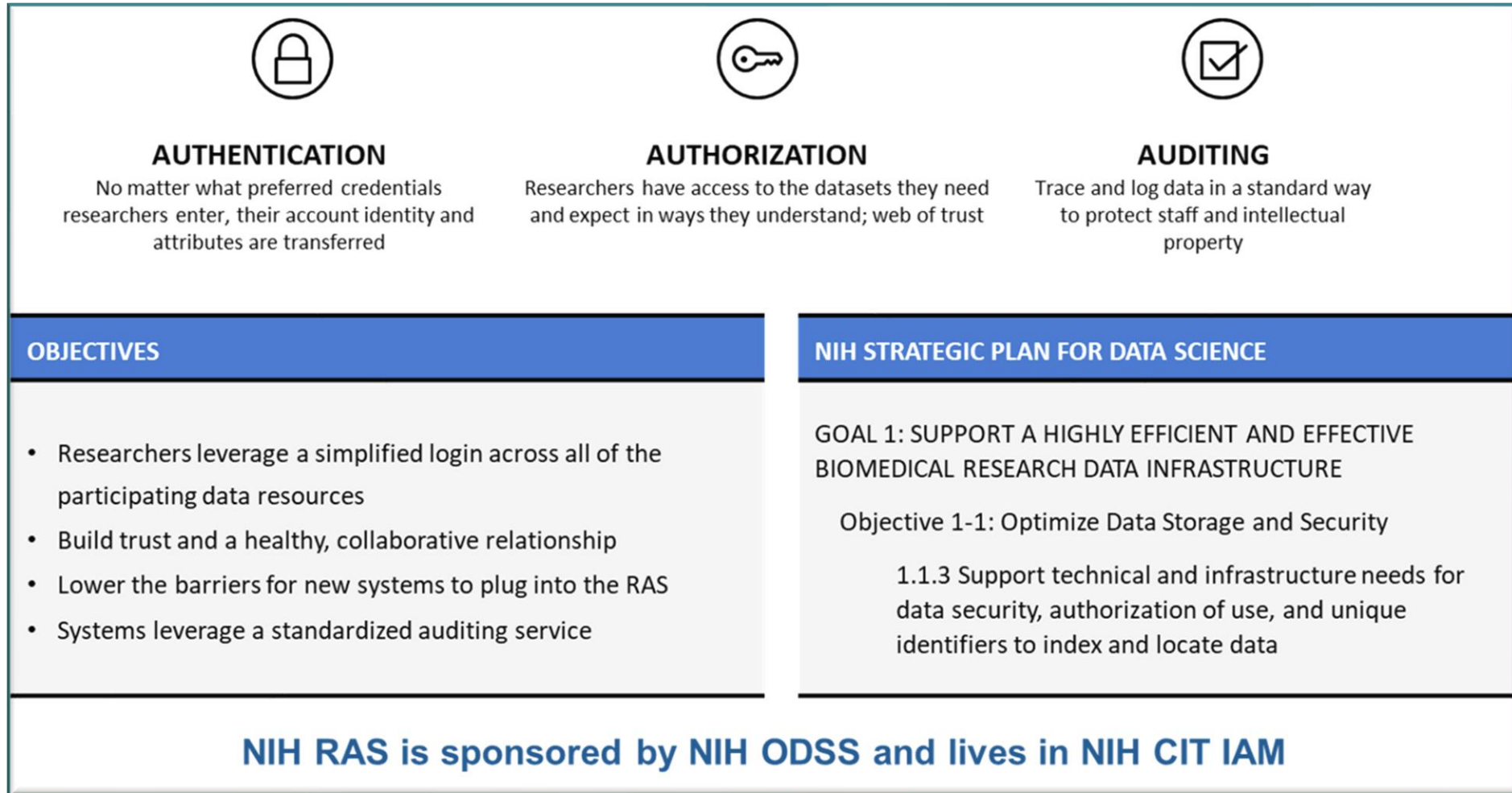
NCPI: NIH Cloud Platforms for Interoperability


Connecting with a Greater Data Ecosystem



RAS: NIH Researcher Auth Services

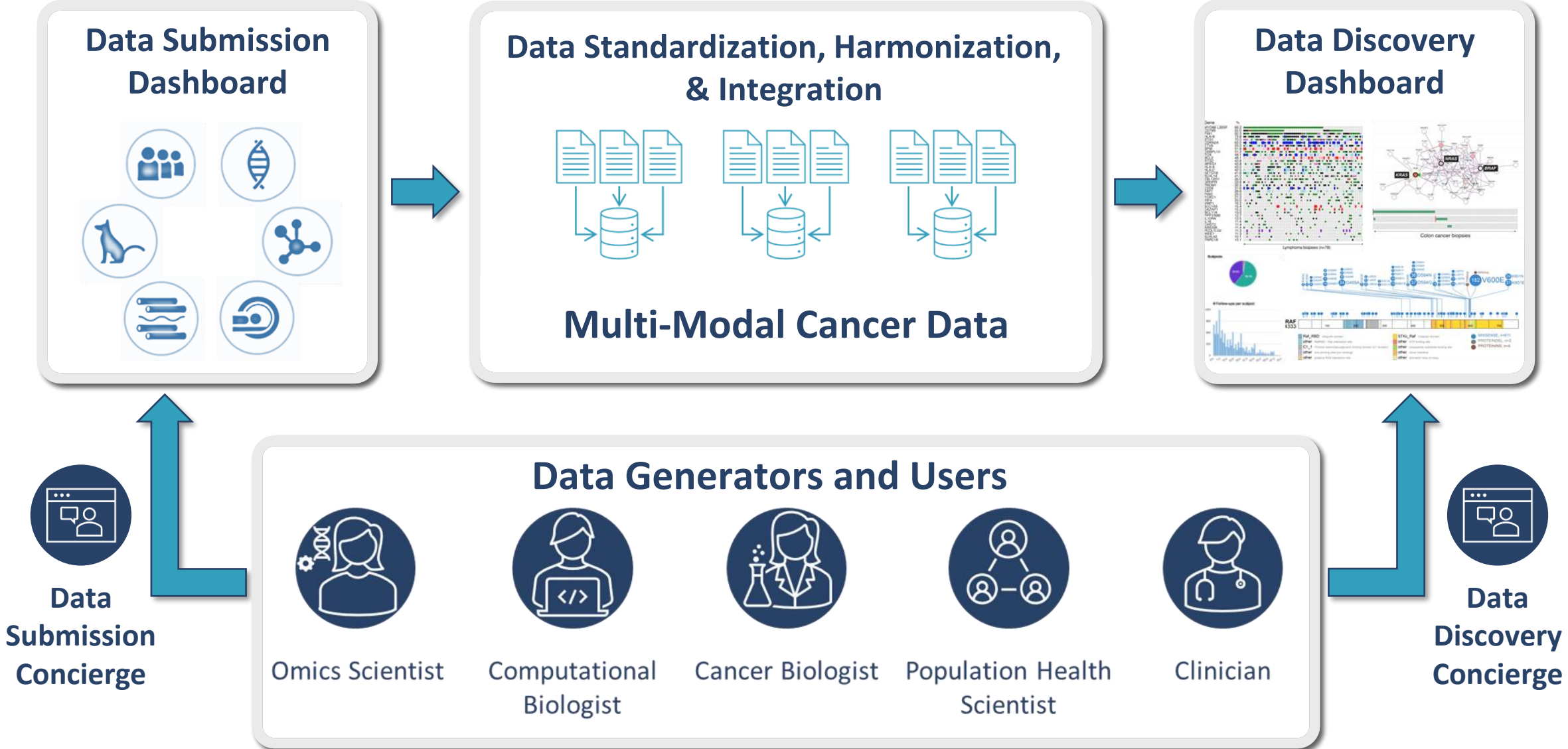
Towards Single “Sign on” Across NIH Data Resources





CRDC Ecosystem: Future State

CRDC Data Portals



CRDC Data Portal: Centralized Data Submission

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons
Data Hub

Home Submission Requests Data Submissions CRDC Data Commons About

Submission Request Form

The following set of high-level questions are intended to provide insight to the CRDC Data Hub, related to data storage, access, secondary sharing needs and other requirements of data submitters.

Status: IN PROGRESS Last updated: 8/8/2023

Data Types

DATA DELIVERY AND RELEASE DATES

Targeted Data Submission Delivery Date 08/07/2023 Expected Publication Date 08/07/2023

DATA TYPES

Indicate the major types of data included in this submission. For each type listed, select Yes or No. Describe any additional major types of data in Other (specify)

Clinical Trial Immunology
 Genomics Proteomics

Principal Investigator and Contact
Program & Study Registration
Data Access and Disease
Data Types
Review and Submit

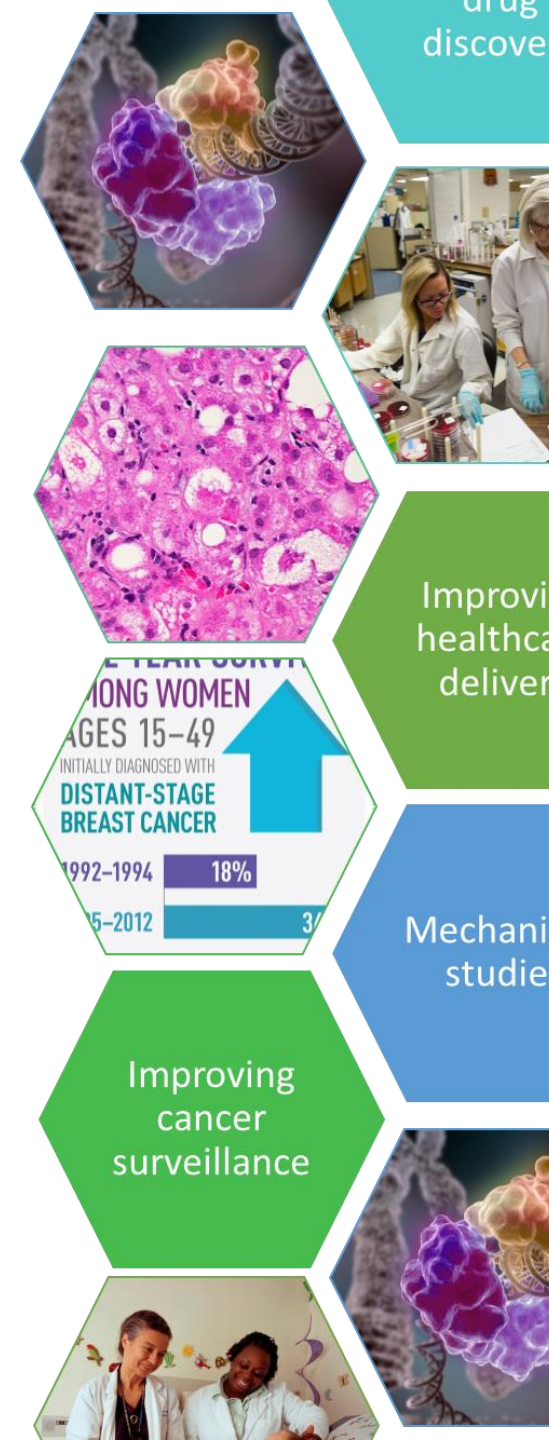
- **Step 1 - Submission Request:** All CRDC Data Commons accepting submission requests through the new Submission Portal
- **Step 2 – Submit Data:**
 - **Currently Accepting Data through the Portal:** Cancer Data Service, Clinical and Translational DC, Integrated Canine DC
 - **Future Integration:** Genomic DC, Proteomic DC, Imaging DC



**Data Submission
Concierge**

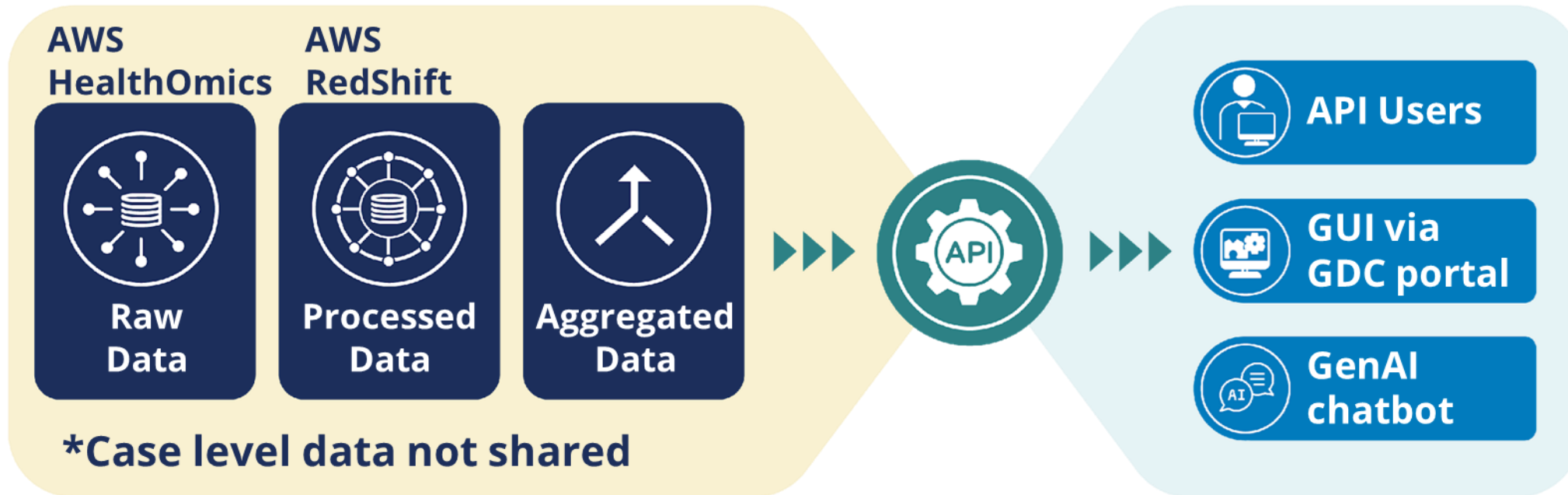
CRDC for Enabling AI in Cancer Research

- **AI in Imaging**
 - AI used in automated de-identification of images
 - Model Hub AI repository
- **AI Data Readiness (AIDR)**
 - Expands FAIR principles to make data accessible for use in AI future applications
 - Request for Information
 - Community Challenge



NCI Genomic Enclave

Data Donation from Tempus: 3300 cases of genomic profiling data, DNA and RNA sequencing



ARPA-H Biomedical Data Fabric (BDF) Toolbox

In collaboration with the National Cancer Institute, ARPA-H will advance the next-generation of tools to synthesize and speed use of health research data, **starting with cancer.**



**Make
biomedical
research data
easier to use**



**Reduce effort
for data
integration**



**Develop new
data fabric
capabilities &
tools**



**Build health science
models to realize the
potential of
the Cancer Data
Ecosystem**

Acknowledgements

- NCI's CRDC Program & Partners:
 - CBIIT Informatics and Data Science (IDS) Program
 - Data Ecosystems Branch (DEB)
 - Clinical & Translational Research Informatics Branch (CTRIB)
 - Computational Genomics & Bioinformatics Branch (CGBB)
 - Center for Cancer Genomics (CCG): GDC
 - DCTD, Office of Cancer Clinical Proteomics Research (OCCPR): PDC
 - DCTD, Developmental Therapeutics Program (DTP): ICDC
 - Division of Cancer Control and Population Sciences (DCCPS): PSDC
 - DCTD, Cancer Imaging Program (CIP): The Cancer Imaging Archive
 - DCTD, Cancer Therapy Evaluation Program (CTEP): NCTN Data Archive
- NCI Frederick National Lab Team (FFRDC)
- All CRDC Subcontractor teams
- All Partners throughout NHI/NCI and data contributors

