

# CRDC

## NCI Cancer Research Data Commons

### What is the **NCI CRDC**?

[NCI's Cancer Research Data Commons](#) (CRDC) is a cloud-based data science infrastructure that connects data sets with analytical tools, providing a foundation for the cancer research community to make new scientific discoveries and lower the burden of cancer. The CRDC is one major component of the broader National Cancer Data Ecosystem that supports the [Cancer Moonshot<sup>SM</sup> Blue Ribbon Panel](#) recommendation to enhance data sharing. The CRDC includes the [Genomic Data Commons \(GDC\)](#), the [Proteomic Data commons \(PDC\)](#), the [NCI Cloud Resources \(CR\)](#), and the [Data Commons Framework \(DCF\)](#), and serves as a central location to support public data sharing for NCI-funded programs.

Currently, more than 23 data sets are available to the cancer research community (see page 3). Starting 2020, new data sets from Cancer Moonshot research projects, such as [Human Tumor Atlas Network](#) (HTAN) and [Immuno-oncology Translational Network](#) (IOTN), will be available through the CRDC for public data access.

### What is a **Data Commons**?

Data Commons co-locates data, storage, and computing infrastructure in the cloud with tools for analyzing and sharing data to create an interoperable resource for the research community.

### Goals of the **NCI CRDC**

- Enable the cancer research community to share diverse data types across programs and institutions
- Provide secure access to data
- Facilitate the generation of innovative tools
- Help NCI-funded Data Coordinating Centers sustain and share data publicly
- Build in an open and modular way to make components extendable and reusable
- Adhere to FAIR principles of data stewardship: Findable, Accessible, Interoperable, and Reusable

# What Does the NCI CRDC Include?

## Data Repositories

### Genomic Data Commons (GDC)

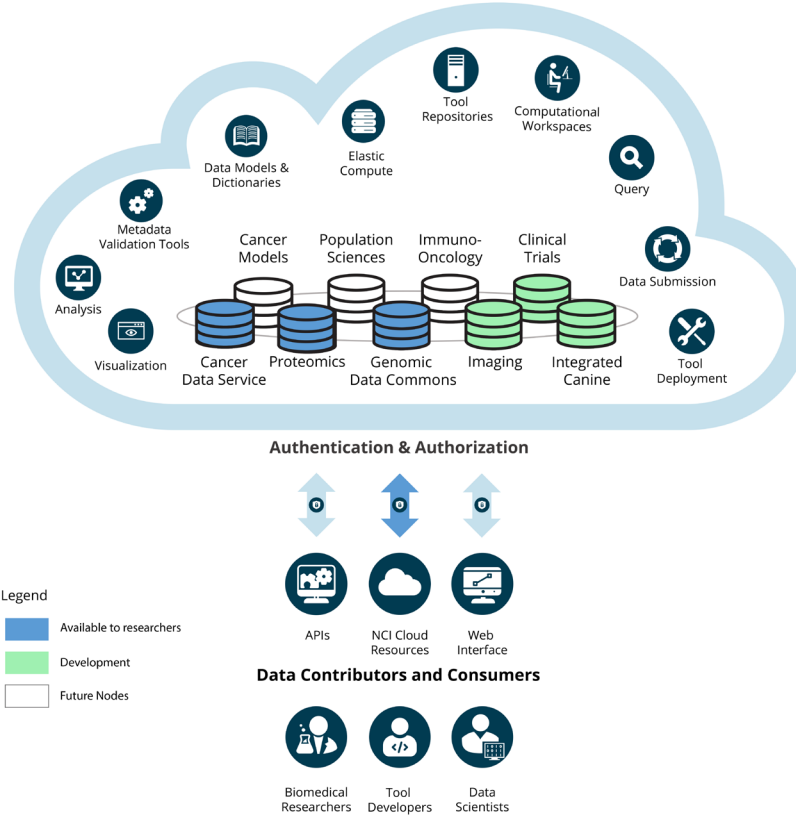
A resource for sharing, analyzing, and visualizing harmonized genomic data, including [TCGA](#), [TARGET](#), and [CPTAC](#).  
<https://gdc.cancer.gov/>

### Proteomic Data Commons (PDC)

A resource for sharing, analyzing, and visualizing proteomic data, such as [CPTAC](#) and [ICPC](#), with new data sets to be included other over time.  
<https://pdc.esacinc.com/pdc/>

### Imaging Data Commons (IDC)

A resource for sharing, analyzing, and visualizing multi-modal imaging data from both clinical and basic cancer research studies. The resource is expected to launch in mid 2020.



### Cancer Data Service (CDS)

A resource for sharing NCI-funded data that is currently not hosted by other repositories. The prototype launched in 2018.

### Integrated Canine Data Commons (ICDC)

A resource for sharing data from canine clinical trials including [PRE-medical Cancer Immunotherapy Network Canine Trials](#) (PRECINCT) and [Comparative Oncology Program](#). The resource is expected to launch in early 2020.

### Clinical Trial Data Commons

A resource to make data from NCI Clinical Trials, such as [Molecular Analysis for Therapy Choice](#) (MATCH) Trial publicly available. The resource is expected to launch in early 2020.

## Infrastructure

### Data Commons Framework (DCF)

Provides secure user Authentication and Authorization and permanent digital object identifiers for data objects.  
<https://dcf.gen3.org/>

### Center for Cancer Data Harmonization (CCDH)

Provides semantic services and tools that facilitate interoperability of the data across the CRDC. The CCDH was launched in 2019 and will work with the repositories listed above.

### Cancer Data Aggregator (CDA)

Enables users to query and connect data distributed across CRDC for integrative analysis. The CDA is expected to launch in August 2020.

### NCI Cloud Resources (CR)

Provides access to cancer data sets to perform large scale analysis using the elastic compute of commercial cloud platforms. The three resources include: **Seven Bridges Cancer Genomics Cloud** (<http://www.cancer-genomics-cloud.org/>) **Institute for Systems Biology Cancer Genomics Cloud** (<http://isb-cgc.org/>) **Broad Institute FireCloud** (<http://firecloud.terra.bio/#>)

# Data and Tools

## What data sets are available in the NCI CRDC?

A variety of data sets are available in the NCI CRDC. Users can bring their own data to combine with the existing data to perform novel analyses through the NCI Cloud Resources.

Data sets	Description	Cancer Types	Cases
<b>The Cancer Genome Atlas (TCGA)</b>	TCGA is a landmark cancer genomics program that molecularly characterized more than 20,000 primary cancer and their matched normal samples spanning 33 cancer types. <a href="https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga">https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga</a>	33	11315
<b>Therapeutically Applicable Research to Generate Effective Treatments (TARGET)</b>	The TARGET program applies a comprehensive genomic approach to determine molecular changes that drive childhood cancers. The goal is to use data to guide the development of effective, less toxic therapies. <a href="https://ocg.cancer.gov/programs/target">https://ocg.cancer.gov/programs/target</a>	5	5062
<b>Clinical Proteomic Tumor Analysis Consortium (CPTAC)</b>	CPTAC integrates proteogenomic and imaging data from 11+ cancer types to identify proteomic-centric subtypes and improve understanding of cancer-relevant pathways through posttranslational modifications. <a href="https://proteomics.cancer.gov/programs/cptac">https://proteomics.cancer.gov/programs/cptac</a>	11 (as of 2019)	2208 (as of 2019)

## What tools and resources are available in the NCI CRDC?

The NCI CRDC makes a wide range of popular and innovative analytic and visualization tools available for all stages of data analysis. Users can also bring custom tools to the NCI Cloud Resources.

The three NCI Cloud Resources provide search, compute, and analytical tools, with more than 1000 tools and workflows including RStudio and Jupyter notebooks.

	Resources	Access	Analysis Tools and Workflows							Data Visualization
		Portal & Workspace	Variant Calling	Mapping	Transcriptome	Epigenomic	Imaging	Proteomic	Multi-omic	
Data Repository	GDC	x	x	x						x
	PDC	x		x				x		x
	IDC	P					P			P
Cloud Resources	SB-CGC	x	x	x	x	x	x	x	x	x
	FireCloud	x	x	x	x	x	x	x	x	x
	ISB-CGC	x	x	x	x	x	x	x	x	x

Currently available features are represented by 'X' and features that are planned are indicated by 'P'.

# Submitting Data to the NCI CRDC And Other NIH Repositories

The NCI CRDC provides repositories to store diverse data types. Currently, data submission requests are made by completing an application at each target repository's website or by contacting repository's helpdesk.  
*Table summarizing target repositories for respective data types as of 11/2019.*

Data Domains	Subcategory	Target Repository*	Submit
Genomic	Non-harmonized genomic data	CDS	<a href="mailto:CDSHelpDesk@nih.gov">CDSHelpDesk@nih.gov</a>
	Harmonized genomic data (e.g., tumor-normal pair)	GDC	<a href="https://gdc.cancer.gov/data-submission-request-form">https://gdc.cancer.gov/data-submission-request-form</a>
Proteomic	Mass Spectrometry	PDC	<a href="https://pdc.esacinc.com/pdc/about">https://pdc.esacinc.com/pdc/about</a>
	Other (e.g., Reverse-phase protein arrays)	CDS	<a href="mailto:CDSHelpDesk@nih.gov">CDSHelpDesk@nih.gov</a>
Imaging	Radiology	TCIA** (IDC in 2020)	<a href="https://www.cancerimagingarchive.net/primary-data/">https://www.cancerimagingarchive.net/primary-data/</a>
	Digital Pathology	TCIA** (IDC in 2020)	<a href="https://www.cancerimagingarchive.net/primary-data/">https://www.cancerimagingarchive.net/primary-data/</a>
	Other (e.g., Cellular imaging)	CDS (IDC in 2021)	<a href="mailto:CDSHelpDesk@nih.gov">CDSHelpDesk@nih.gov</a>
Clinical	Demographics, phenotypes, etc.	NCBI dbGaP CRDC as appropriate	<a href="https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/">https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/</a> <a href="mailto:ncicrdc@mail.nih.gov">ncicrdc@mail.nih.gov</a>
Other	Flow cytometry, videos, etc.	Figshare (up to 20 GB)	<a href="https://nih.figshare.com/f/about">https://nih.figshare.com/f/about</a>

\*CRDC is under active development and target repositories may change as new components become available.  
 Check the [NCI CRDC website](#) for the latest information.

\*\* All clinical imaging data will need to be de-identified through The Cancer Imaging Archive (TCIA) before storing on the IDC.

## Data Sharing Policy Resources

NCI has policies and guidelines regarding sharing NCI-funded data including NCI's Genomic Data Sharing (GDS) Policy. NCI's Office of Data Sharing oversees and coordinates the sharing of NCI-funded data.

### Learn more about

NCI GDS Policy Page: <https://www.cancer.gov/grants-training/grants-management/nci-policies/genomic-data>

GDS FAQs: <https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing-faqs/>

Moonshot supplemental guideline: <https://datascience.cancer.gov/data-sharing/policies>

Last updated 11-05-19

page 4



NATIONAL  
CANCER  
INSTITUTE