NIH NATIONAL CANCER INSTITUTE

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 to provide a foundation for the cancer research community to make new scientific discoveries and lower the burden of cancer. As a major component of the broader National Cancer Data Ecosystem supporting the Cancer Moonshot (SM) Blue Ribbon Panel's Recommendation to enhance data sharing, NCI CRDC serves as a coordinated resource for public data sharing of NCI-funded programs.

Starting in 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 NCI CRDC for public data access.

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

Data and Tools

Datasets Currently Available Through 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. There are 23+ datasets available.

- The Cancer Genome Atlas (TCGA)
- Therapeutically Applicable Research to Generate Effective Treatments (TARGET)
- Cancer Cell Line Encyclopedia (CCLE)
- Clinical Proteomic Tumor Analysis (CPTAC)

The NCI CRDC makes a wide range of analytic and visualization tools available for all stages of data analysis. Users can also bring custom tools to the NCI Cloud Resources. Currently over 1000 tools, workflows, and applications are available, including RStudio and Jupyter Notebooks.

Tools	GDC	PDC	IDC	SB-CGC	ISB-CGC	FireCloud
Portal & Workspace	X	X	P	X	X	X
Variant Calling	X			X	X	X
Reference Mapping	X	X		X	X	X
Transcriptomic				X	X	X
Epigenomic				X	X	X
Imaging			P	X	X	X
Proteomic		X		X	X	X
Multi-omic				X	X	X
Data Visualization	X	X	P	X	X	X

Currently available features are represented by 'X' and features that are planned are indicated by 'P'.
Please see page two for acronym definitions.

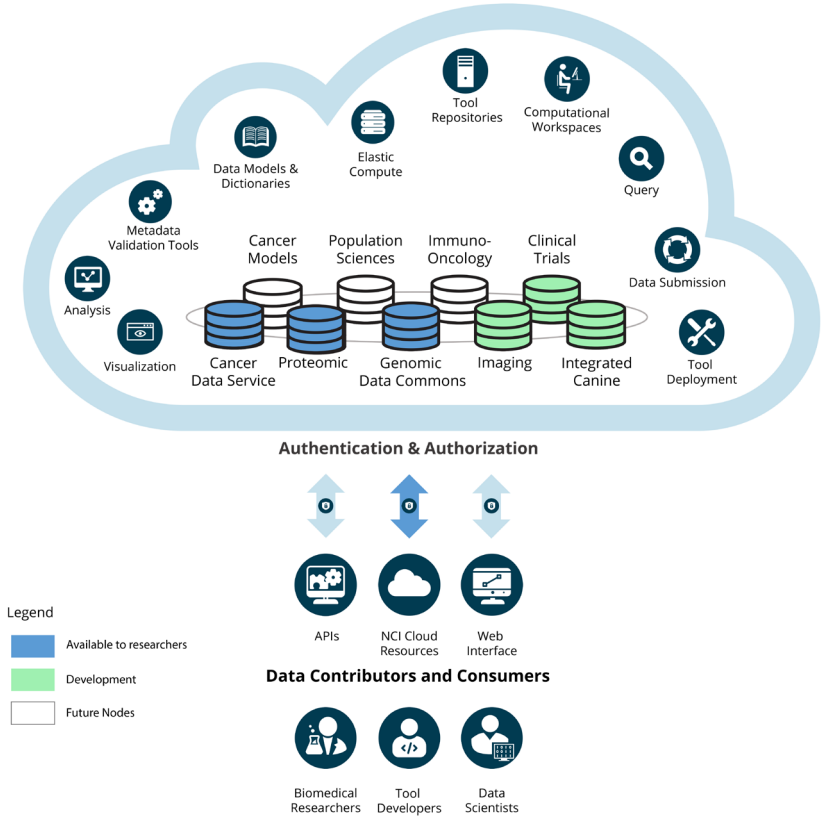
What Does the NCI CRDC Include?

Data Repositories

Genomic Data Commons (GDC)
Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

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

Imaging Data Commons (IDC)
Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies. The resource is expected to launch in 2020.



Cancer Data Service (CDS)
Share NCI-funded data that is currently not hosted by other repositories.

Integrated Canine Data Commons (ICDC)
Share 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 2020.

Clinical Trial Data Commons (CTDC)
Stores data from NCI Clinical Trials, such as Molecular Analysis for Therapy Choice (MATCH) trial publicly available. The resource is expected to launch in 2020.

Infrastructure and Services

Data Commons Framework (DCF)
Provides secure user authentication and authorization and permanent digital object identifiers for data objects.

Center for Cancer Data Harmonization (CCDH)
Provides semantic services and tools that facilitate interoperability of the data across the NCI CRDC.

Cancer Data Aggregator (CDA)
Enables users to query and connect data distributed across NCI CRDC for integrative analysis. The CDA is expected to launch in 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 CGC (SB-CGC)** (http://www.cancergenomicscloud.org/)
- **Institute for Systems Biology CGC (ISB-CGC)** Genomics Cloud (http://isb-cgc.org/)
- **Broad Institute FireCloud** (http://firecloud.terra.bio/#)

Submit Data to the NCI CRDC

https://datascience.cancer.gov/submitcrdc