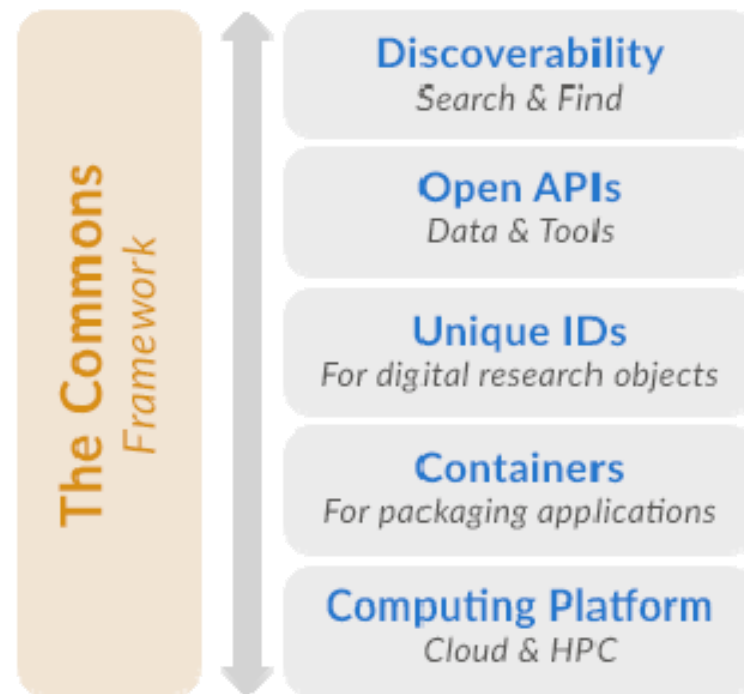


# Data Commons Framework



# Cancer Research Data Ecosystem – Cancer Moonshot BRP

Discovery

Proteogenomics  
Imaging data  
Clinical trials

Well characterized  
research data sets

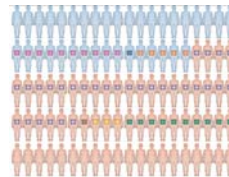


**GDC**  
Research information  
donor

Patient engaged  
Research

Clinical Research  
Observational studies

Cancer cohorts



Active research  
participation

Surveillance  
Big Data  
Implementation research

EHR, Lab Data, Imaging,  
PROs, Smart Devices,  
Decision Support

Patient data



**SEER**  
Learning from every  
cancer patient

# Genomic Data Commons

The Cancer Genomic Data Commons (**GDC**) is an existing effort to standardize and simplify submission of genomic data to NCI and follow the principles of **FAIR** – Findable, Accessible, Attributable, Interoperable, Reusable, and Provide Recognition.

The GDC is part of the NIH Big Data to Knowledge (**BD2K**) initiative and an example of the **NIH Commons**

*Microattribution, nanopublications (DOIs), tracking the use of data, annotation of data, use of algorithms, supports the data /software /metadata life cycle to provide credit and analyze impact of data, software, analytics, algorithm, curation and knowledge sharing*

Force11 white paper

<https://www.force11.org/group/fairgroup/fairprinciples> 3

# GDC overview

The screenshot shows the GDC Data Portal interface. At the top, there's a navigation bar with 'Home', 'Projects', 'Data', and 'Analysis' links. A search bar and 'Login' button are also present. The main content area features a 'Harmonized Cancer Datasets' section with 'Genomic Data Commons Data Portal' and buttons for 'Projects' and 'Data'. Below this, there are search examples for kidney cancer, brain cancer, and TCGA-GBM project. A bar chart titled 'Cases by Primary Site' shows the distribution of cases across various cancer types. At the bottom, there are summary statistics for Projects (39), Primary Site (29), Cases (14,531), and Files (274,821). The footer contains sections for Infrastructure, Documentation, and GDC Applications.

**Harmonized Cancer Datasets**  
**Genomic Data Commons Data Portal**

Get Started by Exploring:

- [Projects](#)
- [Data](#)

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	182 Cases	1,514 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files

**Cases by Primary Site**

Primary Site	Approximate Cases
Kidney	1,500
Brain	1,100
Nervous System	1,100
Breast	1,100
Lung	1,100
Blood	900
Colorectal	600
Uterus	600
Ovary	600
Head and Neck	500
Thyroid	500
Prostate	500
Stomach	400
Skin	400
Bladder	400
Bone	400
Liver	300
Cervix	300
Adrenal Gland	300
Soft Tissue	300
Bone Marrow	200
Pancreas	200
Esophagus	200
Testis	200
Thymus	200
Pleura	100
Eye	100
Lymph Nodes	100
Bile Duct	100

**DATA PORTAL SUMMARY**  
 Data Release 4.0 - October 31, 2016

<b>PROJECTS</b>	<b>PRIMARY SITE</b>	<b>CASES</b>	<b>FILES</b>
39	29	14,531	274,821

**Infrastructure**  
 Data is continuously being processed and harmonized by the GDC.  
 View GDC system statistics:

<b>Compute Infrastructure</b>	12,800 Cores	87.96 TB RAM
<b>Storage Infrastructure</b>	4.98 PB Used	5.42 PB Total

[View Data Download Statistics Report »](#)

**Documentation**  
 Learn how to use the GDC Data Portal to its full potential with common topics such as:

- Browse Data using Facet Search
- Search Data with Advanced Search Technology
- Project Based Data Availability
- Controlled Access Data
- [Visit the Documentation Website »](#)

**GDC Applications**  
 The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

- Data Portal
- Website
- Data Transfer Tool
- API
- Data Submission Portal
- Documentation
- Legacy Archive
- GDC cBio Portal

https://gdc-docs.nci.nih.gov

## NCI Genomic Data Commons

- **The GDC went live on June 6, 2016 with approximately 4.1 PB of data.**
- 577,878 files about 14194 cases (patients), in 42 cancer types, across 29 primary disease sites, 400 clinical data elements
- 10 major data types, ranging from Raw Sequencing Data, Raw Microarray Data, to Copy Number Variation, Simple Nucleotide Variation and Gene Expression.
- Data are derived from 17 different experimental strategies, with the major ones being RNA-Seq, WXS, WGS, miRNA-Seq, Genotyping Array and Expression Array.
- **Foundation Medicine announced the release of 18,000 genomic profiles to the GDC at the Cancer Moonshot Summit, June 29<sup>th</sup>, 2016**
- **The Multiple Myeloma Research Foundation announced it would be releasing its CoMMpass study of more than 1000 cases of Multiple Myeloma on Sept 29, 2016.**

## GDC Monthly Usage

URL	Month	Unique Visitor	Number of Visits	Pages	Hits	Data Volume over F5
gdc.cancer.gov	December 2016	9,383	16,363	70,690	617,051	28.29GB
gdc-api.nci.nih.gov		11,643	26,220	2,284,424	2,284,425	3667.11GB
gdc-portal.nci.nih.gov		11,065	21,541	473,283	618,056	8.11GB
docs.gdc.cancer.gov		3,566	5,519	50,592	301,019	26GB
cbioportal.gdc.cancer.gov		1,791	2,289	96,937	373,021	8.96GB

## GDC storage

Solution	Size	Available	Used	% Used
IBM COS	8.75PiB	2.26PiB	6.49PiB	74%
CephB	760TB	223TB	536.7TB	71%

## GDC data download statistics (since launch)

Disease Type	# Requests	File Size
Breast Invasive Carcinoma	4,630,548	289.76 TB
Glioblastoma Multiforme	1,820,418	267.80 TB
Kidney Renal Clear Cell Carcinoma	1,369,223	240.78 TB
Kidney Renal Papillary Cell Carcinoma	604,056	136.60 TB
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	546,272	76.91 TB
Rectum Adenocarcinoma	363,502	73.09 TB
Sarcoma	335,382	23.18 TB
Pancreatic Adenocarcinoma	335,078	20.12 TB
Esophageal Carcinoma	247,170	38.75 TB
Adrenocortical Carcinoma	218,321	3.14 TB
Kidney Chromophobe	119,718	62.34 TB
Mesothelioma	89,720	819.94 GB
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	79,542	17.31 TB
Not Applicable	38,116	211.71 GB
Chronic Lymphocytic Leukemia	5,430	6.47 TB
Multiple Myeloma	1,532	1.59 TB
Neuroblastoma	441	6.08 TB
Rhabdoid Tumor	300	294.89 GB



## Foundation Medicine, Inc. (FMI)

FMI agreed to donate data to GDC (press release June 29, 2016)

### Data

- 18,004 cases
- Clinical data: age, gender and diagnosis
- Genomic data: mutation calls (no raw sequence data)

### Current status

- All data have been transferred to GDC
- GDC finished lift-over from HG19 to HG38 so genomic coordinates are compatible with other genomic data in GDC
- Final phase of data QC



## Multiple Myeloma Research Foundation

MMRF agreed to donate data to GDC (Press release September 28, 2016)

GDC and MMRF are currently working together on

- Clinical data elements mapping and harmonization
- Biospecimen metadata ETL (Excel format to XML)
- Genomic data migration testing
  - Sample DNaseq BAM submission and re-alignment successful
  - Sample RNAseq FASTQ submission and re-alignment successful