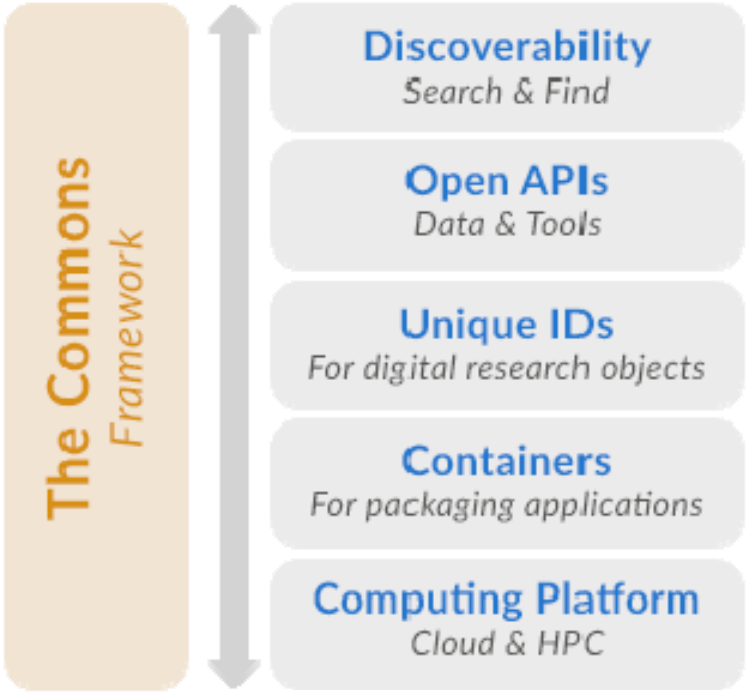
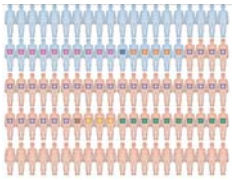
Data Commons Framework



Cancer Research Data Ecosystem – Cancer Moonshot BRP

Discovery	Patient engaged Research	Surveillance Big Data Implementation research
Proteogenomics Imaging data Clinical trials	Clinical Research Observational studies	EHR, Lab Data, Imaging, PROs, Smart Devices, Decision Support
Well characterized research data sets	Cancer cohorts	Patient data
 **GDC** Research information donor	 Active research participation	 **SEER** Learning from every cancer patient

Genomic Data Commons

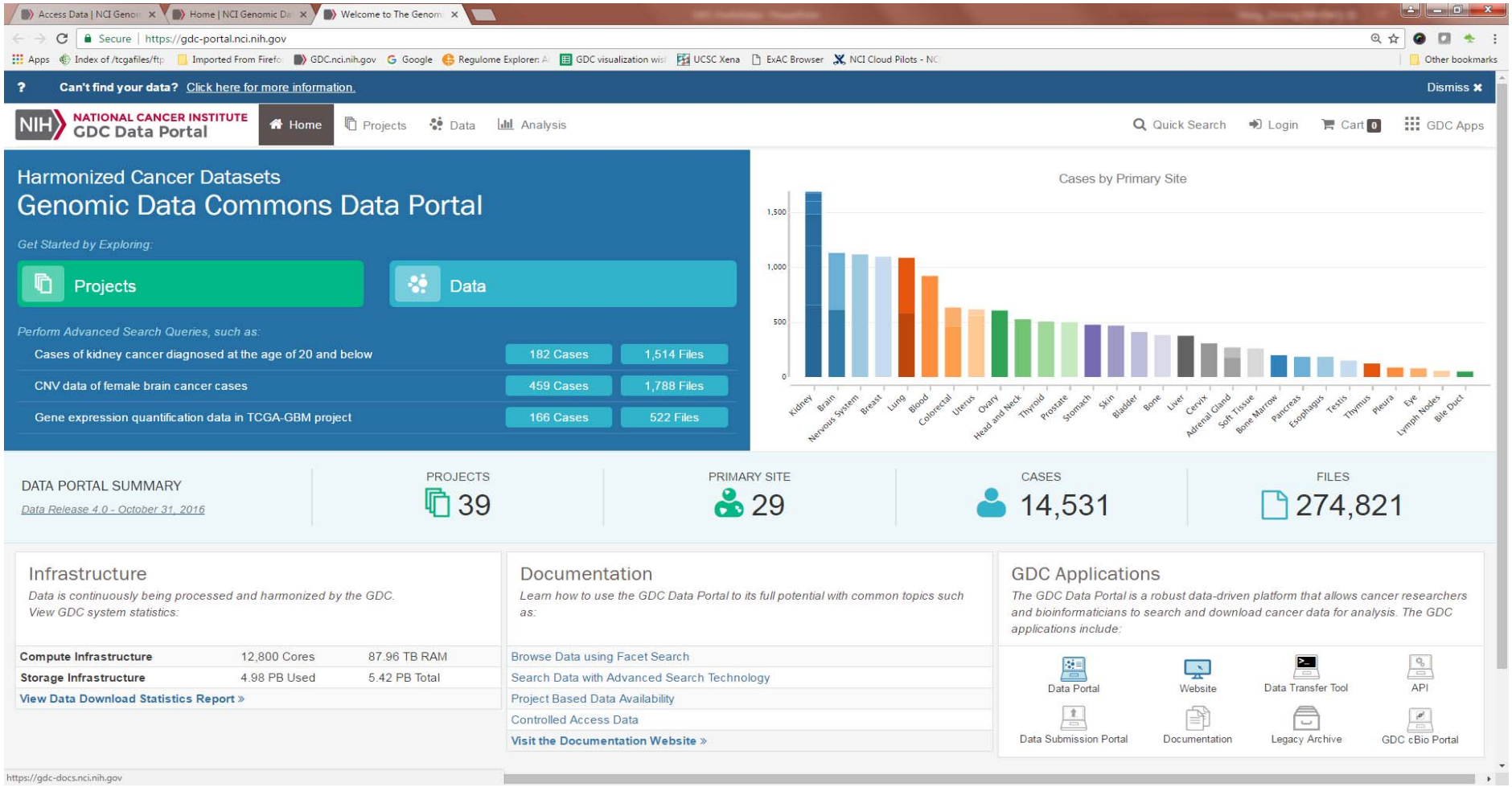
The Cancer Genomic Data Commons (**GDC**) is an existing effort to standardize and simplify submission of genomic data to NCI and follow the principles of **FAIR** – Findable, Accessible, Attributable, Interoperable, Reusable, and Provide Recognition.

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

GDC overview



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 29th, 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