

Genomic Data Commons

NCI Cloud Pilots

Louis Staudt, MD, PhD

Warren Kibbe, PhD



@wakibbe

Changing the conversation around data sharing

NIH Data Commons



- How do we find data, software, standards?
- How can we make data, annotations, software, metadata accessible?
- How do we reuse data standards
- How do we make more data machine readable?

*Data commons co-locate data, storage and computing infrastructure, and commonly used tools for analyzing and **sharing data** to create an **interoperable** resource for the research community.*

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson, A Case for Data Commons Towards Data Science as a Service, to appear. Source of image: Interior of one of Google's Data Center, www.google.com/about/datacenters/.



FAIR –
Making data
Findable,
Accessible,
Attributable,
Interoperable,
Reusable,
and provide Recognition

Force11 white paper

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

NIH Genomic Data Sharing Policy

<https://gds.nih.gov/>

Went into effect January 25, 2015

NCI guidance:

<http://www.cancer.gov/grants-training/grants-management/nci-policies/genomic-data>

Requires public sharing of genomic data sets

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, Interoperable, Reusable.

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

Microattribution, nanopublications, 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

Genomic Data Commons

- Unified knowledge base that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology
- Contains standardized data from approximately 14,500 patients, derived from NCI programs, including:
 - The Cancer Genome Atlas (TCGA)
 - Therapeutically Applicable Research to Generate Effective Treatment (TARGET)
 - Cancer Genome Characterization Initiative (CGCI)
 - The Cancer Line Encyclopedia (CCLE)

Genomic Data Commons
went live at ASCO June 6, 2016

The New York Times

Biden Unveiling Public Database for Clinical Data on Cancer

FOX NEWS Health

Biden unveiling public database for clinical data on cancer

HealthData
Management

NCI launches open access resource to spur cancer research

HOUSTON CHRONICLE

Biden unveils searchable government cancer database

SCIENTIFIC AMERICAN **REUTERS**
PUBLIC HEALTH

Biden Unveils Major Database to Advance Cancer Research

THE HUFFINGTON POST
INFORM • INSPIRE • ENTERTAIN • EMPOWER

Biden Announces Crucial Piece Of His Cancer Moonshot Initiative

The Washington Post

Biden unveils launch of major, open-access database to advance cancer research



FORTUNE

Joe Biden Just Announced a Huge New National Cancer Database

Biden announces U.S. project to promote cancer data sharing

REUTERS

CHICAGO SUN-TIMES

VP Joe Biden in Chicago to promote Moonshot Initiative vs. cancer

THE CANCER LETTER

Biden Designates NCI's Genomic Data Commons As Foundation of Cancer Moonshot

Daily **Mail**

New US data system to centralize cancer information

genomeweb

NCI Launches Genomic Data Commons for Cancer Data Sharing

HealthITAnalytics
News and Resources for Healthcare Analytics Pros

NIH Launches Genomic Data Commons Supporting Cancer Moonshot

fedSCOOP

Biden launches data portal to back Cancer Moonshot

Genomic Data Commons (GDC)

was highlighted in the June 29th Cancer Moonshot Summit at Howard University in the US

Foundation Medicine announced the release of 18,000 genomic profiles to the GDC at the Cancer Moonshot Summit

NCI Genomic Data Commons

- The GDC went live with approximately 4.1 PB of data.
- This includes: 2.6 PB of legacy data;
- and 1.5 PB of “harmonized” data.
- 577,878 files about 14194 cases (patients), in 42 cancer types, across 29 primary sites.
- 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.

Genomic Data Commons Data Portal

NATIONAL CANCER INSTITUTE
GDC Data Portal

Projects
Data
Annotations
Reports

Quick Search
Login
Cart 0
GDC Apps

Harmonized Cancer Datasets Genomic Data Commons Data Portal

Get Started by Exploring:

📁 Projects

📄 Data

Perform Advanced Search Queries, such as:

Kidney cancer cases under the age of 20 at diagnosis	128 Cases	1,159 Files
CNV data of female brain cancer cases	459 Cases	7,809 Files
Germline mutation data in TCGA-OV project	423 Cases	1,700 Files

Cases by Primary Site

Primary Site	Approximate Number of Cases
Kidney	1,500
Hematopoietic System	1,400
Bladder	1,300
Breast	1,200
Lung	1,150
Colon	1,100
Rectum	1,050
Stomach	1,000
Head and Neck	950
Thyroid	900
Prostate	850
Esophagus	800
Bladder	750
Leuk	700
Cervix	650
Melanoma	600
Pancreas	550
Endometrium	500
Troch	450
Thyroid	400
Rectum	350
Adipose Tissue	300
Melanoma	250
Esophagus	200
Uterus	150
Prostate	100
Bladder	50

Data Portal Summary

Latest Release #0 - May 2, 2016

PROJECTS

46

PRIMARY SITES

29

CASES

14,194

FILES

594,527

Infrastructure

Data is continuously being processed and harmonized by the GDC. System stats:

Storage Infrastructure	1.73 PB used	7.42 PB available
Compute Infrastructure	13,120 Cores	81,920 GB RAM
Internet Facing Bandwidth	1.2 Gbps in	18.2 Mbps out
Unique Visitors	20 daily	1500 Total
Downloads to Date	355,024	11.94 TB

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

Data Transfer Tool

API

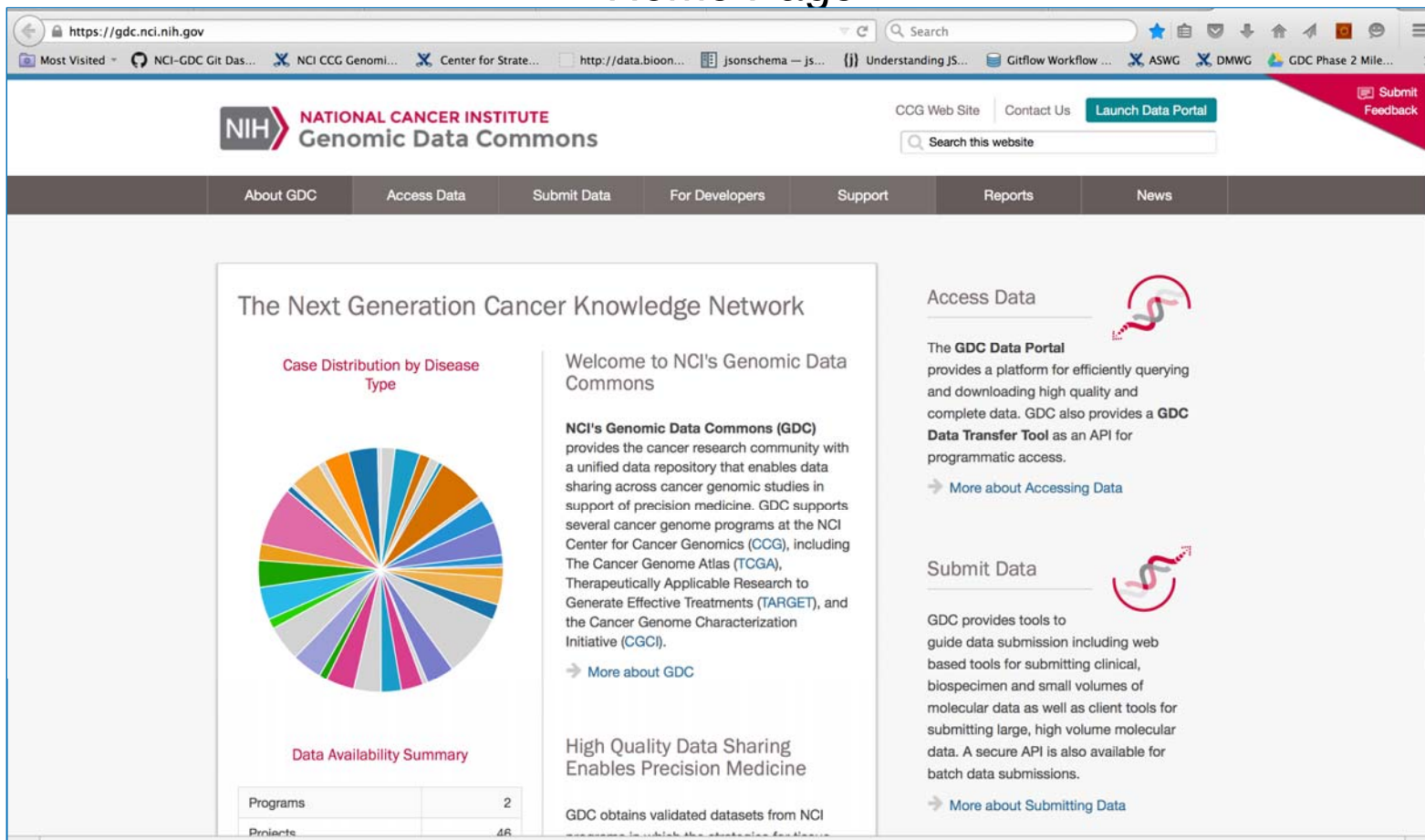
Data Submission Portal

Documentation

Website

Legacy Archive

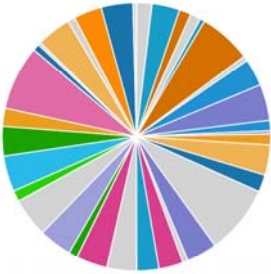
The NCI Genomic Data Commons User Interface Home Page



The screenshot shows the NCI Genomic Data Commons (GDC) home page. The browser address bar displays <https://gdc.nci.nih.gov>. The page header includes the NIH logo, the text "NATIONAL CANCER INSTITUTE Genomic Data Commons", and navigation links for "CCG Web Site", "Contact Us", and "Launch Data Portal". A search bar is also present. A dark navigation bar contains links for "About GDC", "Access Data", "Submit Data", "For Developers", "Support", "Reports", and "News".

The main content area features a large section titled "The Next Generation Cancer Knowledge Network". On the left, there is a "Case Distribution by Disease Type" pie chart and a "Data Availability Summary" table. On the right, there are sections for "Access Data" and "Submit Data", each with a DNA helix icon and a "More about" link.

Case Distribution by Disease Type



Data Availability Summary

Programs	2
Projects	46

Welcome to NCI's Genomic Data Commons

NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine. GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), and the Cancer Genome Characterization Initiative (CGCI).

[→ More about GDC](#)

High Quality Data Sharing Enables Precision Medicine

GDC obtains validated datasets from NCI

Access Data

The **GDC Data Portal** provides a platform for efficiently querying and downloading high quality and complete data. GDC also provides a **GDC Data Transfer Tool** as an API for programmatic access.

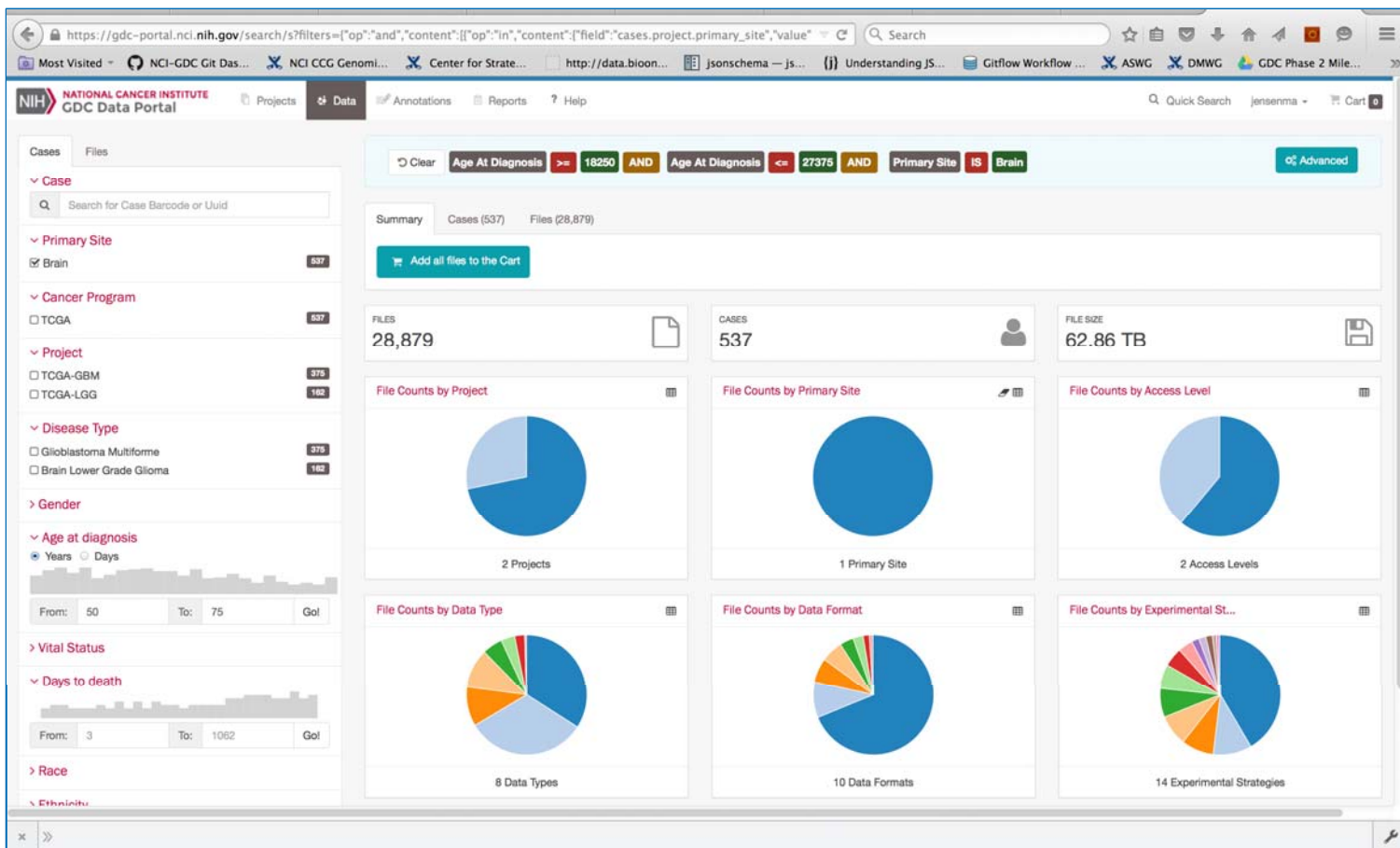
[→ More about Accessing Data](#)

Submit Data

GDC provides tools to guide data submission including web based tools for submitting clinical, biospecimen and small volumes of molecular data as well as client tools for submitting large, high volume molecular data. A secure API is also available for batch data submissions.

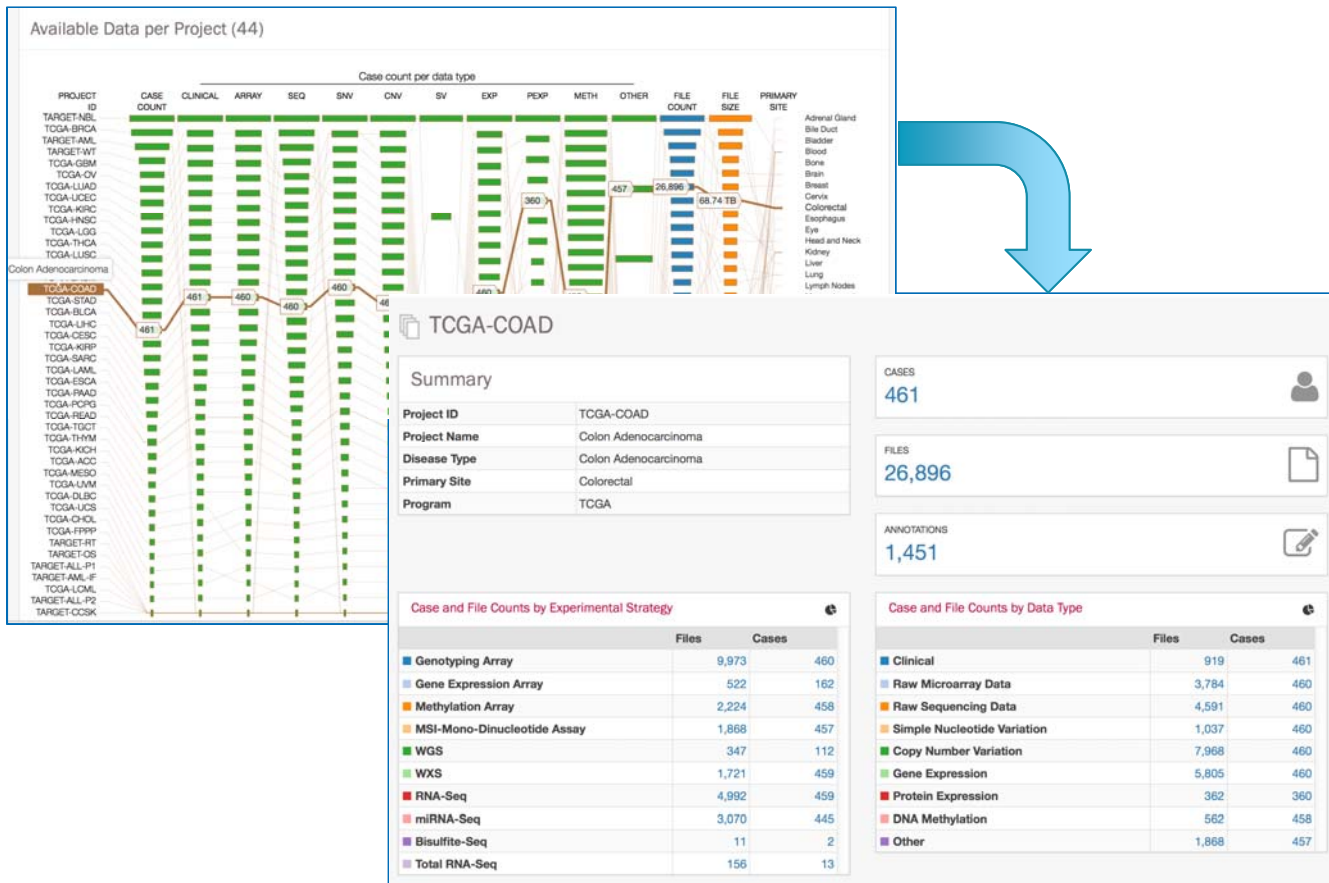
[→ More about Submitting Data](#)

The NCI Genomic Data Commons User Interface Sample Browser



The NCI Genomic Data Commons User Interface

Sample Selection



The NCI Genomic Data Commons User Interface Data Submission Dashboard

NIH NATIONAL CANCER INSTITUTE GDC Data Submission Portal TCGA-DEV3 Search Help VFERRETTI

Home » TCGA DEV3 » Dashboard Dashboard Browse Dictionaries

Clinical data

25
39% of 64 Cases

[DETAILS](#)

Biospecimen data

25
39% of 64 Cases

[DETAILS](#)

Molecular data

2
3% of 64 Cases

[DETAILS](#)

Files uploaded

23
32% of 71 Files

[DETAILS](#)

1. Upload & Validate 2. Submit 3. Release

1. UPLOAD AND VALIDATE DATA

To submit data to GDC, you first need to upload and validate your clinical, biospecimen and experimental data to the project workspace. For additional instructions, please see the [GDC Data Submission Portal User's Guide](#)

A. Data must be compliant with the project

BROWSE OR DOWNLOAD DATA

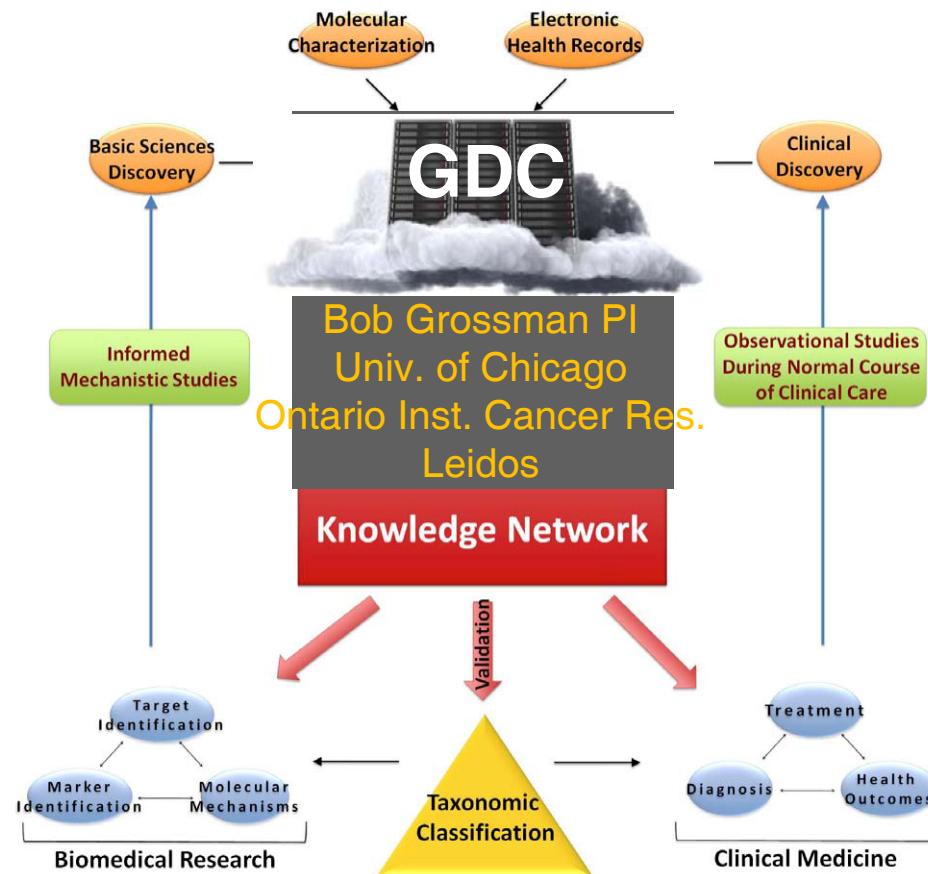
Use the following links to browse or download data stored in the project workspace. Data will be exported in TSV format. These file can be used to:

- Review data before

TRANSACTIONS

ID	Type	Date Created
129	Release	Dec 14, 2015
128	Submission	Dec 14, 2015
127	Submission	Dec 14, 2015
126	Submission	Dec 14, 2015

Development of the NCI Genomic Data Commons (GDC) To Foster the Molecular Diagnosis and Treatment of Cancer

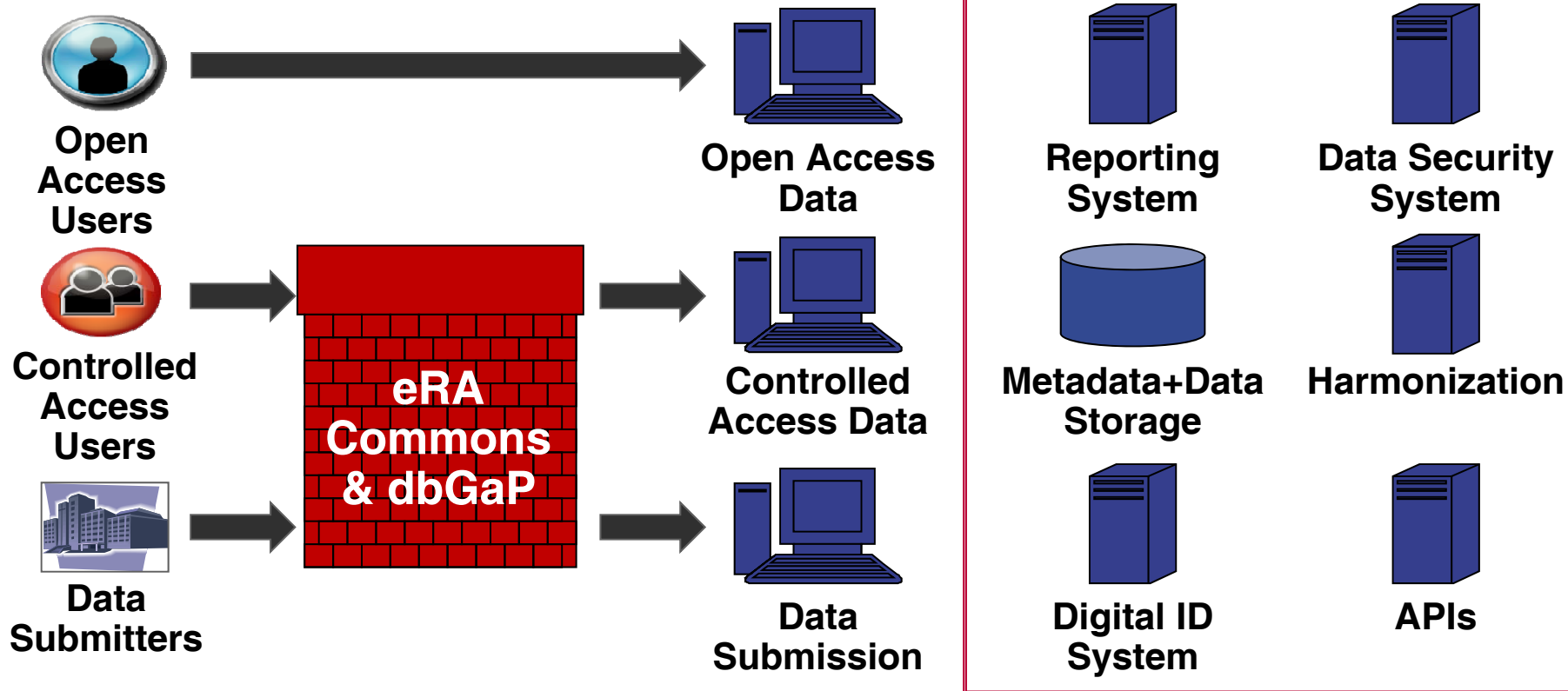


Institute of Medicine
Towards Precision Medicine
2011

GDC Infrastructure and Functionality




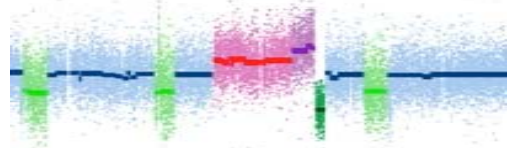
GDC Users

GDC System Components



GDC Data Harmonization

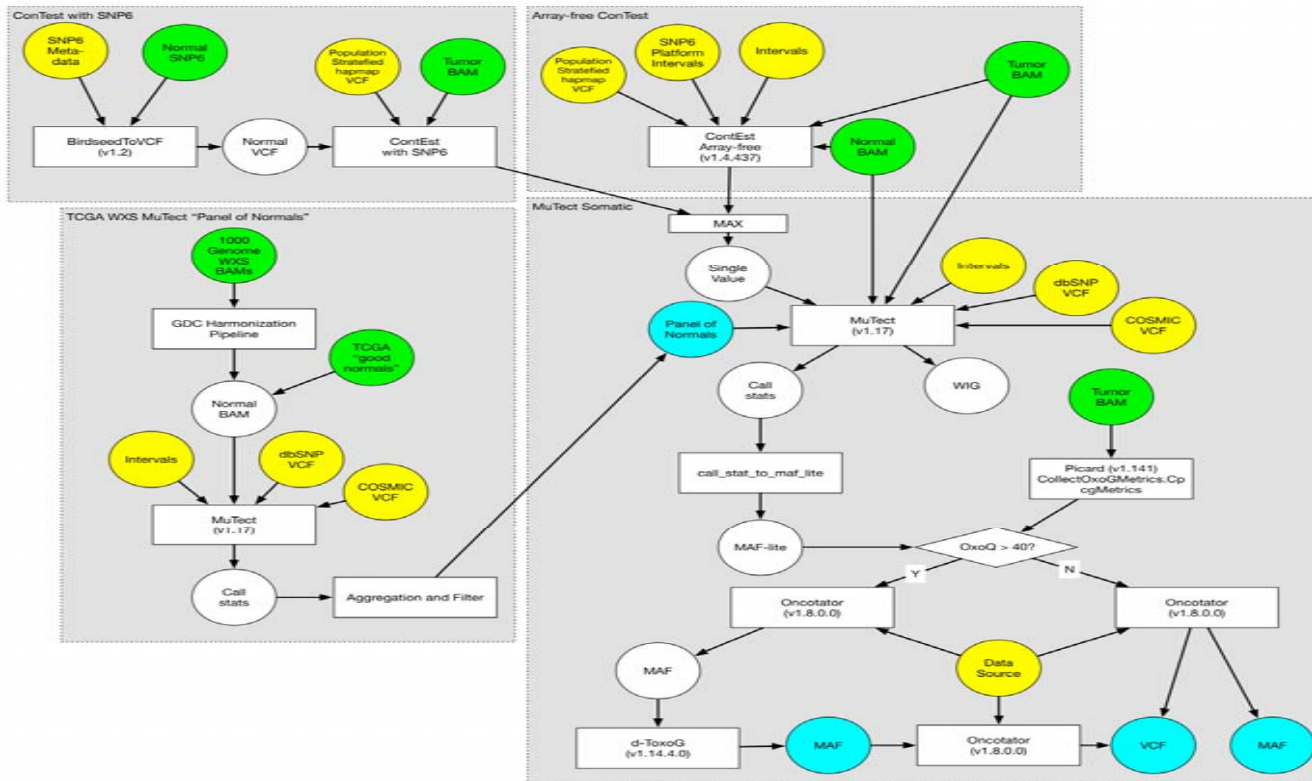
Multiple data types and levels of processing

	<u>1° processing</u>	<u>2° processing</u>	<u>3° processing</u>
<p>Exome-seq</p> 	Genome alignment	Mutations	Oncogene vs. Tumor suppressor
<p>Whole genome-seq</p> 	Genome alignment	Mutations + structural variants	Translocations
<p>RNA-seq</p> 	Genome alignment	Digital gene expression	Relative RNA levels Alternative splicing
<p>Copy number</p> 	Data segmentation	Copy number calls	Gene amplification/ deletion

GDC Data Harmonization

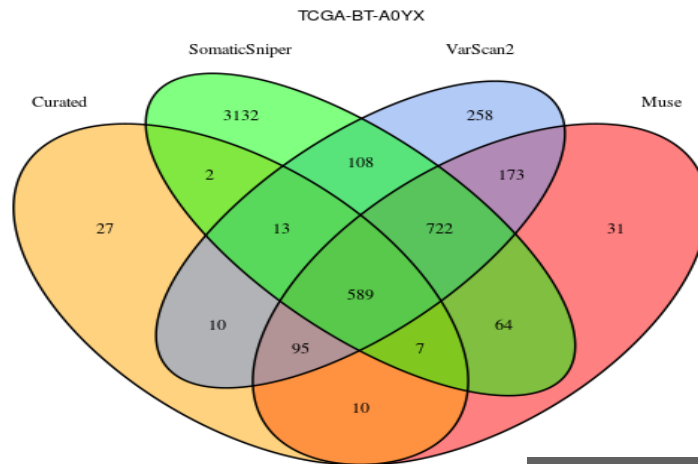
Open Source, Dockerized Pipelines

Mutect2
pipeline



GDC Data Harmonization

Multiple pipelines needed to recover all variants



GDC variant calling
pipelines
 Wash U
 Baylor
 Broad

	Recovery rate (% true positives)
SomaticSniper	81.1%
VarScan	93.9%
MuSE	93.1%
All Three	96.4%

GDC Content

Current

- ❖ TCGA 11,353 cases
- ❖ TARGET 3,178 cases

Coming soon

- ❖ Foundation Medicine 18,000 cases
- ❖ Cancer studies in dbGAP ~4,000 cases

Planned (1-3 years)

- ❖ NCI-MATCH ~3,000 cases
- ❖ Clinical Trial Sequencing Program ~3,000 cases
- ❖ Cancer Driver Discovery Program ~5,000 cases
- ❖ Human Cancer Model Initiative ~1,000 cases
- ❖ APOLLO – VA-DoD ~8,000 cases

~56,000 cases

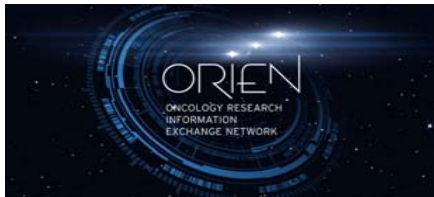


What Makes GDC Special?

- ❖ Stores **raw genomic data**, allowing continuous reanalysis as computation methods and genome annotations improve
- ❖ Utilizes **shared bioinformatic pipelines** to facilitate cross-study comparisons and integrated analysis of multiple data types
- ❖ Maintains **harmonized clinical data** in a highly structured and extensible schema
- ❖ NCI commitment to maintain **long-term storage** of cancer genomic data in the GDC with free access to researchers
- ❖ Enables researchers to comply with the **NIH Genomic Data Sharing policy** as well as journal requirements for data sharing
- ❖ The explanatory power of data in the GDC will grow over time as it accrues more cases => **GDC will promote precision oncology**



Other Cancer Data Sharing Efforts



Signature Efforts

BRCA Challenge
Somatic variant sharing

Precision medicine questions
Somatic variant sharing

Clinical trial
Public-private partnerships

Clinical trial access
Clinical/genomic data
aggregation

Clinical oncology standards

Data

Isolated genetic variants
No raw sequencing data

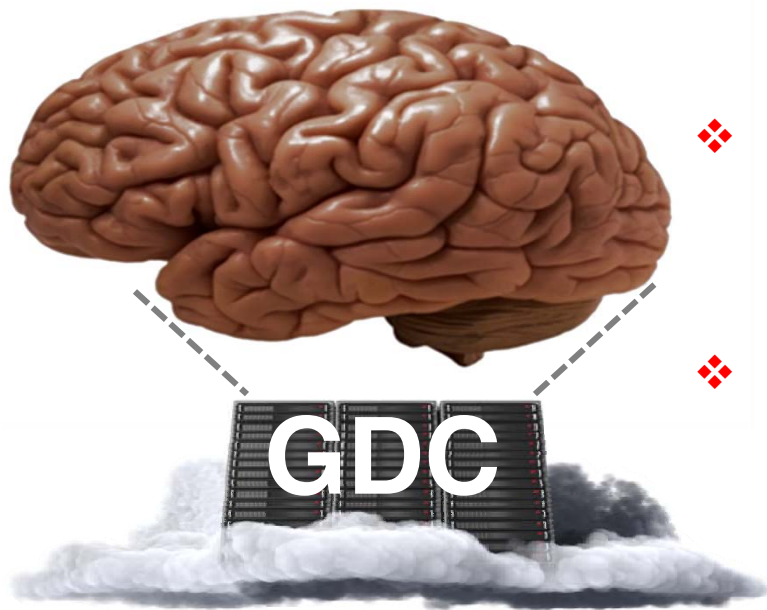
Panel gene resequencing
Clinical response

Comprehensive genomics
Detailed clinical
phenotype data

EHR data
Clinical sequencing

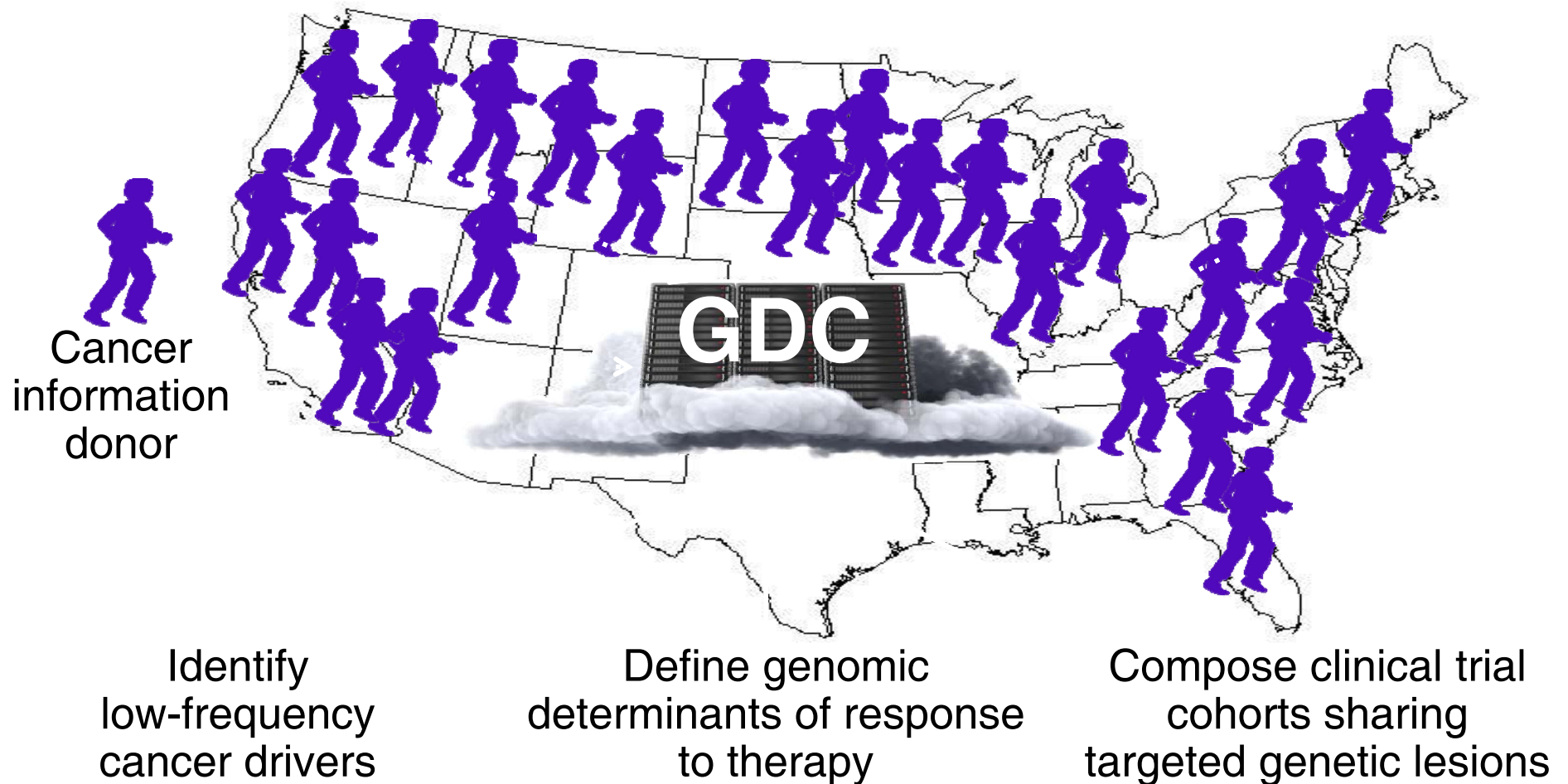
EHR data
Clinical sequencing

Towards a Cancer Knowledge System



- ❖ Continue genomic investigations of cancer
 - ⇒ Need > 100,000 cases analyzed
 - ⇒ Embrace all genomic platforms
 - ⇒ Relationship of relapse and primary biopsies
- ❖ Incorporate associated clinical annotations
 - ⇒ Clinical trial data
 - ⇒ Observational, longitudinal standard-of-care data
 - ⇒ N-of-1 clinical data
- ❖ Promote and curate biological investigations of cancer genetic variants
 - ⇒ Driver vs. passenger mutations
 - ⇒ Multiple phenotypic assays
 - ⇒ Alterations in regulatory pathways – proteomics
 - ⇒ Mechanisms of therapeutic resistance
 - ⇒ Functional genomic investigations
- ❖ Integrative models for high-dimensional data

Utility of a Cancer Knowledge System

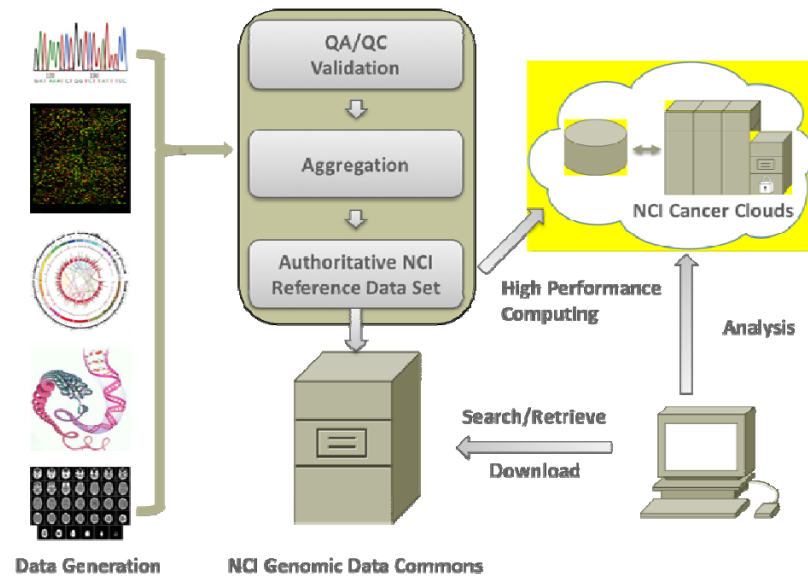




Support the Precision Medicine Initiative

The Genomic Data Commons and Cloud Pilots

- Expand data model to include other data (e.g. imaging and proteomics)
- Allow easy publication of persistent links to data, annotations, algorithms, tools, workflows
- Measure usage and impact
- Change incentives for public contributions



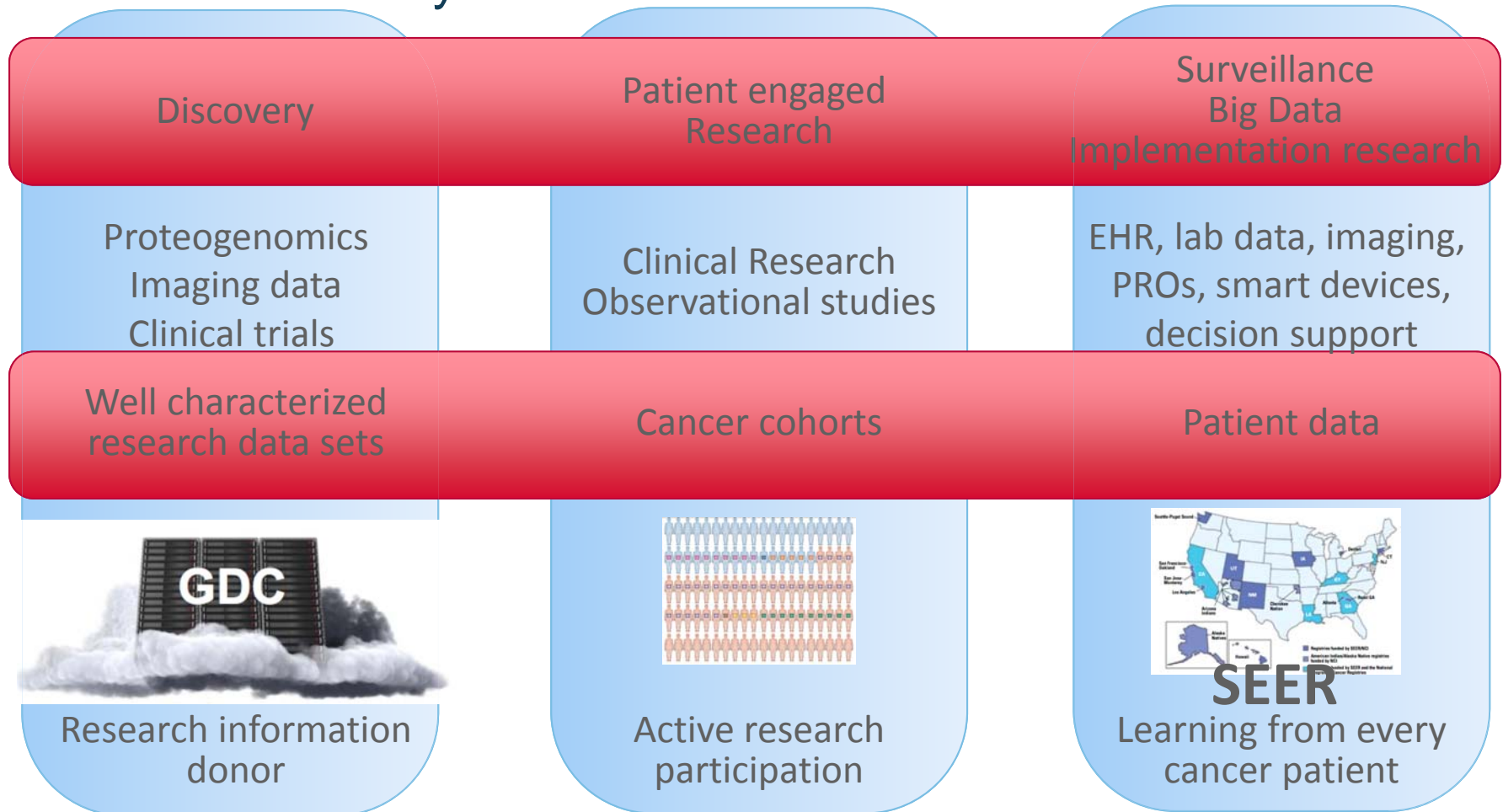
PMI – Oncology, the GDC and the Cloud Pilots Goals

- **Support** precision medicine-focused clinical research
 - Enable researchers to deposit well-annotated (**Interoperable**) genomic data sets with the GDC
 - Provide a single source (and single dbGaP access request!) to **Find** and **Access** these data
 - Enable effective analysis and meta-analysis of these data without requiring local downloads – data **Reuse**
 - Understand **Contributions**, **Assess** value through usage, and give **Attribution** to all users

PMI – Oncology, the GDC and the Cloud Pilots Goals

- **Provide** a data integration platform to allow multiple data types, multi-scalar data, temporal data from cancer models and patients through **open APIs**
- Work with the Global Alliance for Genomics and Health (**GA4GH**) to **define** the next generation of **secure, flexible, meaningful, interoperable, lightweight interfaces – open APIs**
- Engage the cancer research community in **evaluating** the **open APIs** for ease of use and effectiveness

Cancer data ecosystem



GDC Acknowledgements

NCI Center for Cancer Genomics

Lou Staudt
Zhining Wang
Martin Ferguson
JC Zenklusen
Daniela Gerhard
Deb Steverson

NCI NCI CBIIT

Tony Kerlavage
Tanya Davidsen

Leidos Biomedical Research

Mark Jensen
Sharon Gaheen
Himanso Sahni

Univ. of Chicago

Bob Grossman
Allison Heath
Mike Ford
Zhenyu Zhang

Ontario Institute for Cancer Research

Vincent Ferretti
Francois Gerthoffert
JunJun Zhang

Cancer Genomics Project Teams

CGC Pilot Team Principal Investigators

- Gad Getz, Ph.D - Broad Institute - <http://firecloud.org>
- Ilya Shmulevich, Ph.D - ISB - <http://cgc.systemsbiology.net/>
- Deniz Kural, Ph.D - Seven Bridges – <http://www.cancergenomicscloud.org>

NCI Project Officer & CORs

- Anthony Kerlavage, Ph.D –Project Officer
- Juli Klemm, Ph.D – COR, Broad Institute
- Tanja Davidsen, Ph.D – COR, Institute for Systems Biology
- Ishwar Chandramouliswaran, MS, MBA – COR, Seven Bridges Genomics

GDC Principal Investigator

- Robert Grossman, Ph.D - University of Chicago
- Allison Heath, Ph.D - University of Chicago
- Vincent Ferretti, Ph.D - Ontario Institute for Cancer Research

Center for Cancer Genomics Partners

- JC Zenklusen, Ph.D.
- Daniela Gerhard, Ph.D.
- Zhining Wang, Ph.D.
- Liming Yang, Ph.D.
- Martin Ferguson, Ph.D.

NCI Leadership Team

- Doug Lowy, M.D.
- Lou Staudt, M.D., Ph.D.
- Stephen Chanock, M.D.
- George Komatsoulis, Ph.D.
- Warren Kibbe, Ph.D.

Questions?

Louis Staudt, M.D., Ph.D.

lstaudt@mail.nih.gov

Warren Kibbe, Ph.D.

Warren.kibbe@nih.gov



@wakibbe





**NATIONAL
CANCER
INSTITUTE**

www.cancer.gov

www.cancer.gov/espanol