Genomic Data Commons

NCI Cloud Pilots

Louis Staudt, MD, PhD

Warren Kibbe, PhD

@wakibbe

NIH NATIONAL CANCER INSTITUTE

Spetember 7th, 2016

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

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

NIH Launches Genomic Data Commons Supporting Cancer Moonshot

REUTERS

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

fedscoop

Biden launches data portal to back Cancer Moonshot

Genomic Data Commons (GDC)

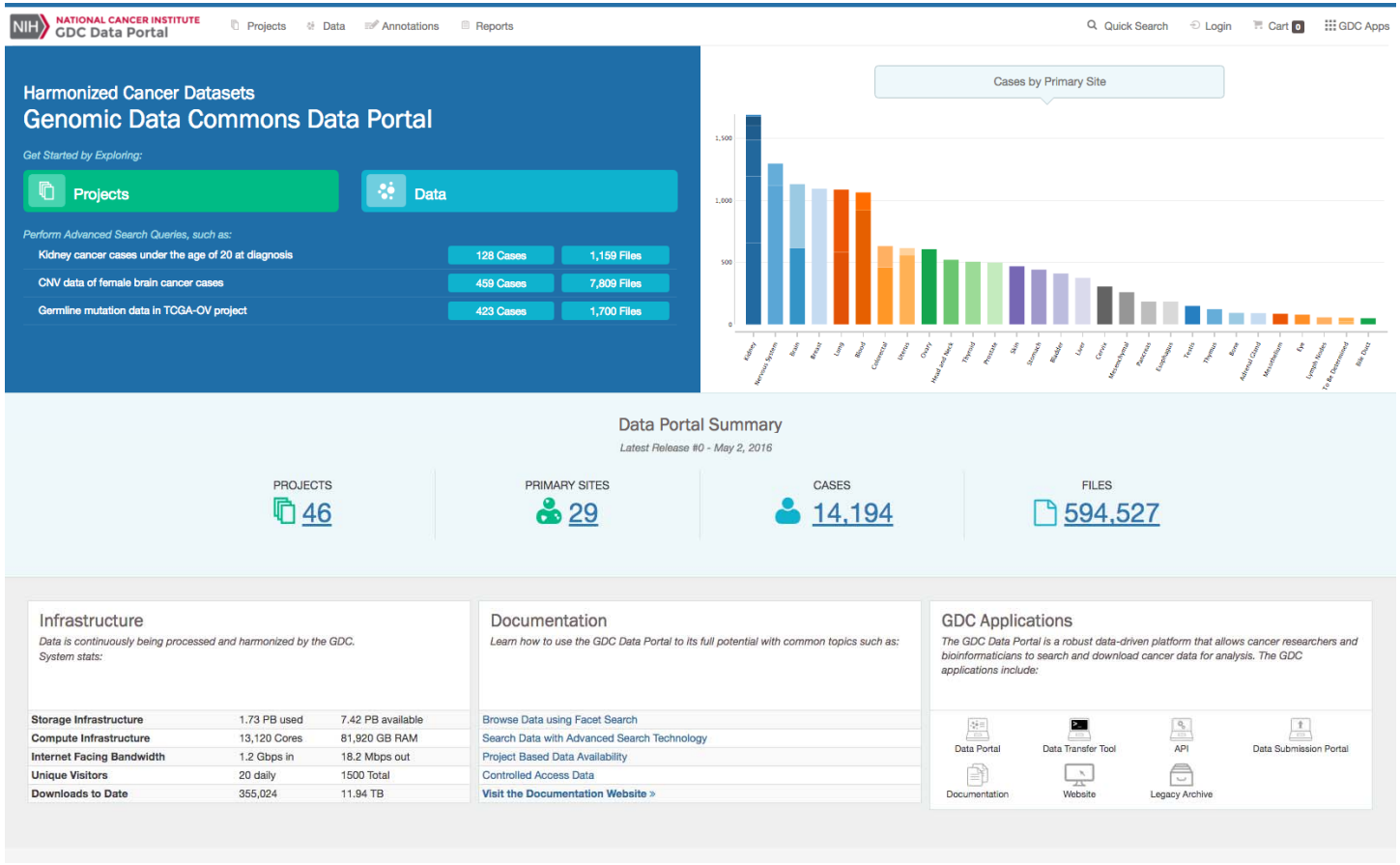
was highlighted in the June 29th Cancer Moonshot Summit at Howard University in the US

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

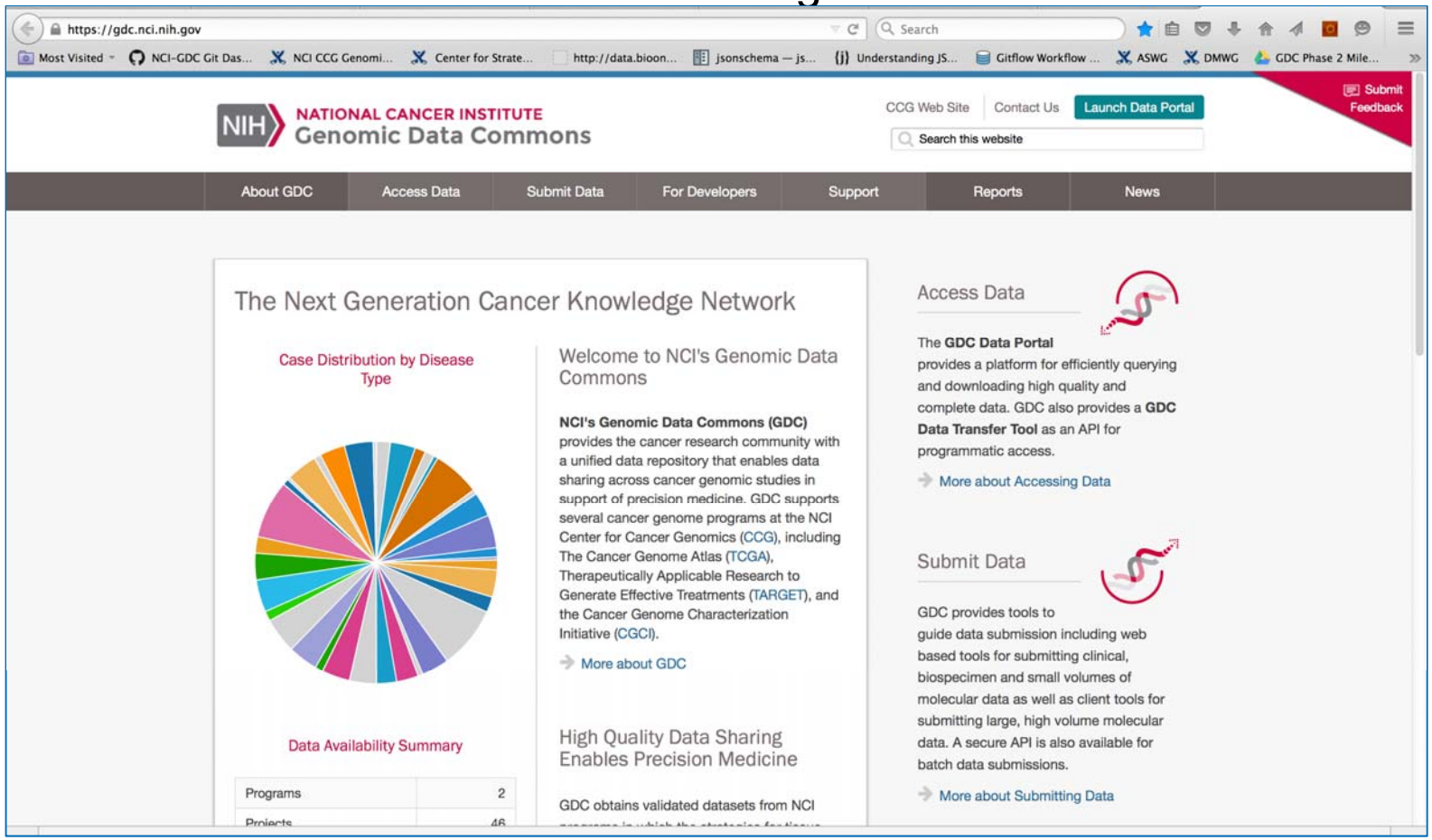
NCI Genomic Data Commons

- 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.

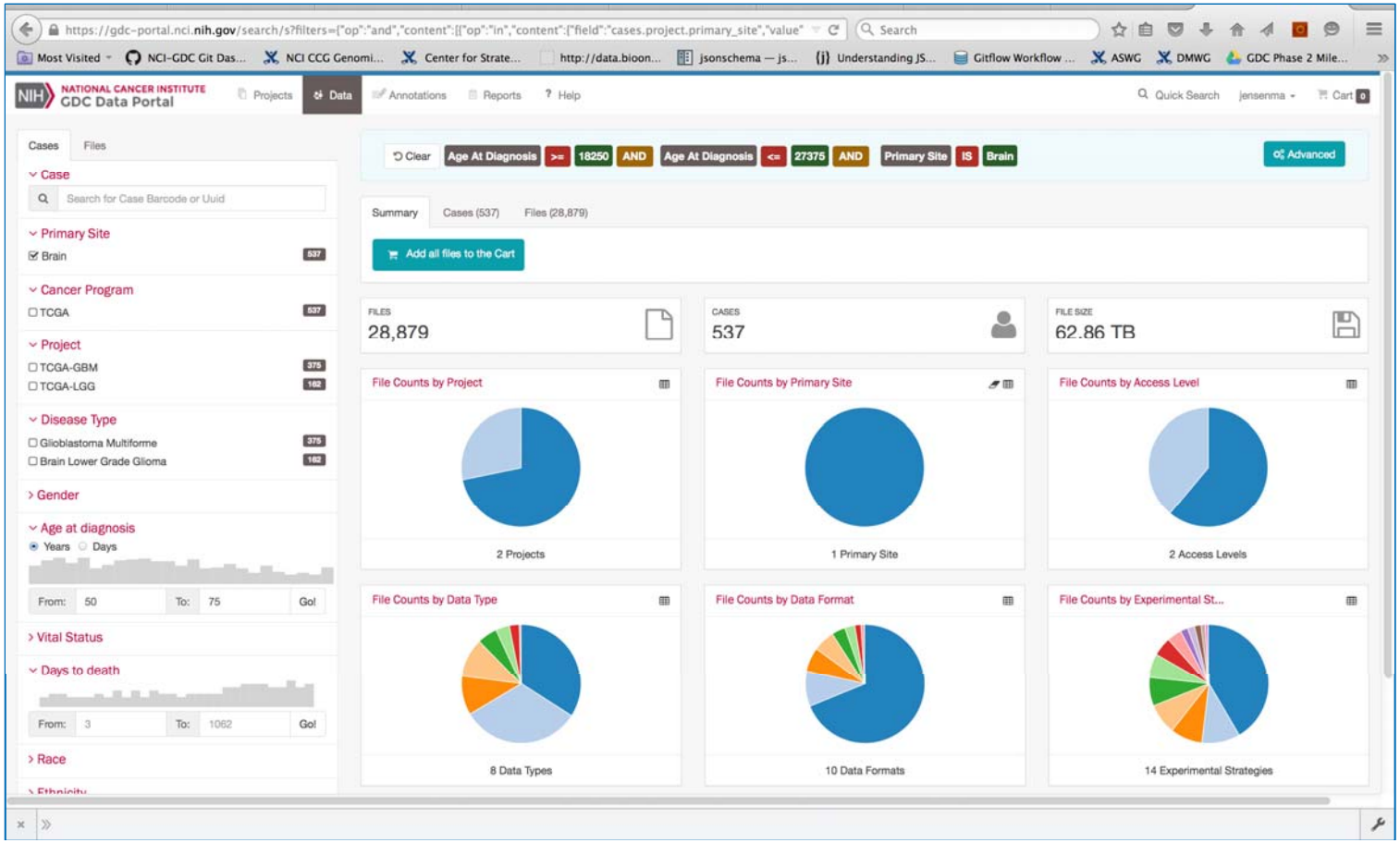
Genomic Data Commons Data Portal



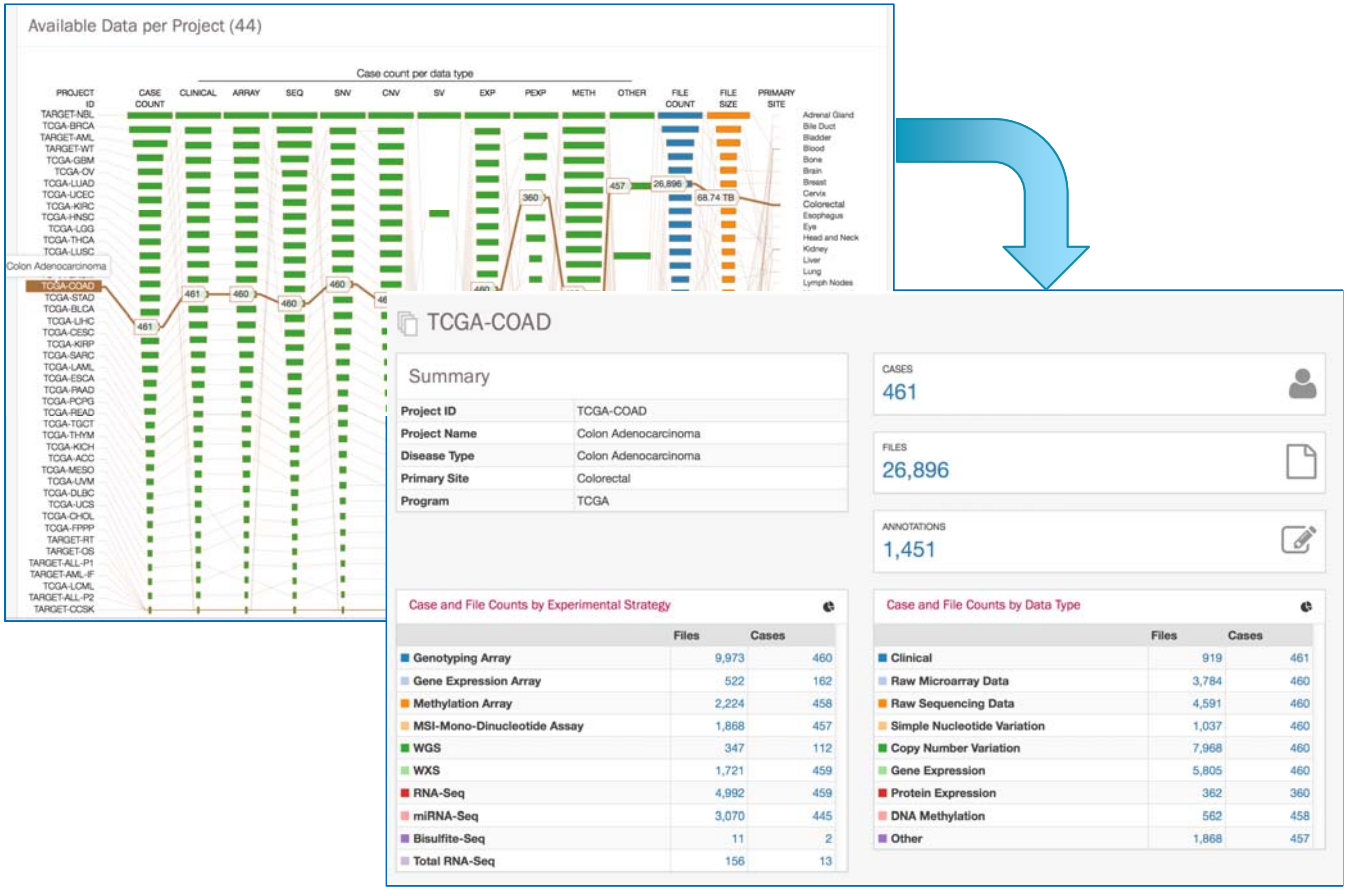
The NCI Genomic Data Commons User Interface Home Page



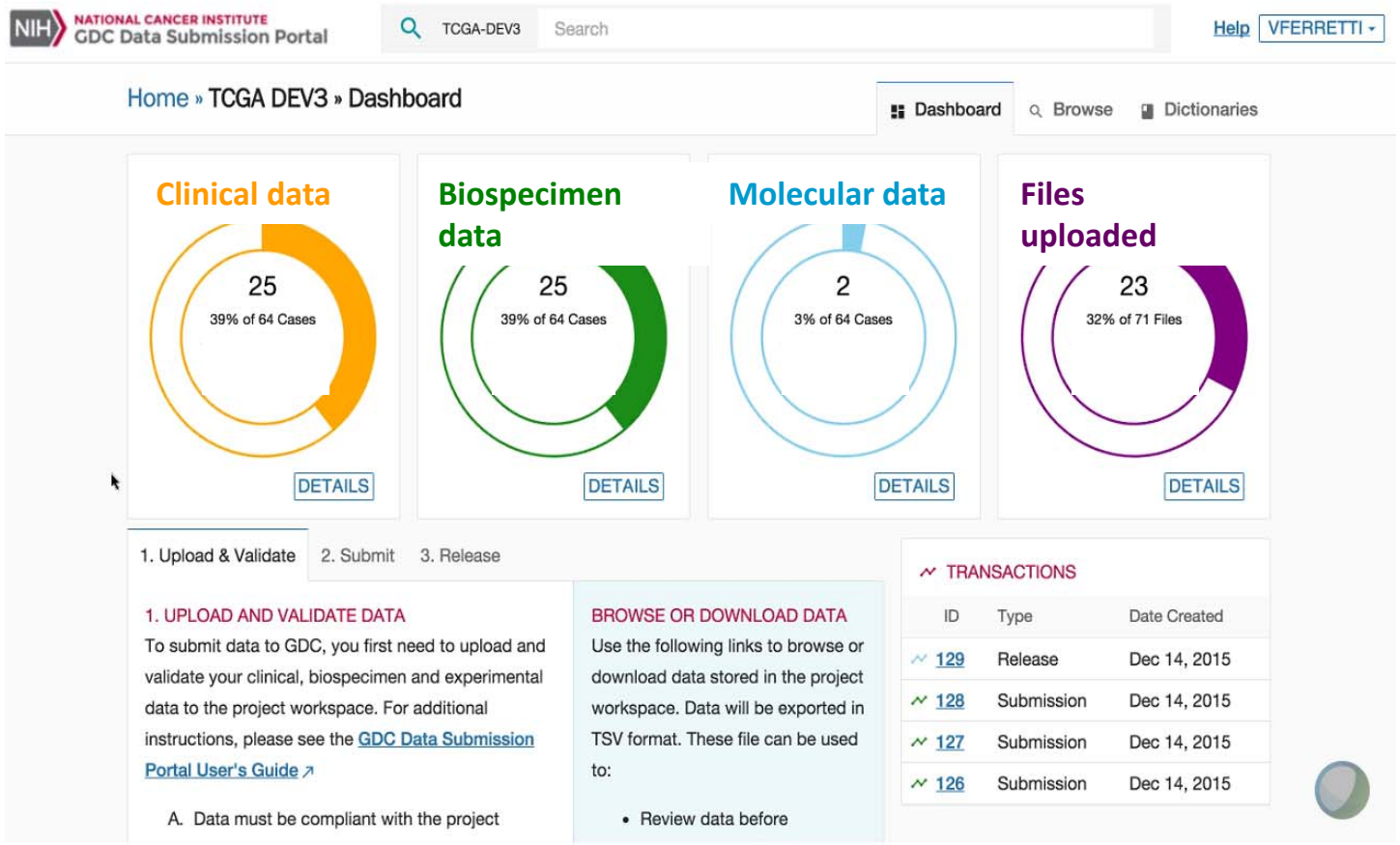
The NCI Genomic Data Commons User Interface
Sample Browser



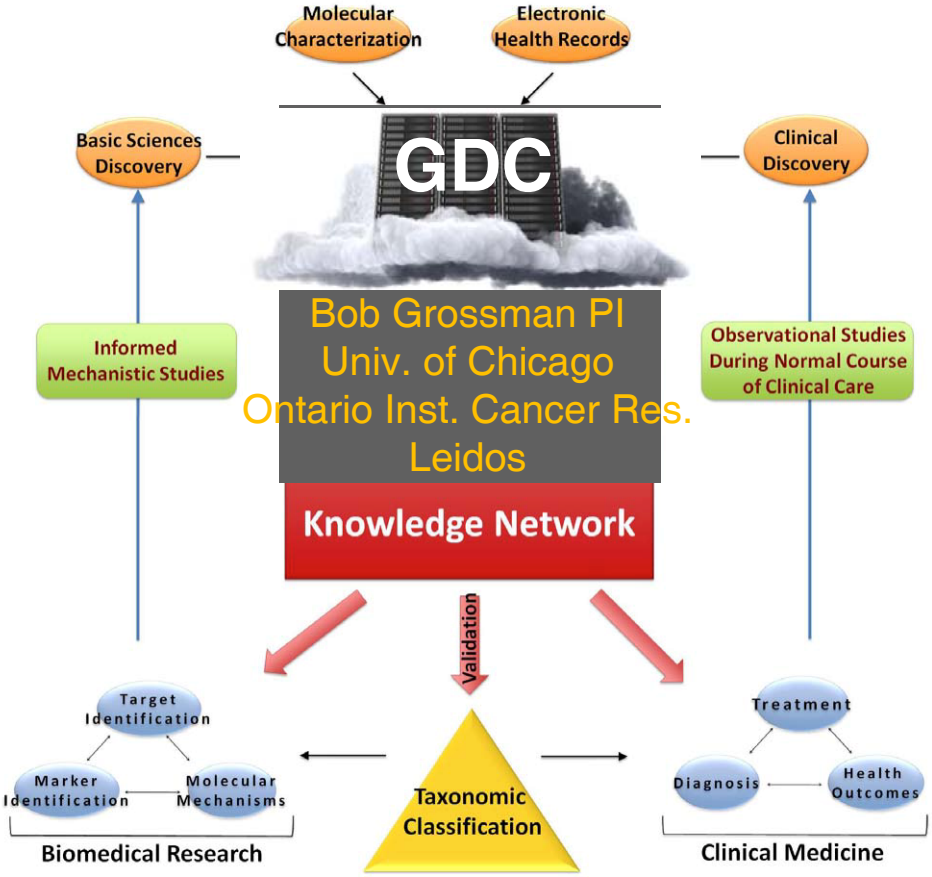
The NCI Genomic Data Commons User Interface
Sample Selection



The NCI Genomic Data Commons User Interface
Data Submission Dashboard

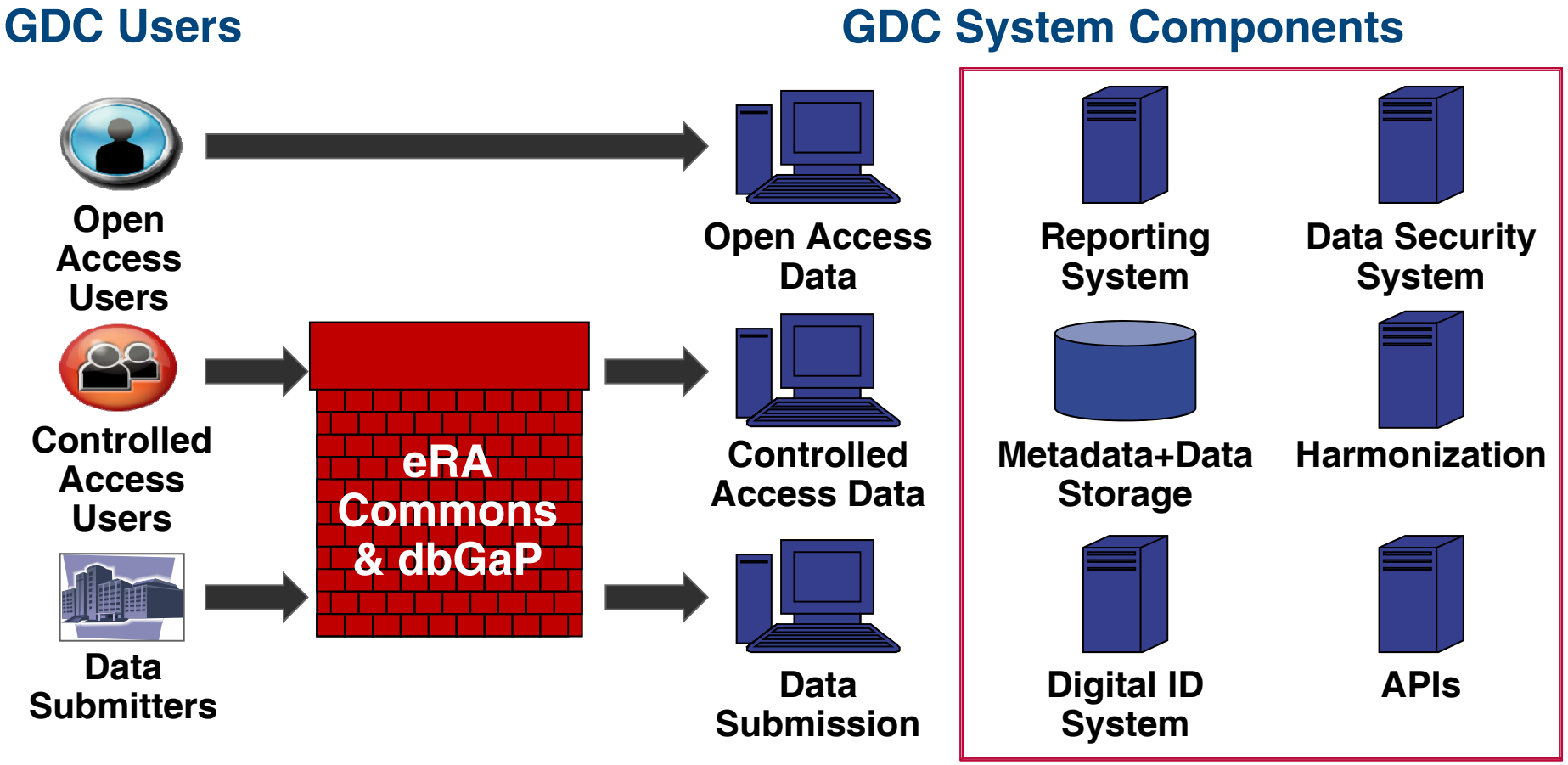


Development of the NCI Genomic Data Commons (GDC)
To Foster the Molecular Diagnosis and Treatment of Cancer

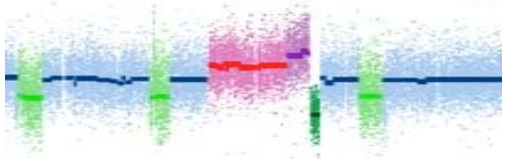


Institute of Medicine
Towards Precision Medicine
2011

GDC Infrastructure and Functionality



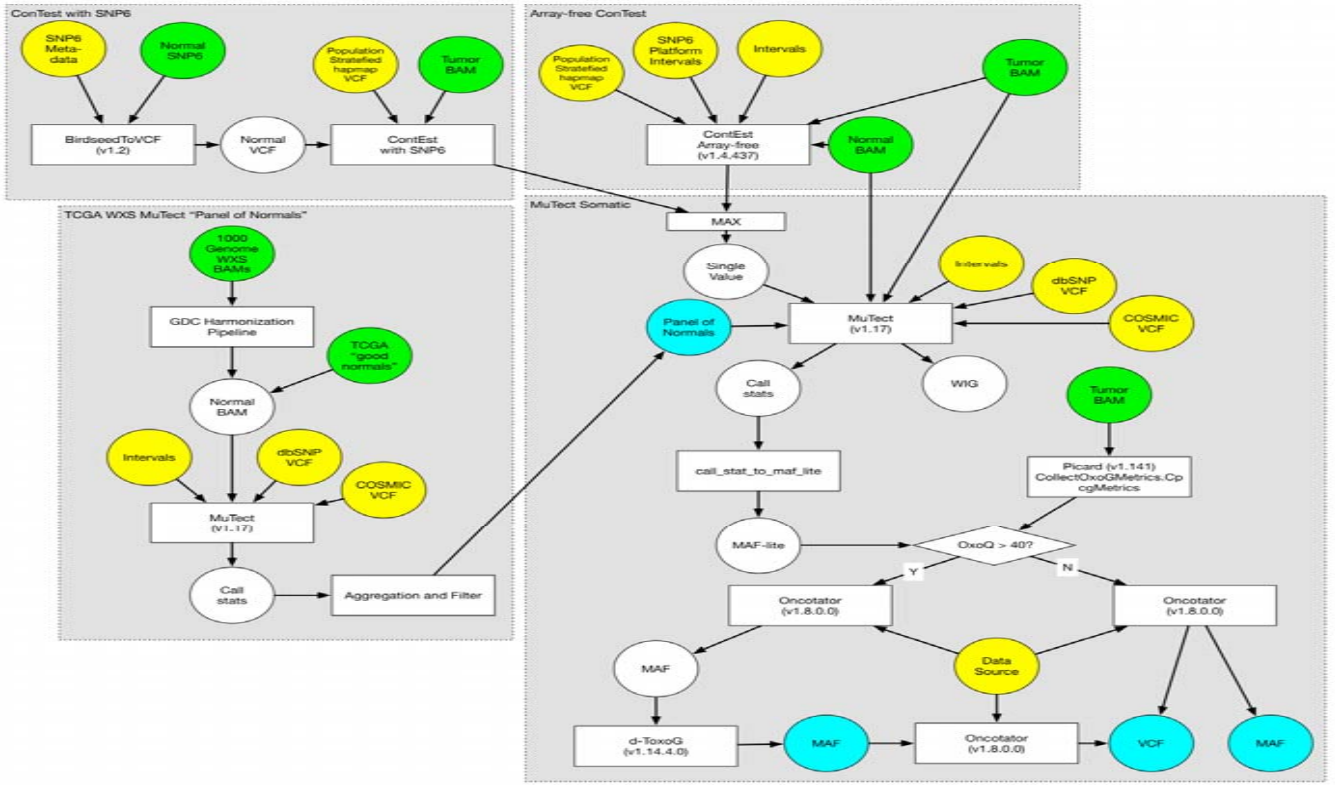
GDC Data Harmonization
Multiple data types and levels of processing

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

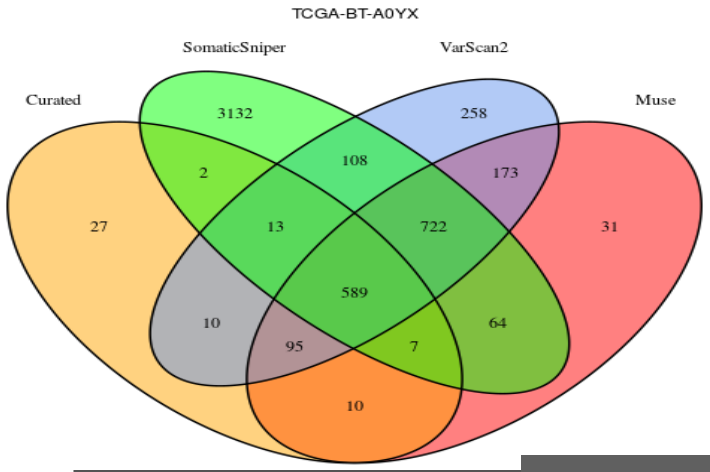
GDC Data Harmonization

Open Source, Dockerized Pipelines

Mutect2 pipeline



GDC Data Harmonization
Multiple pipelines needed to recover all variants



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

GDC variant calling <u>pipelines</u>

Wash U

Baylor

Broad

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

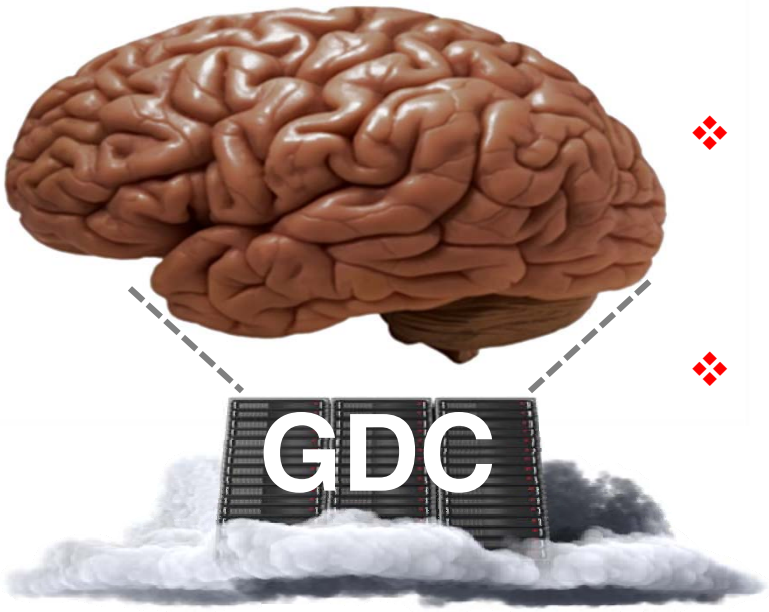
What Makes GDC Special?

- 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	Data
Global Alliance for Genomics & Health	BRCA Challenge Somatic variant sharing	Isolated genetic variants No raw sequencing data
PROJECT GENIE Genomics Evidence Neoplasia Information Exchange	Precision medicine questions Somatic variant sharing	Panel gene resequencing Clinical response
ORIEN Oncology Research Information Exchange Network	Clinical trial Public-private partnerships	Comprehensive genomics Detailed clinical phenotype data
Oncology Precision Network	Clinical trial access Clinical/genomic data aggregation	EHR data Clinical sequencing
ASCO CancerLinQ Learning Intelligence Network for Quality	Clinical oncology standards	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

Cancer information donor

Identify low-frequency cancer drivers

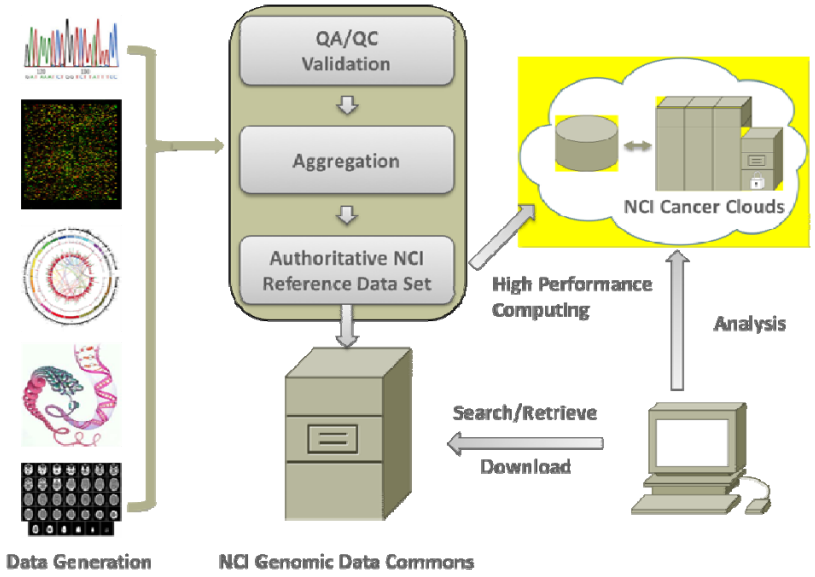
Define genomic determinants of response to therapy

Compose clinical trial cohorts sharing targeted genetic lesions

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



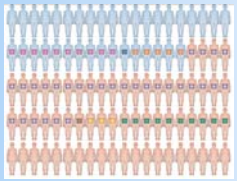
PMI – Oncology, the GDC and the Cloud Pilots Goals

- **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

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
		 SEER
Research information donor	Active research participation	Learning from every cancer patient

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