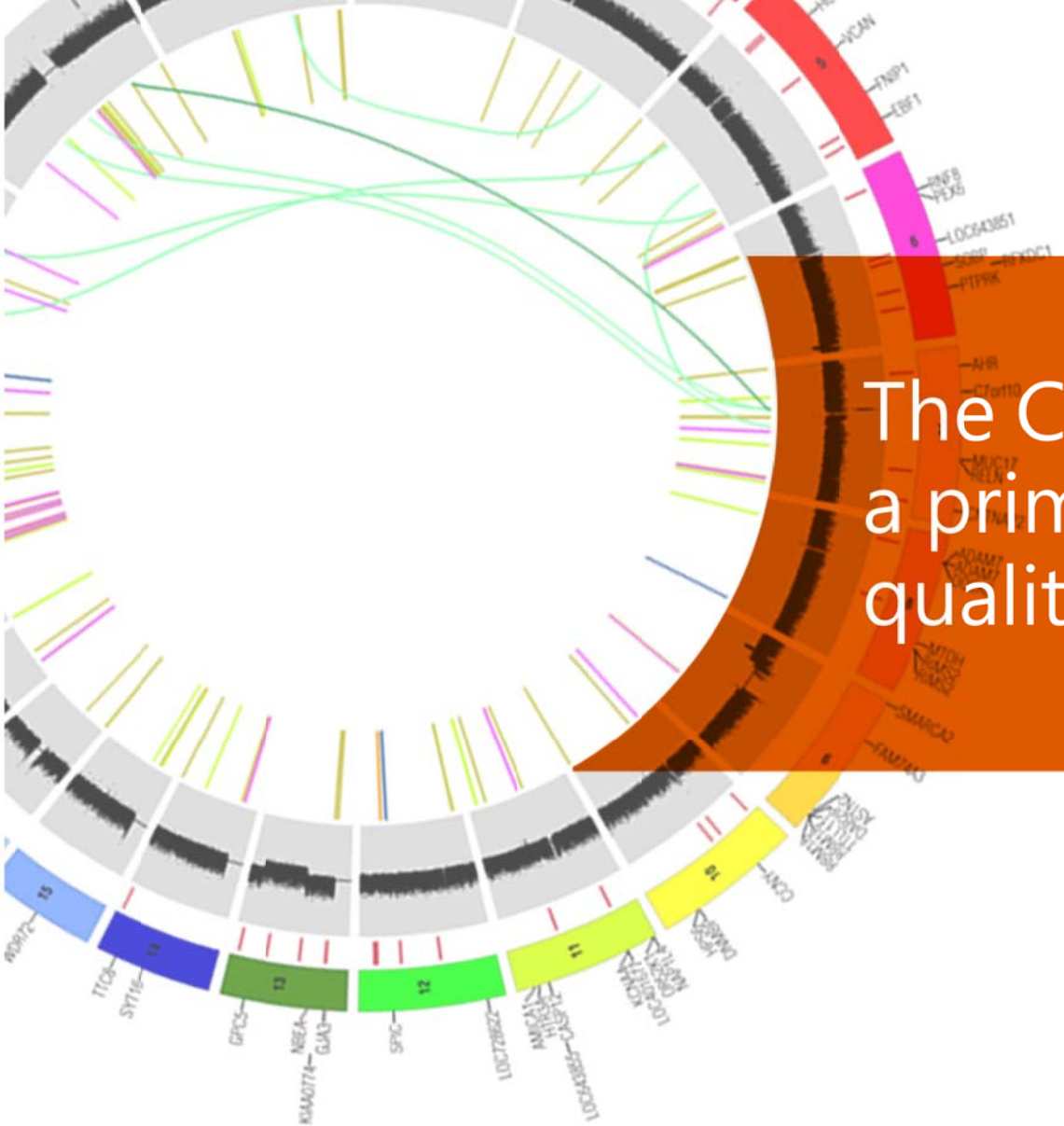


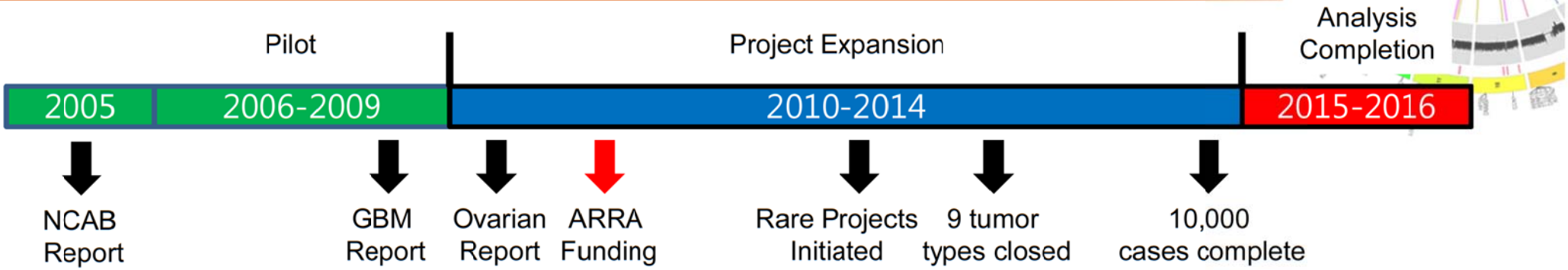
The Cancer Genome Atlas

The Cancer Genome Atlas: a primer on 2.5 Pb of high-quality data

September 2nd, 2016

Jean-Claude ZenKlusen, Ph.D.
Director: TCGA Program Office
zenklusj@mail.nih.gov

TCGA: Timeline



- **Pilot Projects: GBM and Ovarian carcinoma (~500 cases ea.)**
  - Establish infrastructure for effective team science
  - Develop a scalable "pipeline"
  - Demonstrate the feasibility of a large-scale, high throughput approach to identifying the molecular 'parts-list'
  - Make the data publicly and broadly available to the cancer community while protecting patient privacy

- **Expansion 2010 to 2014:**
  - Add 25-35 tumor types
  - Enhancement of sample acquisition & program staff
  - Add Genome Data Analysis Centers
  - Publish "Benchmark Marker Papers"
  - Established FFPE protocols
  - Completely characterize 10,000th case

- **Analysis Completion 2015-2016:**
  - Finish marker papers on rare & "challenging-to-accrue" tumors
  - Complete Pan-Cancer Analysis
  - Broader sharing of tools, analytical methods

TCGA: What's in a Core Data Set?

Data from Tissue Source Sites

- Complete path report
- Paired metastatic samples
- Double normals
- Treatment data

Core Data Set

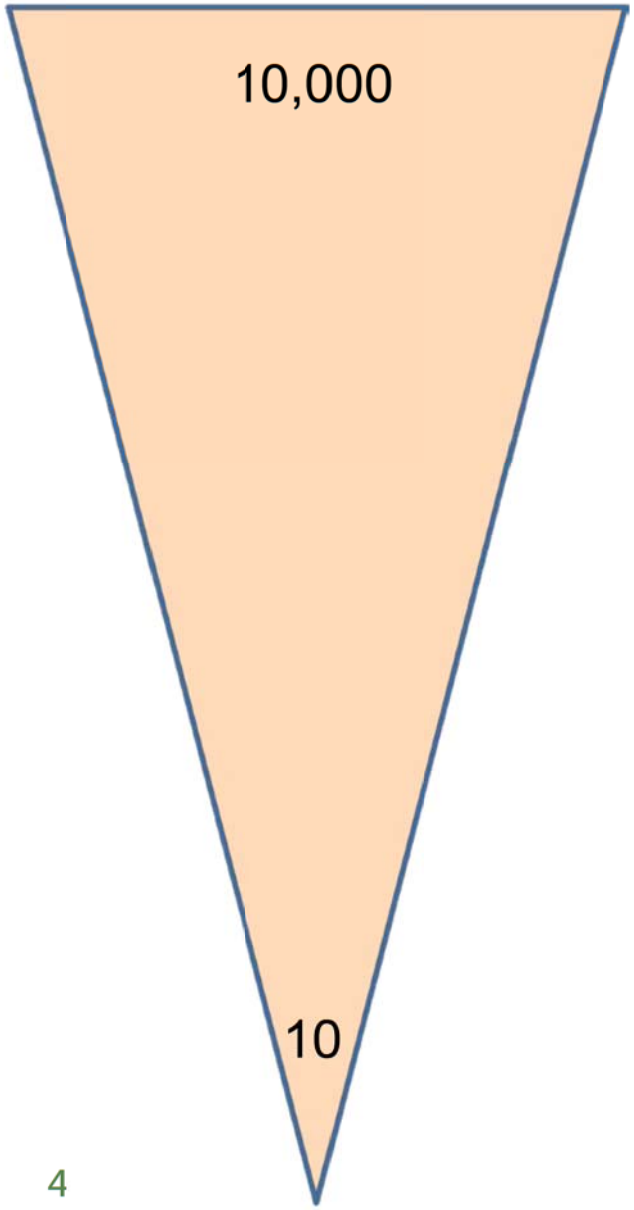
- Synoptic path report
- Histology images
- Required clinical data
- Whole exome
- SNP 6.0 array
- mRNAseq
- miRNAseq
- Methylation array

Data Generated by GCCs &GSCs

- 50X WGS
- 8X WGS
- Methylseq
- RPPA

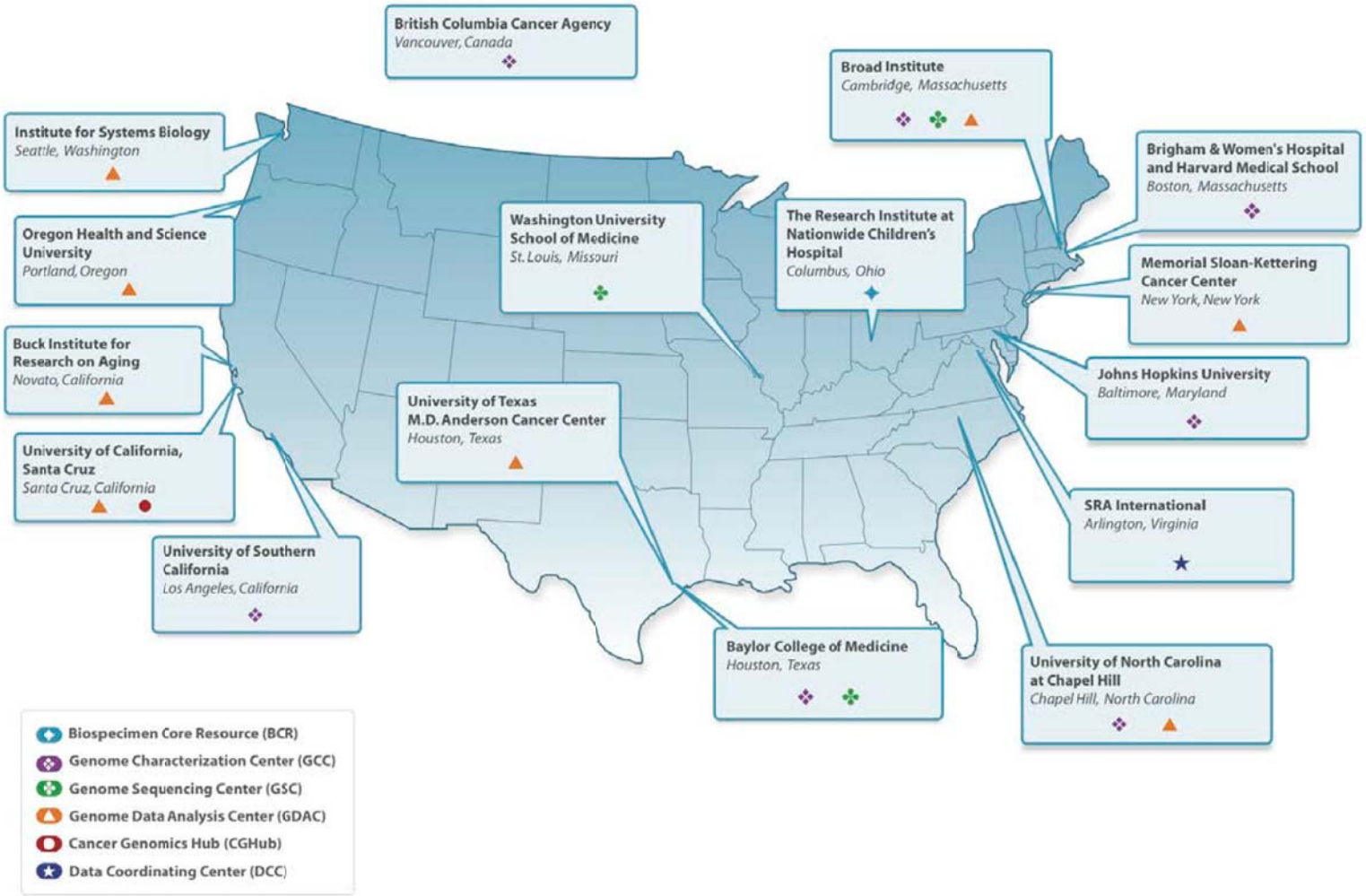
(May not apply to GBM/Ovarian cases collected during the pilot phase)

Sample Criteria Limit 'Askable' Questions

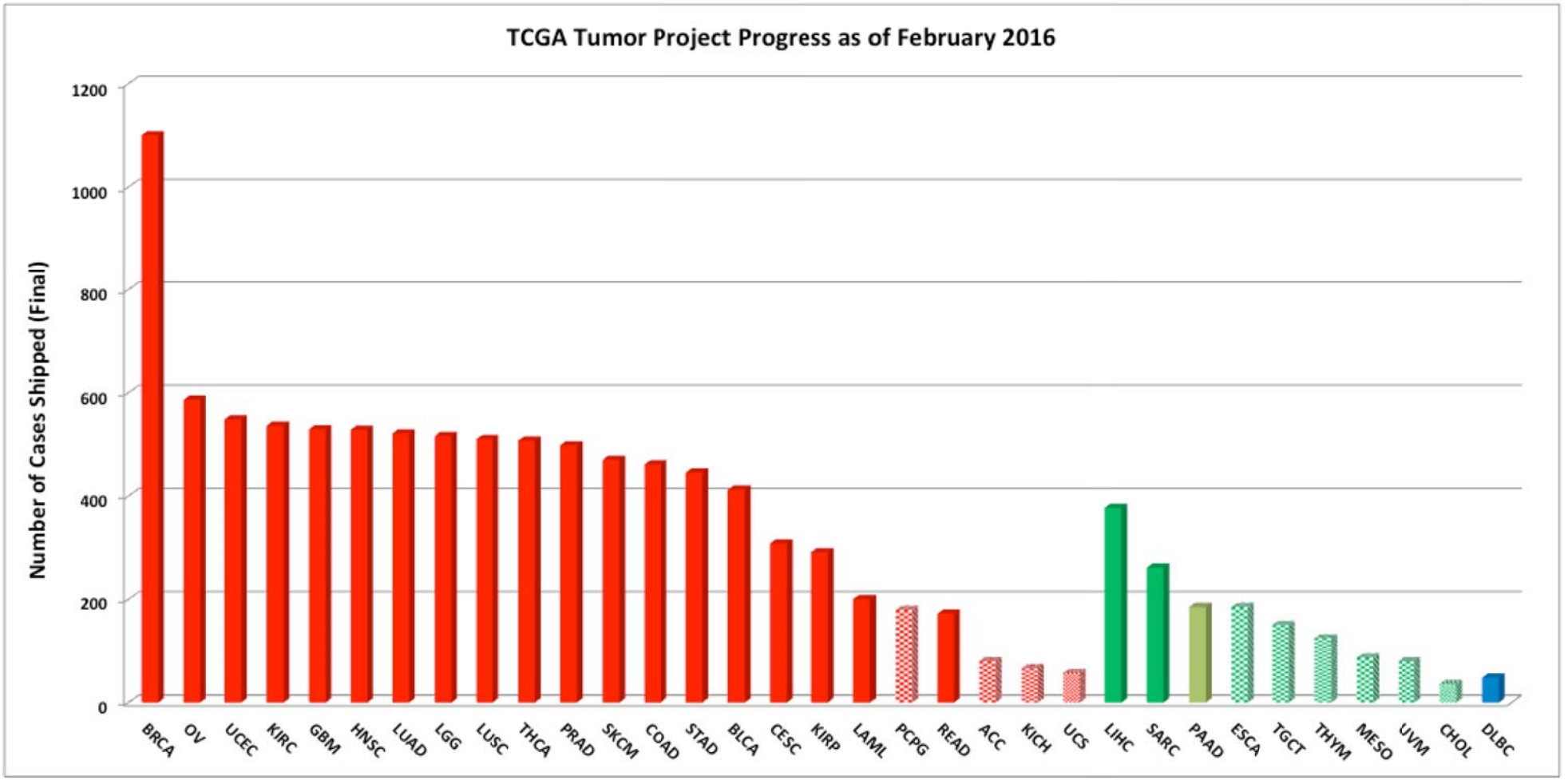


- Primary, *adult* tumors (except for melanoma and triplets)
- Malignant (no *in situ* cases)
- Snap frozen, <60min from clamp to LN2
- ~ 50 mg (biopsies starting to be feasible)
- Pathology review of tissue sent to TCGA
- No more than 20% necrosis ; ≥ 60%* tumor cells
- No prior treatment
- *Matched* source of germline: Blood (buffy coat/white cells)/saliva or skin for liquid tumors
- Clinical annotation; but not pre-analytic variables
- IRB approval for use in TCGA; proactive consenting for genomic studies
- MTA w/out retention of IP

TCGA Network



Tumor Project Progress



TCGA: The Pipeline for Comprehensive Characterization

- Model Informed Consents
- Novel contracting methods to incentivize quality

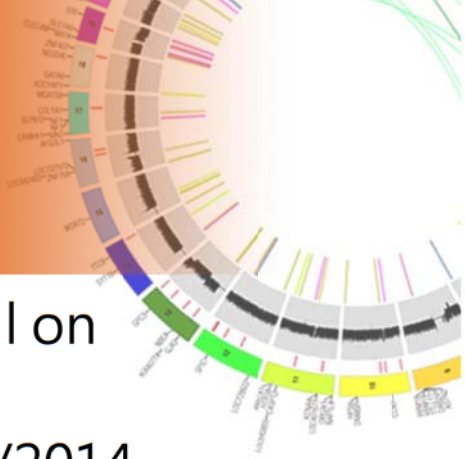
- FFPE isolation protocols
- Remote pathology review
- Reduction of batch effects

- FFPE sequencing/array protocols
- Benchmarking datasets

- Innovating Data storage/release policies
- New data standards (e.g. cancer .vcf)

- Innovating cancer genomes analysis
- Cross-tumor discovery

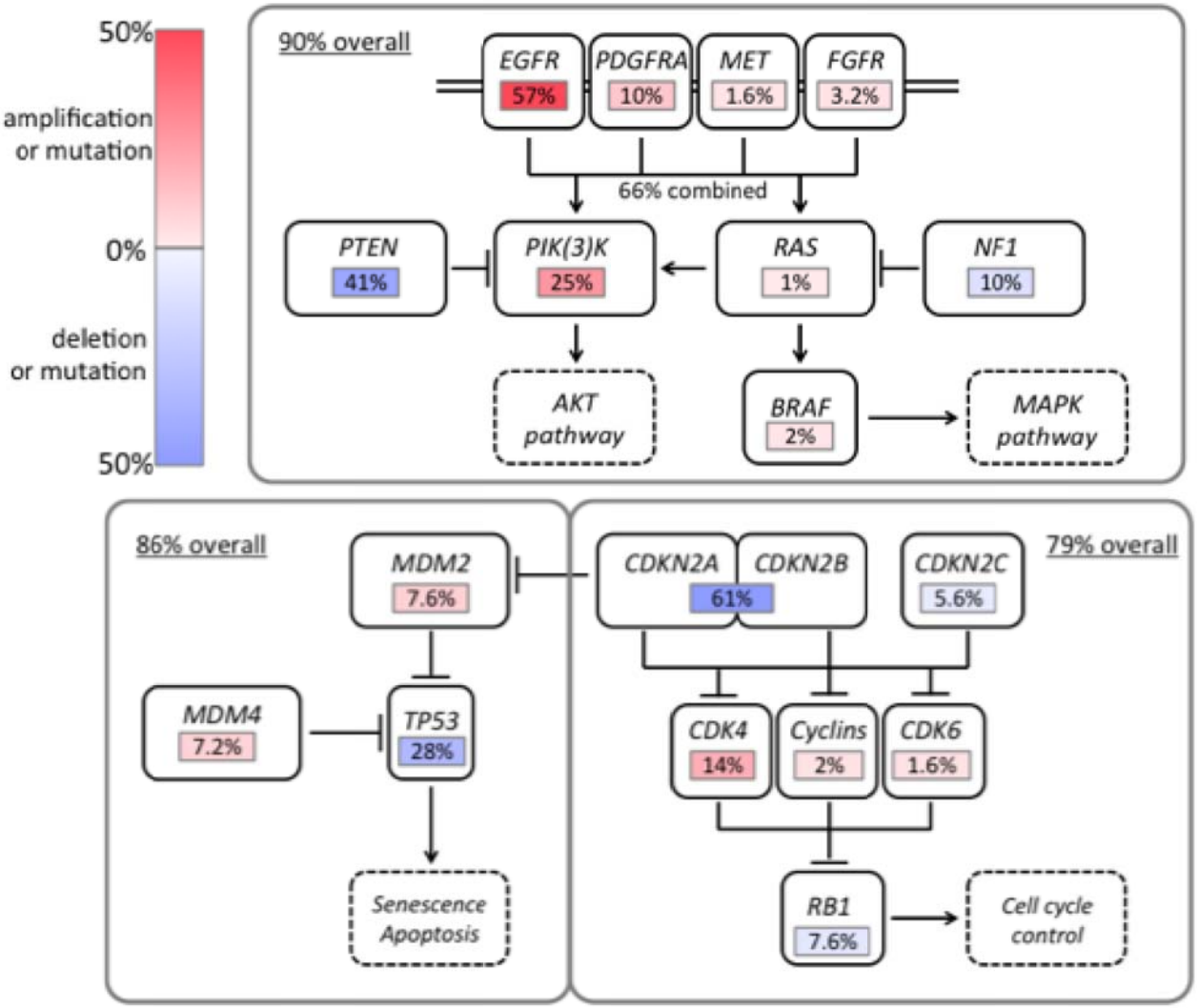
Recent Publications



- Kidney Chromophobe
- Thyroid Papillary
- Head and Neck
- Skin Melanoma
- Low Grade Glioma
- Lobular Breast Carcinoma
- Prostate Adenocarcinoma
- LGG-GBM
- Kidney Papillary Carcinoma
- Pan-Lung
- Adenocortical Carcinoma
- Liver Adenocarcinoma
- Uterine Carcinosarcoma
- Cervical Carcinoma

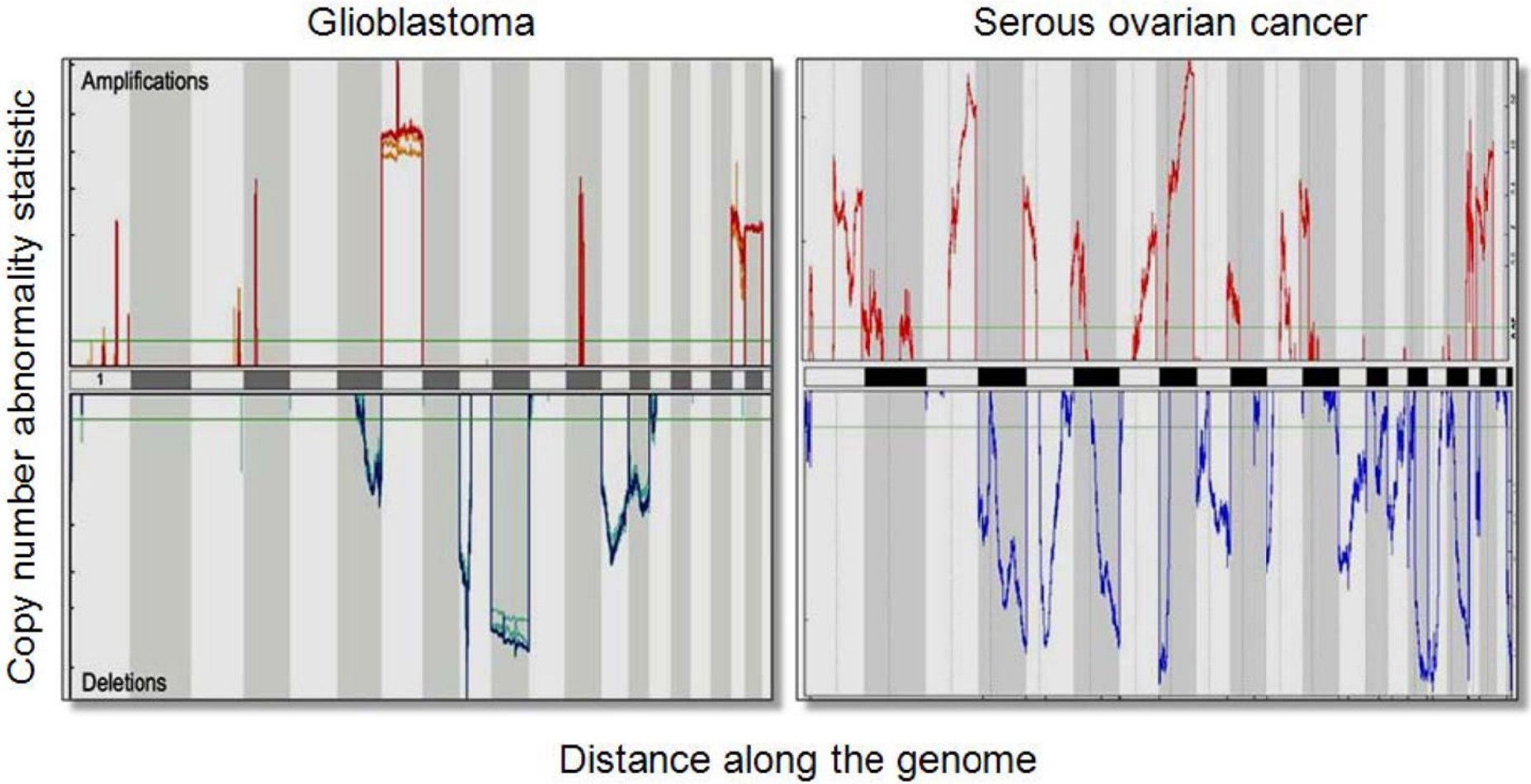
- Published Cancer Cell on 09/08/2014
- Published Cell 10/23/2014
- Published Nature on 01/29/2015
- Published Cell 06/18/2015
- Published NEJM 06/25/2015
- Published Cell 10/08/2015
- Published Cell 11/05/2015
- Published Cell 01/28/2016
- Accepted NEJM 09/2015
- Accepted Nature Genetics
- Submitted to Cancer Cell
- Submitted to Nature Genetics
- Submitted to Nature
- Submitted to Nature

TCGA: Lessons Learned from the Data Pathway Analyses

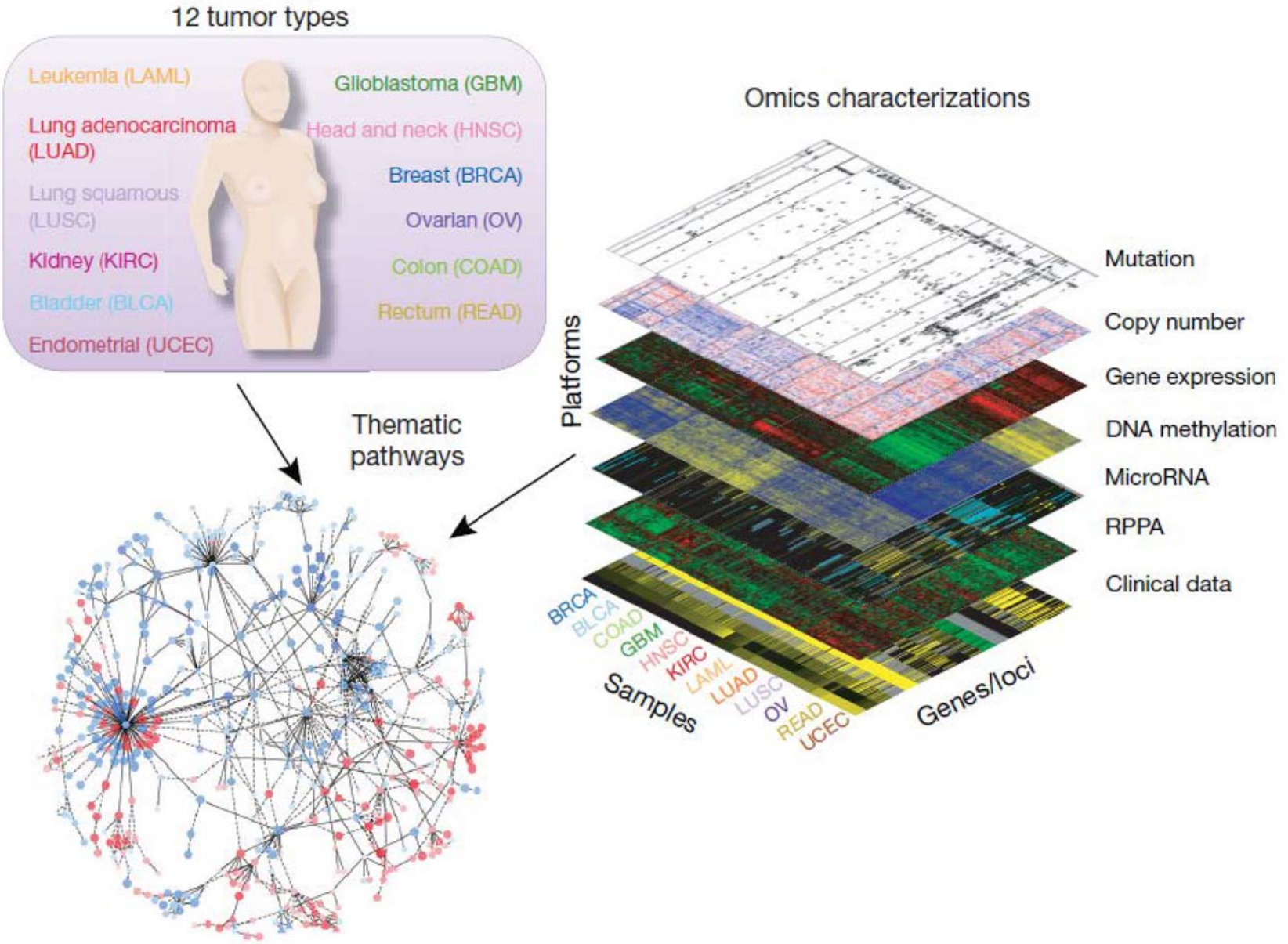


TCGA: Lessons Across Cancers Learned from the Data

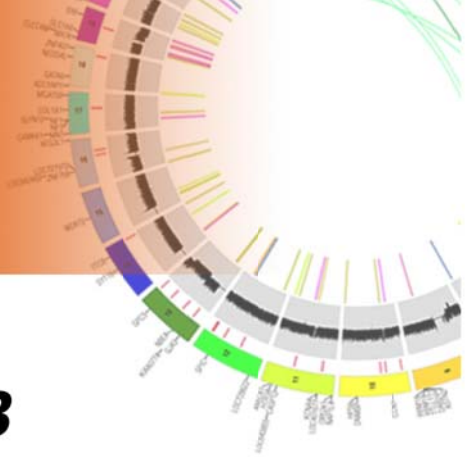
A contrast in copy number complexity



Integration Matters



Most of TCGA Data are Open Access



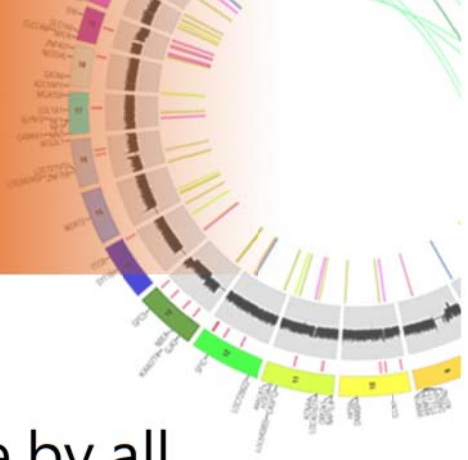
Platform	*Level 1*	*Level 2*	*Level 3*
Copy Number	controlled	controlled	open
Methylation	open	open	open
mRNA array	open*	open	open
mRNAseq	-	controlled	open
miRNA Seq	-	controlled	open
DNA Seq	-	controlled	open
Clinical	open		

Mutations

Sequence Data (BAM file)

*Level 1 data from Affy exon arrays in early GBM & Ov studies are controlled access

Data Use Policy



- TCGA is a community resource project
  - Data are released rapidly (pre-publication) for use by all
  - Ft. Lauderdale principles:
    - Obligation of data users to consider etiquette of publishing research using pre-publication data of others
    - Obligation of data generators to publish findings promptly
- TCGA clarifies policy. There are no limitations on publication using TCGA data sets if:
  - A marker paper has been published on the tumor type, OR
  - 18 months have transpired after 100 cases of a tumor type have entered into data generation, OR
  - The author receives approval from the TCGA Publication Committee in consultation with the tumor analysis working group (tcga@mail.nih.gov).

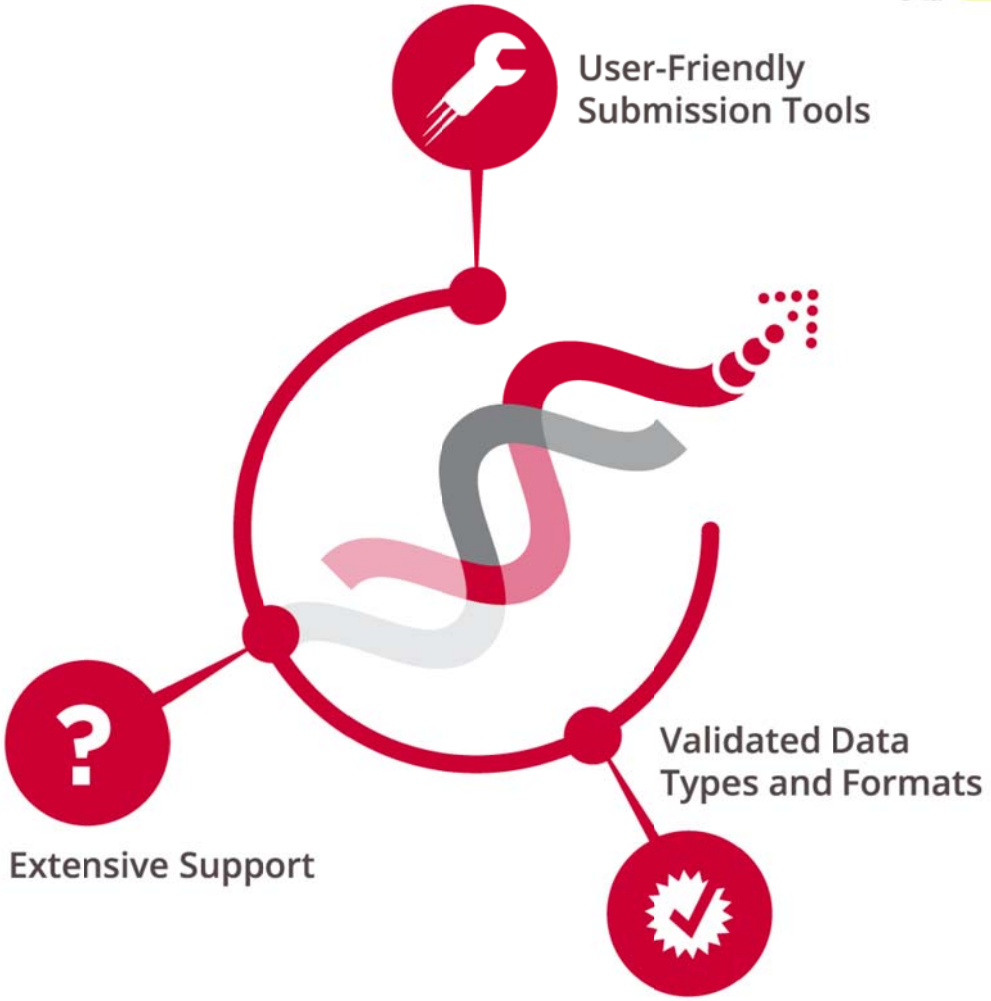
GDC: Mission and Goals

The mission of the GDC is to provide the cancer research community with a **unified data repository** that enables **data sharing** across **cancer genomic studies** in support of **precision medicine**

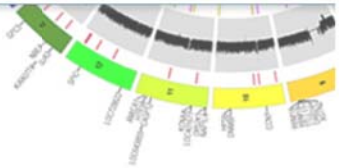
- Provide a cancer knowledge network that:
  - Enables the identification of both high- and low-frequency cancer drivers
  - Assists in defining genomic determinants of response to therapy
  - Informs the composition of clinical trial cohorts sharing targeted genetic lesions
- Support the *receipt*, *quality control*, *integration*, *storage*, and *redistribution* of standardized genomic data sets derived from cancer research studies
  - Harmonization of raw sequence both from existing (e.g. TCGA, TARGET, CGCI) and new cancer research programs
  - Application of state-of-the-art methods of generating high level data

GDC Functions

- GDC Data Submission
- GDC Data Processing
- GDC Data Retrieval

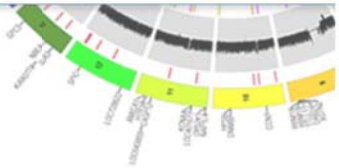


GDC Data Submission: Data Submitter Types



- The GDC supports two types of data submitters:
  - Type 1: Large Organizations
    - Users associated with an institution or group with significant informatics resources
    - Large one-time submission or long-term ongoing data submission
    - Primarily use the GDC Application Programming Interfaces (API) or the GDC Data Transfer Tool (command line interface) for data submission
  - Type 2: Researchers or Individual Laboratories
    - Users associated with a single group (such as a laboratory investigator or researcher), with limited informatics resources
    - One-time or sporadic uploads of low volumes of patient and analysis data with varying levels of completeness
    - Submit via the web-based GDC Data Submission Portal and the GDC Data Transfer Tool that use the GDC API

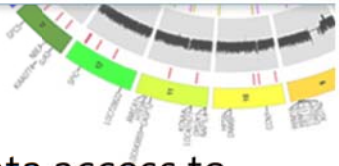
GDC Data Submission: Data Submission Policies



- dbGaP Data Submission Policies
  - GDC data submitters must first apply for data submission authorization through dbGaP
  - Data submission through dbGaP requires institutional certification under NIH's Genomic Data Sharing Policy

- GDC Data Sharing Policies
  - Data Sharing Requirement
    - Data submitted to the GDC will be made available to the scientific community at large according to the data submitter's NCI Genomic Data Sharing Plan. Controlled access data will be made available to members of the community having the appropriate dbGaP Data Use Certification.
    - The GDC will produce harmonized data (raw and derived) based on the originally submitted data. **The GDC will not preserve an exact copy of the originally submitted data**

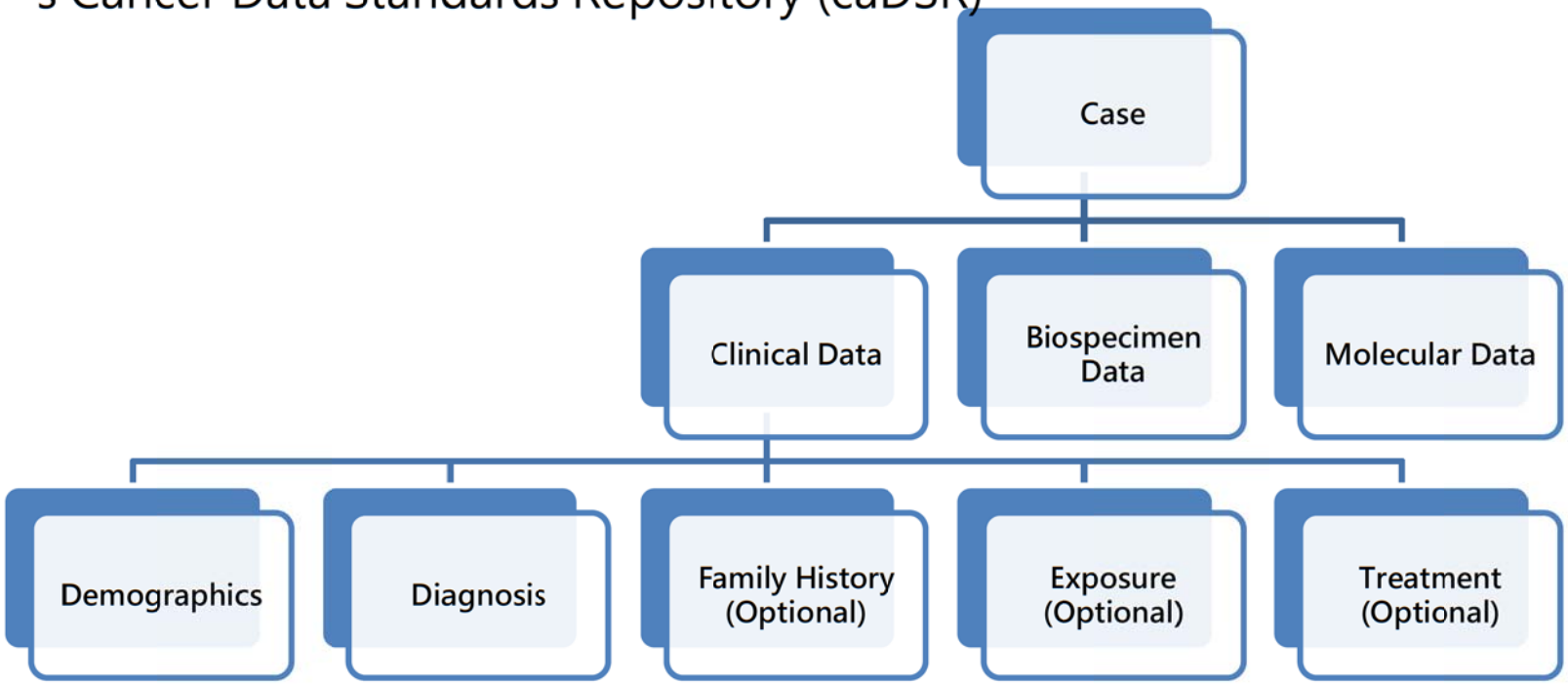
GDC Data Submission: Data Submission Policies



- Data Pre-processing Period
  - For each project, the GDC will afford a pre-processing period of exclusive data access to submitters and their named collaborators. The pre-processing period allows for submitters to perform data cleaning and quality, and submission of revised data before public release.
  - The pre-processing period may generally last up to six months from the date of first submission
- Data Submission Period and Release
  - Once submitted, data will be processed and validated by the GDC. Submitted data will be released and available via controlled access for research that is consistent with the dataset' s "data use limitations" either six months after data submission or at the time of first publication
- Data Redaction
  - The GDC in general will not remove data access in response to submitter requests. GDC will remove data access in the following events: *Data Management Incident, Human Subjects Compliance Issue, Erroneous Data.*

Data Types and File Formats: Clinical Data

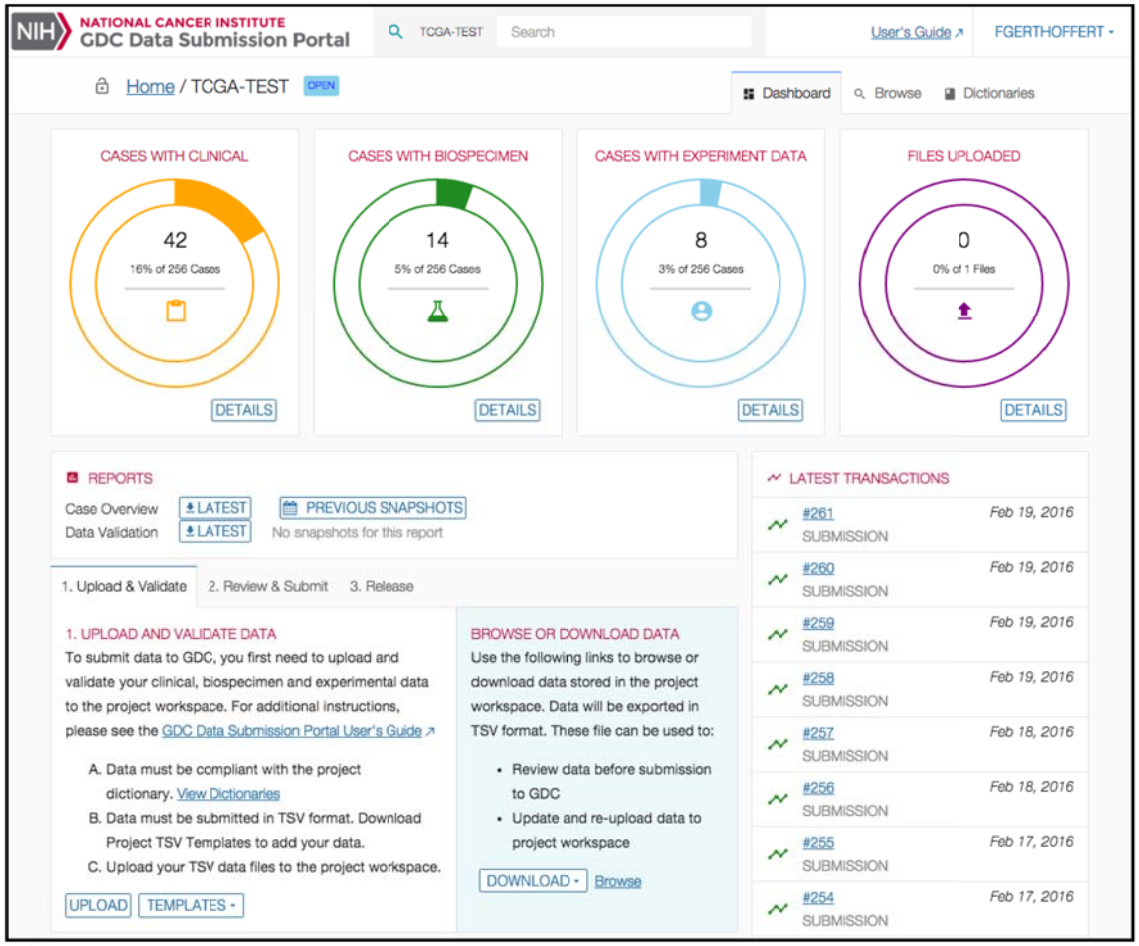
- GDC clinical data types are associated with each case and include required, preferred, and optional clinical data elements for demographics, diagnosis, family history, exposure, and treatment
  - GDC clinical data elements were reviewed with members of the research community
  - Clinical data elements are defined in the GDC dictionary and registered in the NCI's Cancer Data Standards Repository (caDSR)

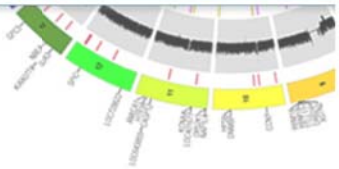


Data Submission Tools: GDC Data Submission Portal

- The GDC Data Submission Portal is a web-based data-driven platform that allows users to validate and submit biospecimen, clinical, and molecular data and metadata

- ✓ Upload clinical, biospecimen, and molecular data using user friendly web-based tools
- ✓ Validate data against GDC standard data types defined in the project data dictionary
- ✓ Obtain information on the status of data submission and processing by project





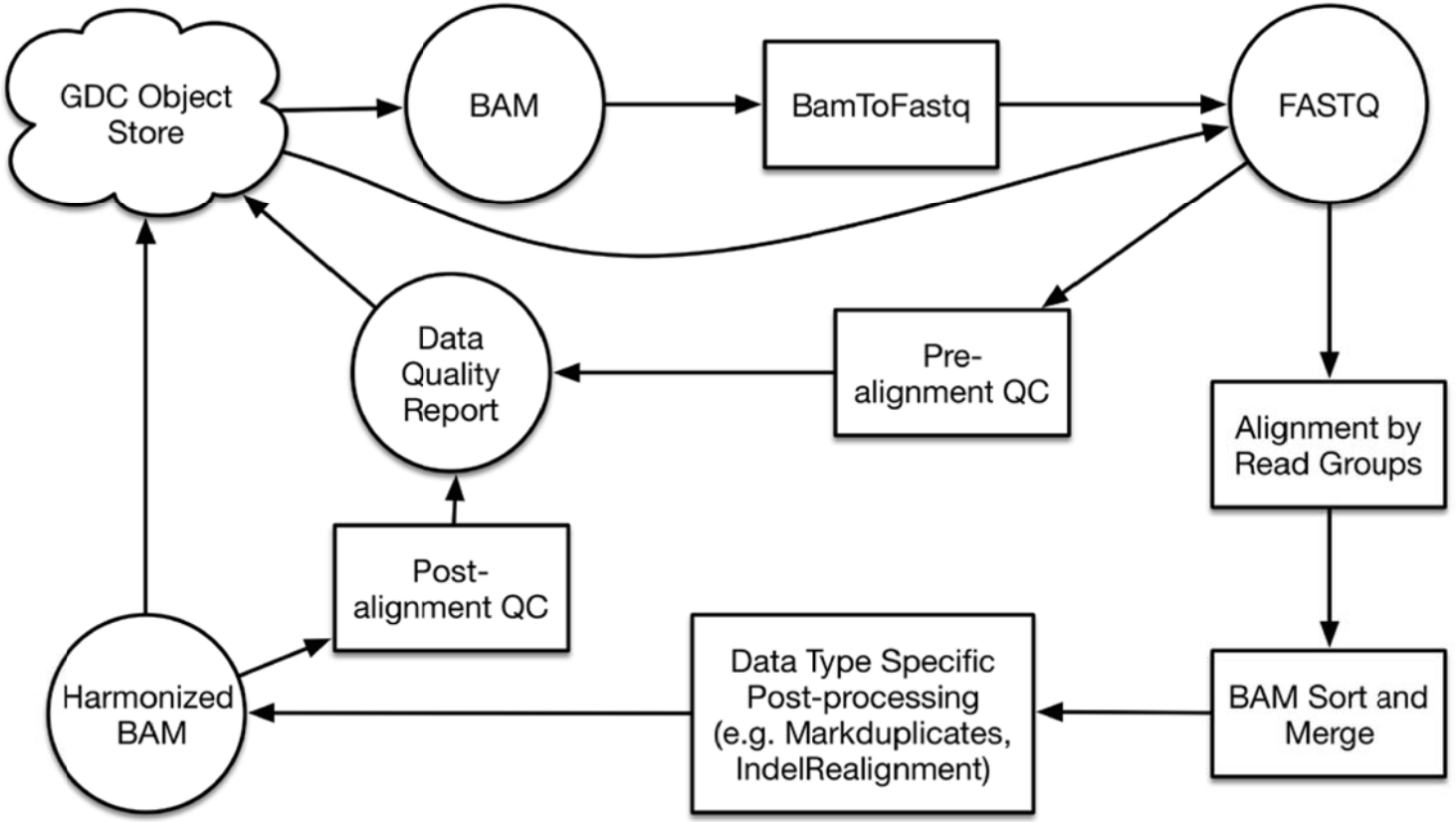
GDC Data Submission

GDC Data Processing

GDC Data Retrieval

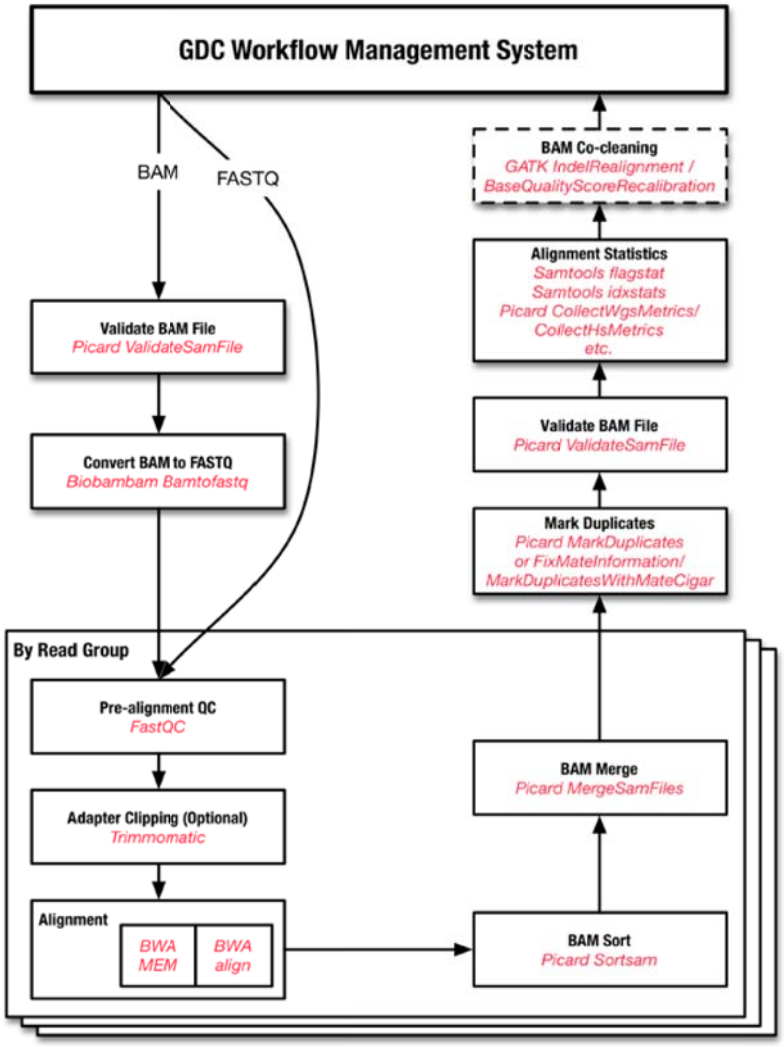
GDC Processing: Harmonization

- GDC pipelines supporting the harmonization of DNA and RNA sequence data against the latest genome build (GRCh38)

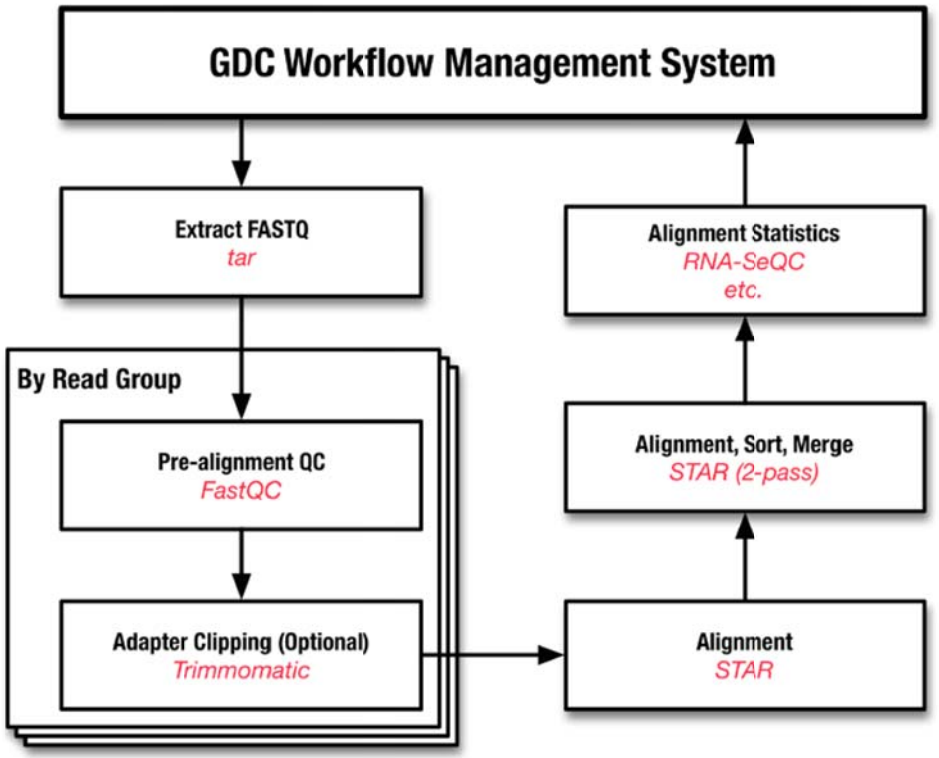


GDC Processing: DNA and RNA Sequence Harmonization Pipelines

DNA Sequence Pipeline

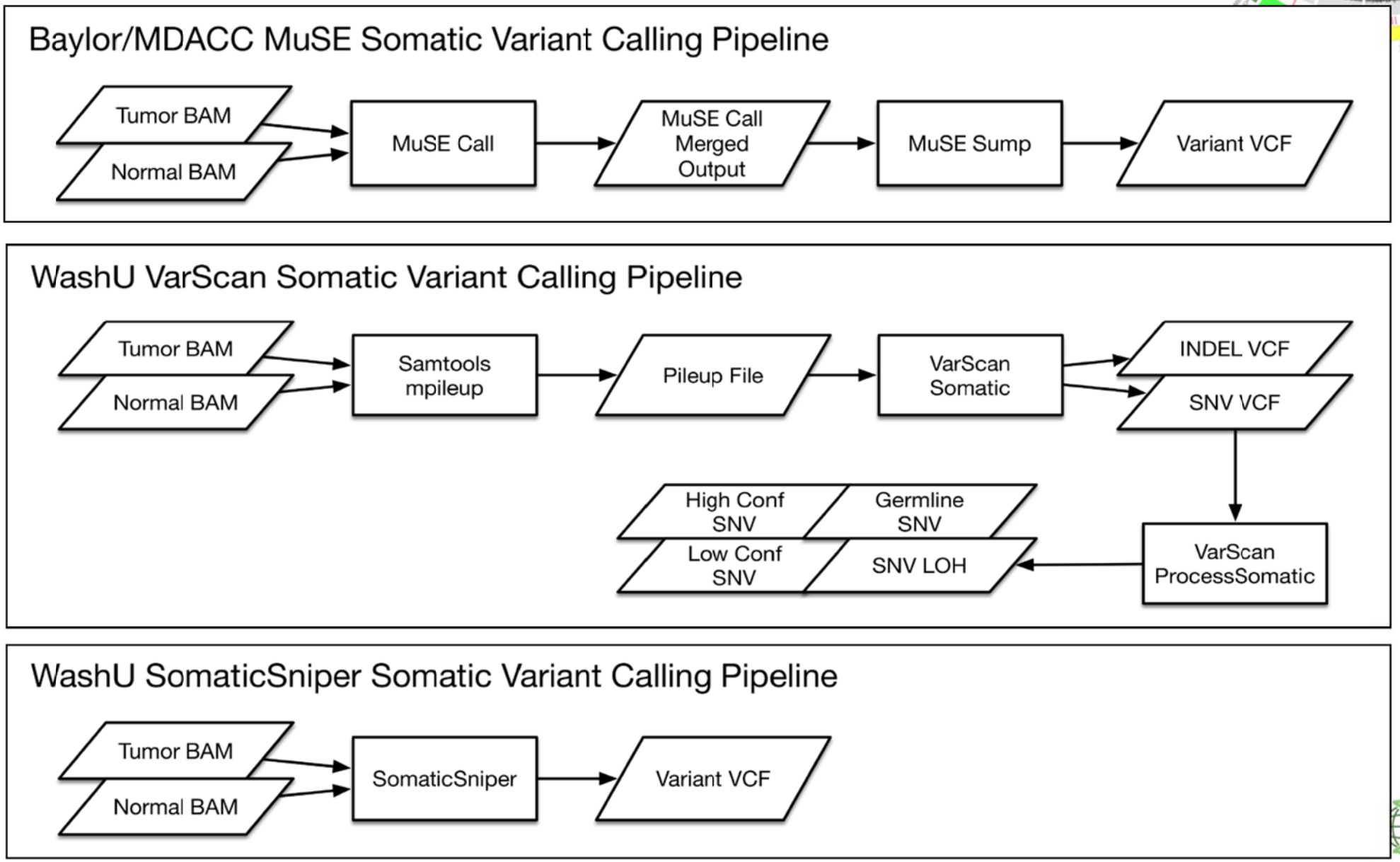


RNA Sequence Pipeline



The Cancer Genome Atlas

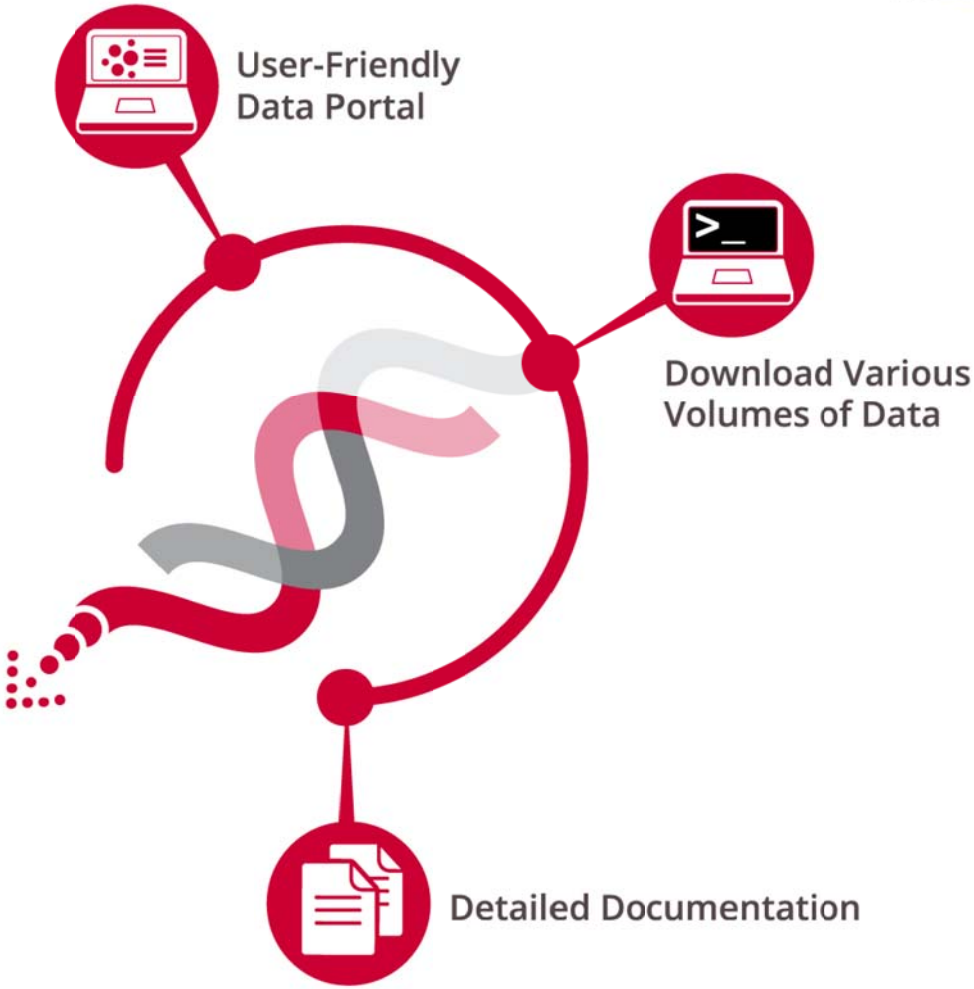
GDC Processing: GDC Variant Calling Pipelines



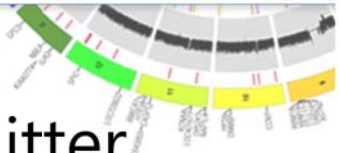
GDC Data Submission

GDC Data Processing

GDC Data Retrieval



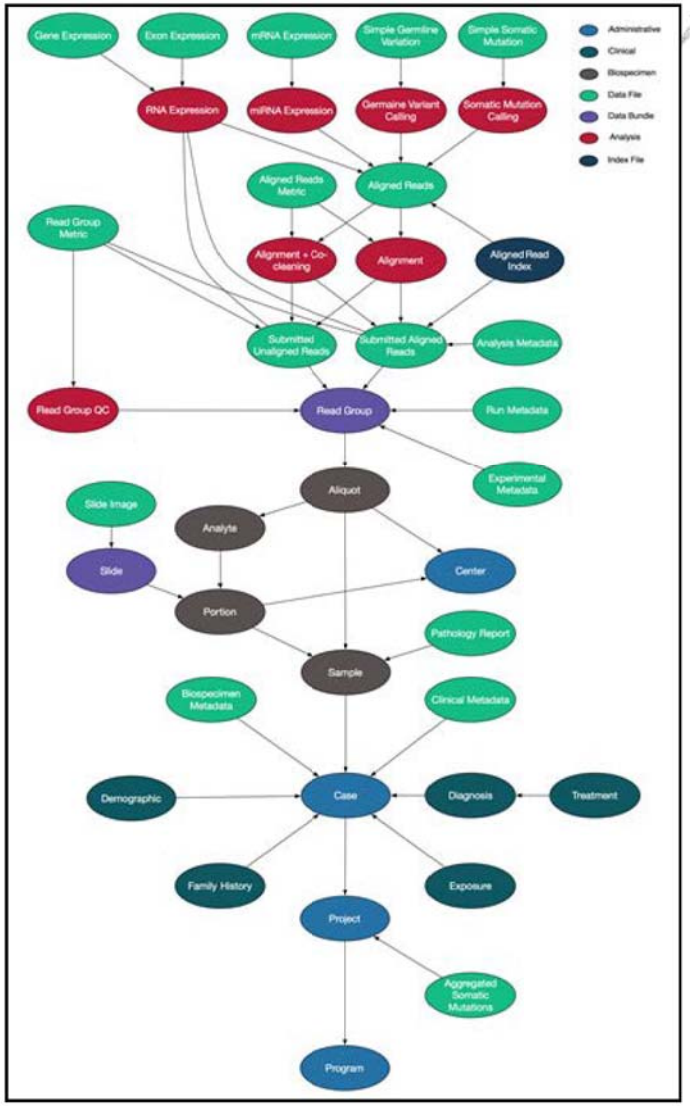
GDC Data Retrieval



- Once data is submitted into GDC and the data submitter signs-off on the data to request data release, GDC performs Quality Control (QC) and data processing
- Upon successful GDC QC and processing, the data is released based on the appropriate dbGaP data restrictions
- Released data is made available for query and download to authorized users via the GDC Data Portal, the GDC Data Transfer Tool, and the GDC API
- Data queries are based on the GDC Data Model

GDC Data Retrieval: GDC Data Model

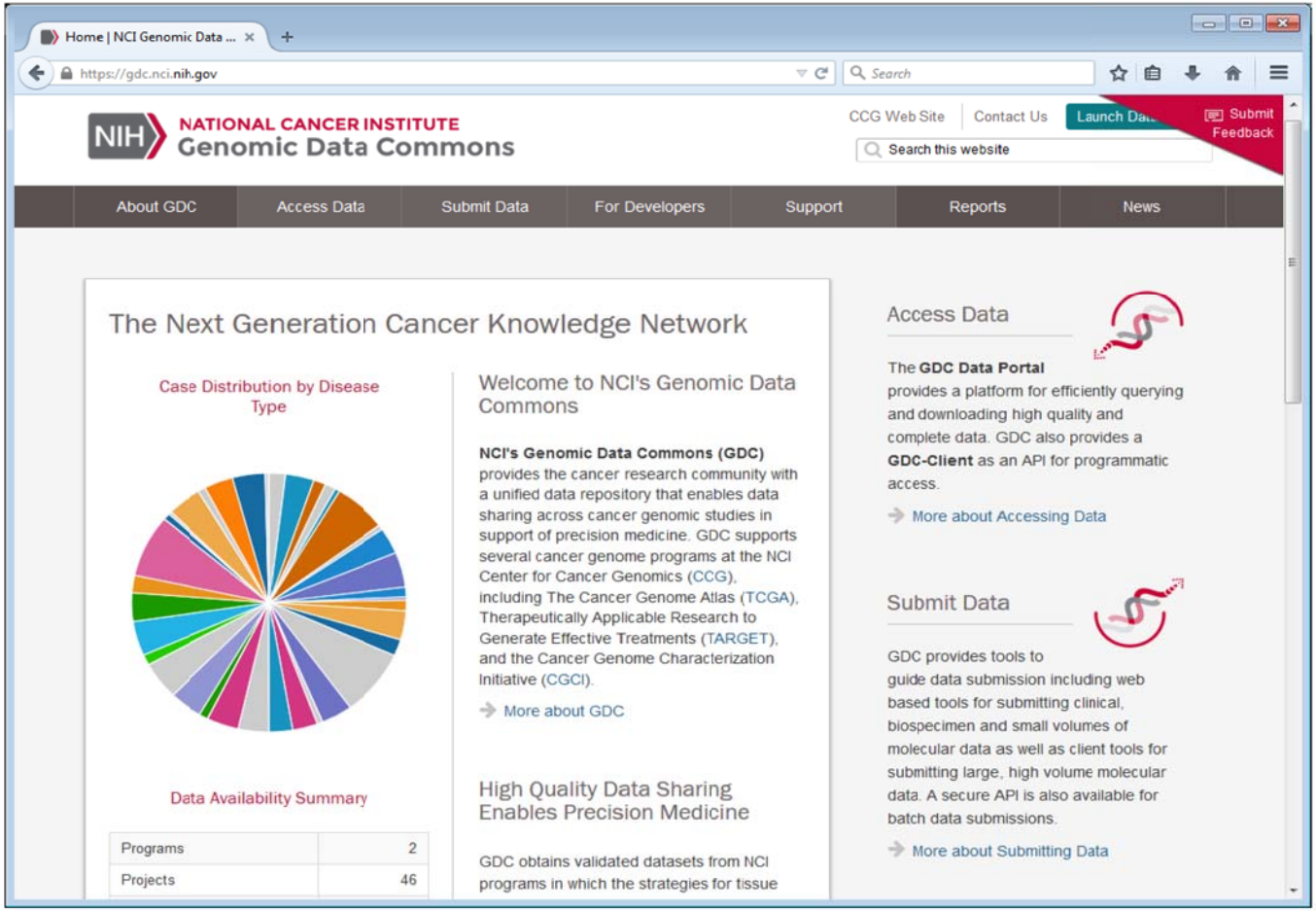
- The GDC data model is represented as a graph with nodes and edges, and this graph is the store of record for the GDC
- The GDC data model maintains the critical relationship between projects, cases, clinical data and molecular data and insures that this data is linked correctly to the actual data file objects themselves, by means of unique identifiers



GDC Data Retrieval: GDC Data Portal

- The GDC Data Portal is a web-based platform that allows users to search and download cancer data sets for analysis, and provides access to GDC reports on data statistics

- ✓ Data browsing by project, file, case, or annotation
- ✓ Visualization allowing users to perform fine-grained filtering of search results
- ✓ Data search using advanced smart search technology
- ✓ Data selection into a personalized cart
- ✓ Data download from cart or a High-performance Data Transfer Tool



References

- GDC Web Site
  - https://gdc.nci.nih.gov
- GDC Documentation Site
  - https://gdc-docs.nci.nih.gov
- GDC Data Portal
  - https://gdc-portal.nci.nih.gov
- GDC Data Submission Portal
  - https://gdc-portal.nci.nih.gov/submission/
  - https://gdc.nci.nih.gov/submit-data/gdc-data-submission-portal
- GDC Data Transfer Tool
  - https://gdc.nci.nih.gov/access-data/gdc-data-transfer-tool
- GDC Application Programming Interface (API)
  - https://gdc.nci.nih.gov/developers/gdc-application-programming-interface-api

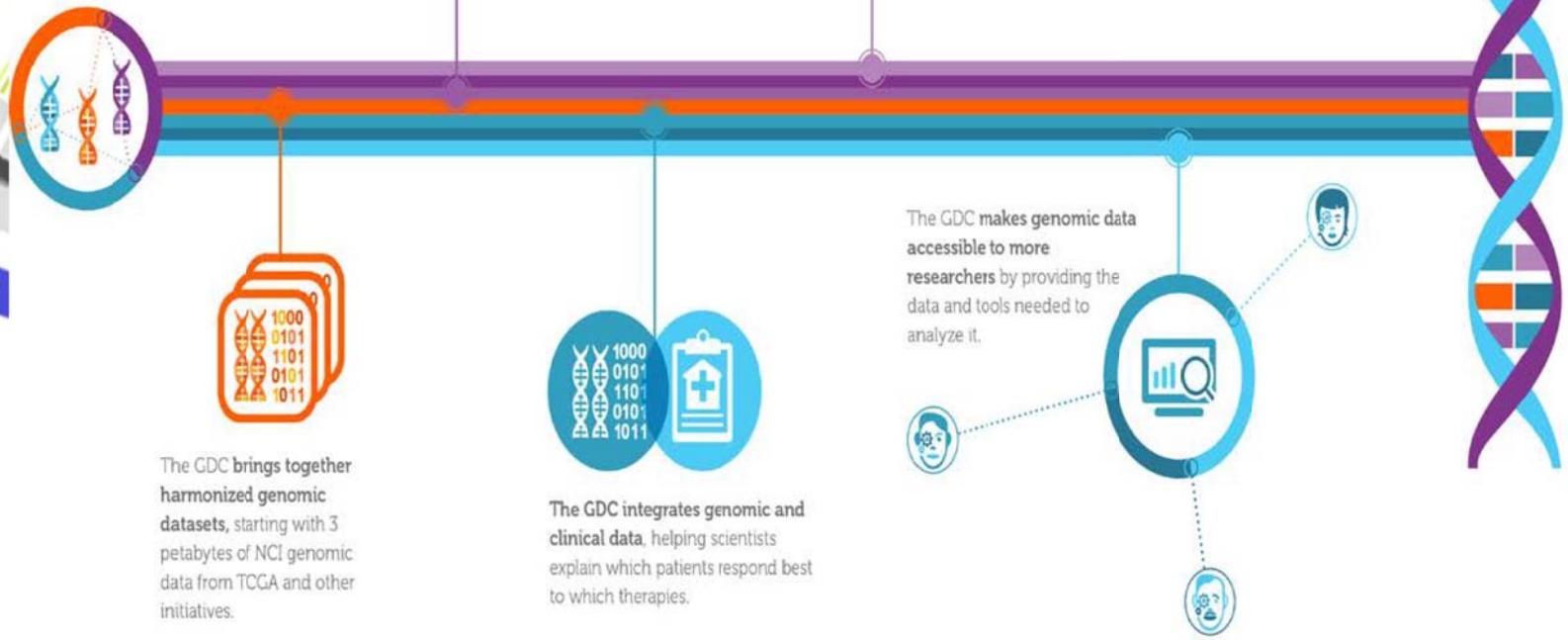
Note: Requires access to the University of Chicago Virtual Private Network

NATIONAL CANCER INSTITUTE
GENOMIC DATA COMMONS

The NCI Genomic Data Commons (GDC) is a knowledge base for cancer that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

Researchers are encouraged to submit their data. The GDC will harmonize all incoming data.

The GDC will offer an **interactive, cloud-based knowledge system** that includes cutting-edge bioinformatics tools. Researchers will be able to compare their own data with GDC data in the cloud without downloading a single file.



The GDC **brings together harmonized genomic datasets,** starting with 3 petabytes of NCI genomic data from TCGA and other initiatives.

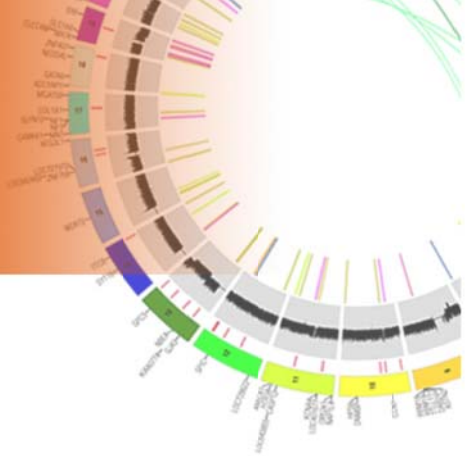
The GDC integrates genomic and clinical data, helping scientists explain which patients respond best to which therapies.

The GDC **makes genomic data accessible to more researchers** by providing the data and tools needed to analyze it.

Expanding access to genomic and clinical data will accelerate cancer research and help improve the diagnosis and treatment of each cancer patient.

www.cancer.gov

Future Programs



- There will be no TCGA 2.0
- A variety of new programs have started or are about to start (ALCHEMIST, Exceptional Responders, CDDP, CTSP) that center on discovering the reasons why some patients respond better than others to therapy and learn more about the molecular underpinnings on the way.
- The big message is cancer genomics is here to stay, but in the context of helping decipher how to better treat the patient.

Acknowledgements

TCGA NCI Office

Amy Blum
Samantha Carter-Johnson
John Demchok
Ina Felau
Martin Ferguson
Roy Tarnuzzer
Zhining Wang
Liming Yang

TCGA NHGRI Office

Carolyn Hutter

Elian Silverman
Heidi Sofia

University of Chicago

Robert Grossman

Allison Heath
Joshua Miller
Zenyhu Zhang
Michael Ford

Ontario Institute of Cancer Research

Vincent Ferretti
François Gerthoffert

Leidos

Sharon Gaheen
Mark Jensen
Himanso Sahni