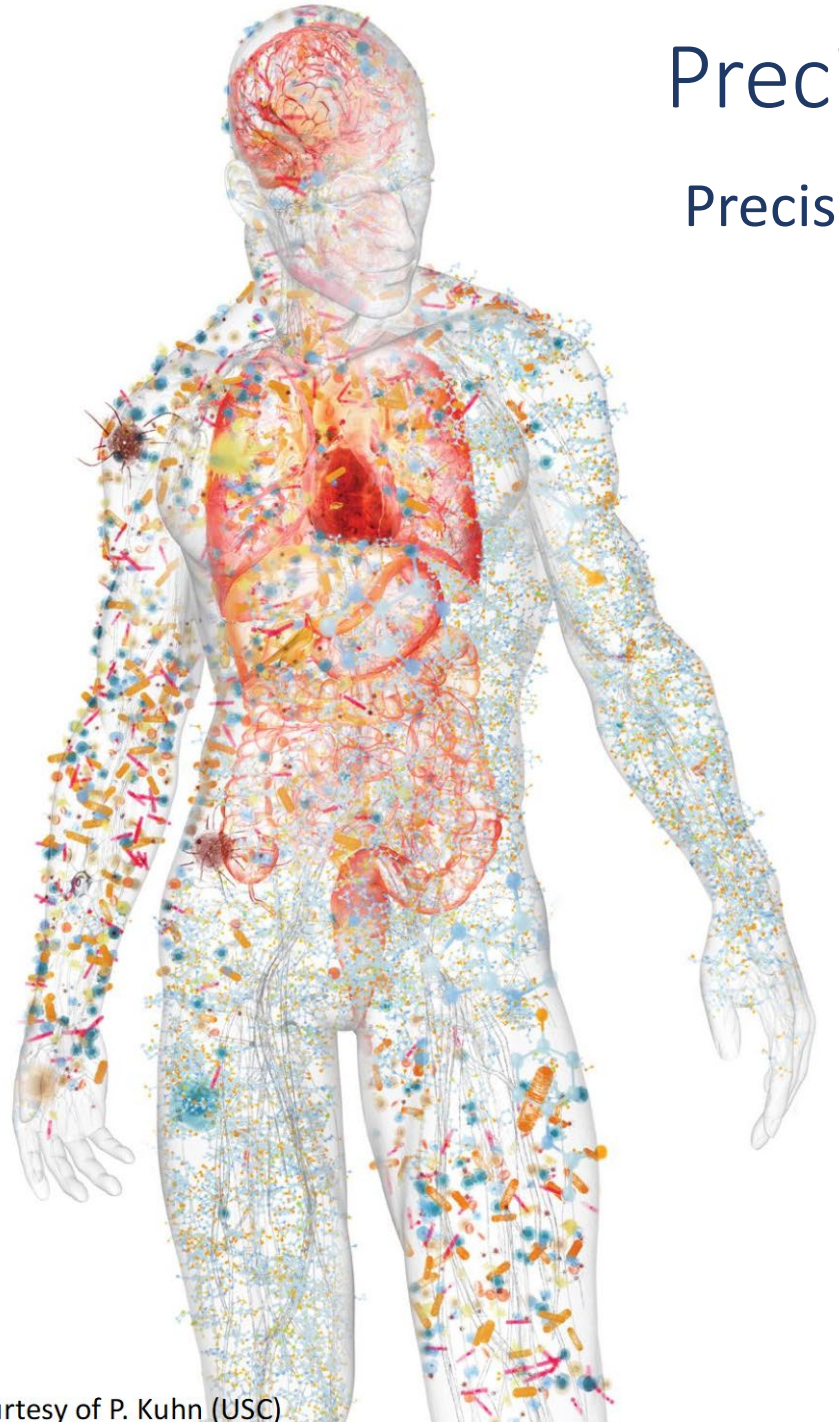
NCI Cancer Research Data Commons (CRDC) and ITCR

Tanja Davidsen, Ph.D.
Center for Biomedical Informatics and Information Technology

ITCR Meeting
May 29, 2019

NIH NATIONAL CANCER INSTITUTE



Courtesy of P. Kuhn (USC)

Precision Medicine Initiative (PMI)

Precision Medicine is a grand challenge, requiring:

- Deep biological understanding
- Advances in scientific methods, instrumentation, and technology
- Advances in data management and computation
- Ability to apply those advances to drive research and treatment
- Ability to securely share data across domains, institutions, and stakeholders

*Cancer research and care generate detailed **data** that are critical to create a learning health system for cancer*

Key tenet of the PMI: secure, responsible access to high-quality data

The PMI was announced during the State of the Union Address, 2015

The Beau Biden Cancer Moonshot[sm]

Overarching goals – Jan, 2016

- Accelerate progress in cancer, including prevention & screening
  - From cutting edge basic research to wider uptake of standard of care
- Encourage greater cooperation and collaboration
  - Within and between academia, government, and private sector
- Enhance data sharing

Blue Ribbon Panel – October, 2016

- Network for Direct Patient Engagement
- Cancer Immunotherapy Translational Science Network
- Therapeutic Target Identification to Overcome Drug Resistance
- A National Cancer Data Ecosystem for Sharing and Analysis
- Fusion Oncoproteins in Childhood Cancers
- Symptom Management Research
- Prevention and Early Detection – Implementation of Evidence-based Approaches
- Retrospective Analysis of Biospecimens from Patients Treated with Standard of Care
- Generation of 3D Human Tumor Atlas
- Development of New Enabling Cancer Technologies
- Full report: www.cancer.gov/brp

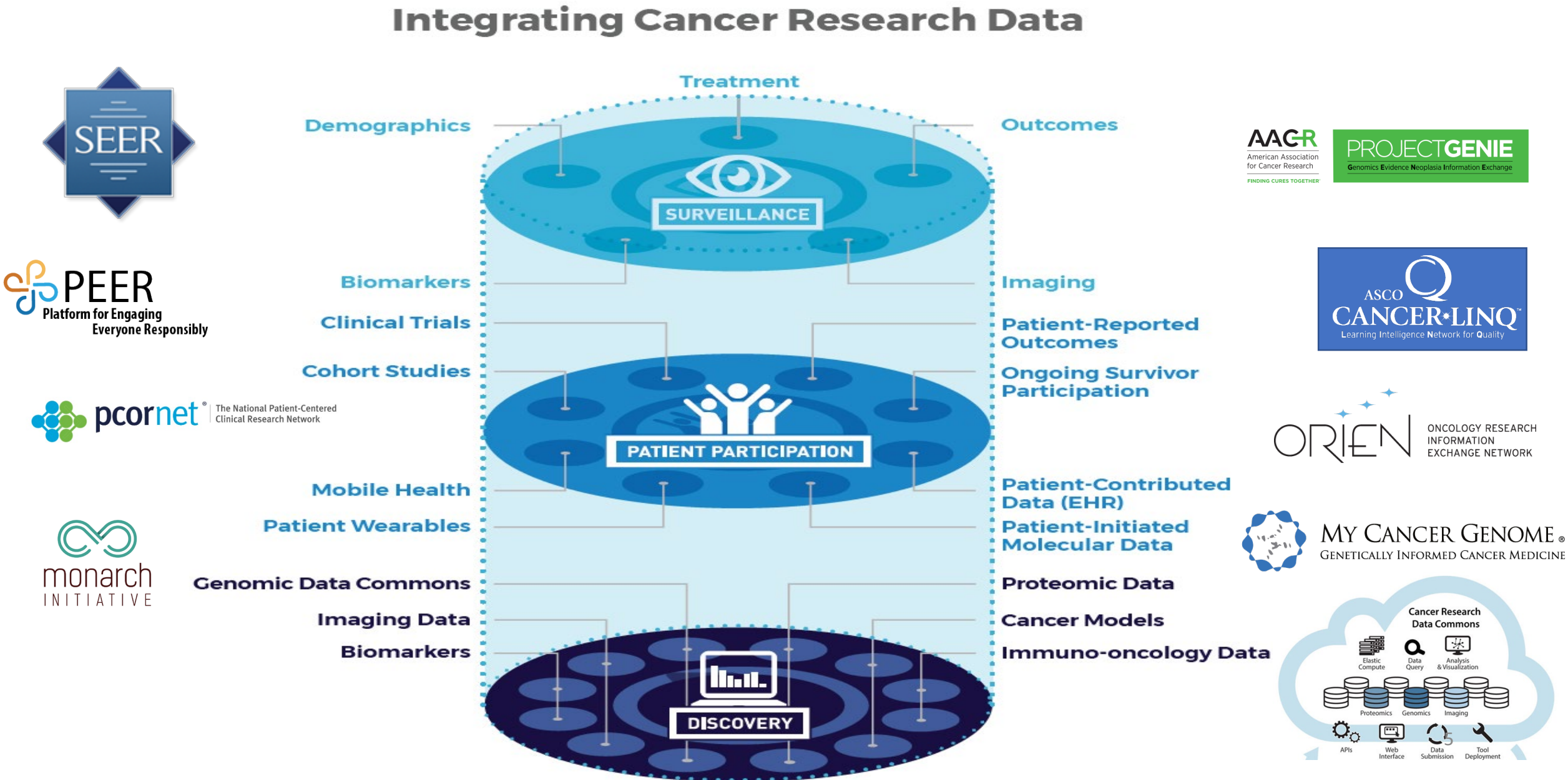
National Cancer Data Ecosystem Recommendations

Overall goal: *"Enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer."*

Recommendations

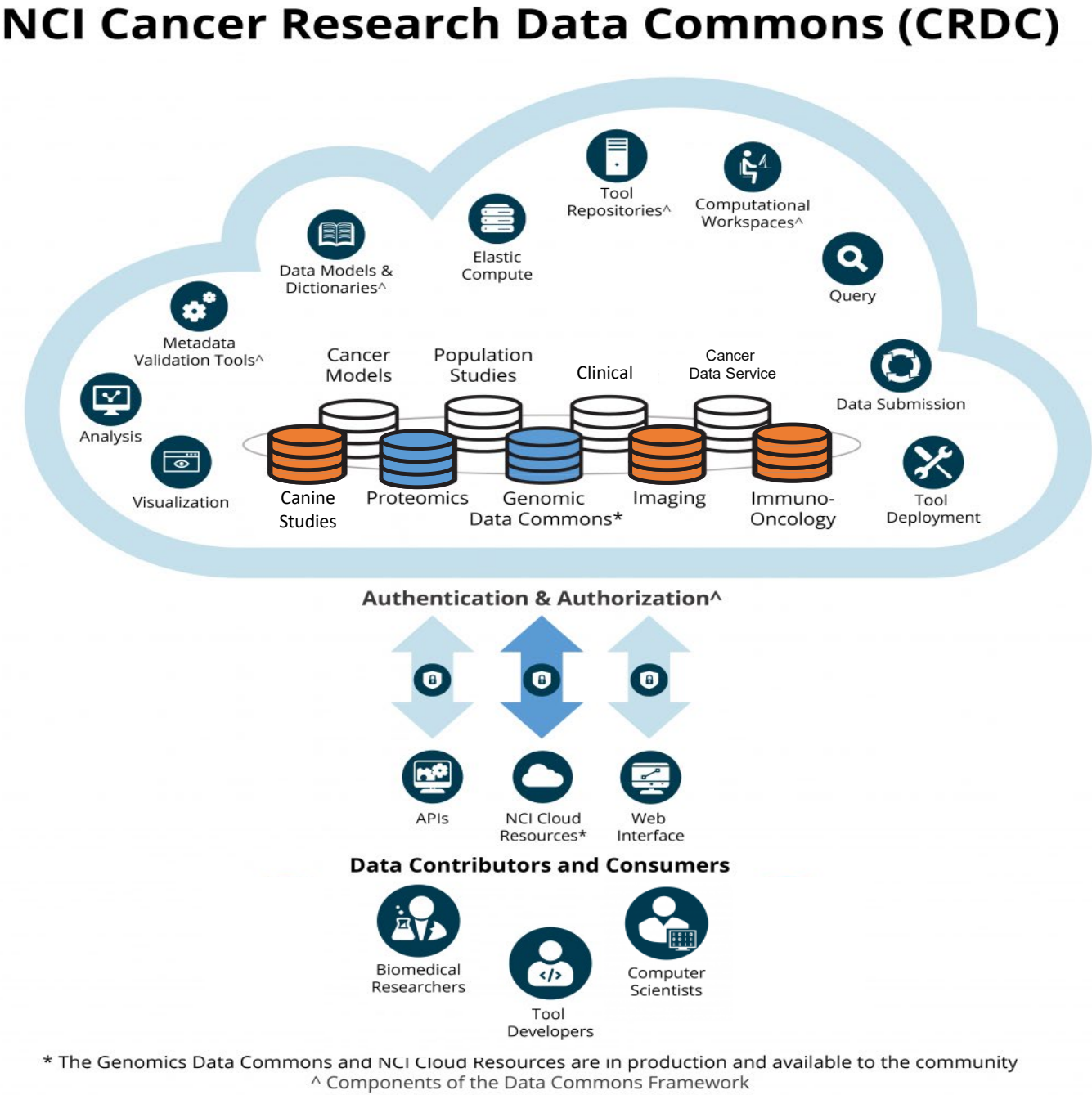
- Build a National Cancer Data Ecosystem
  - Enhanced cloud-computing platforms
  - Essential underlying data science infrastructure and portals for the Cancer Data Ecosystem
  - Services that link disparate information, including clinical, image, and molecular data
  - Develop standards and tools so that data are interoperable
  - Address sustainability and data governance to ensure long-term health of the Ecosystem
- The National Cancer Data Ecosystem is broader than NCI
  - An NCI Cancer Research Data Commons is envisioned as part of the National Cancer Data Ecosystem

National Cancer Data Ecosystem –
Integrating data from basic research through clinical care and surveillance



Cancer Research Data Commons

- Data are stored in domain-specific repositories, called Data Nodes
- The CRDC contains all the resources needed for searching, analyzing, visualizing, and interoperating across diverse data types
- A common Authentication and Authorization mechanism secures controlled access data
- Researchers can bring their own data and tools to the cloud, and combine with the data in the CRDC for integrative analysis



Goals of the NCI CRDC

- Enable the cancer research community to share diverse data types across programs and institutions
- Provide easy access to data, regardless of where it is stored
- Provide mechanisms for innovative tool discovery, access, usage
- Help NCI Data Coordinating Centers sustain and share their data publicly
- Develop a set of reusable components - a framework - for the community to use to build interoperable data commons

Data Commons Framework – What Is It?

Reusable, expandable framework for a Data Commons

Core principles and structures

Set of modular components that can be leveraged across the CRDC

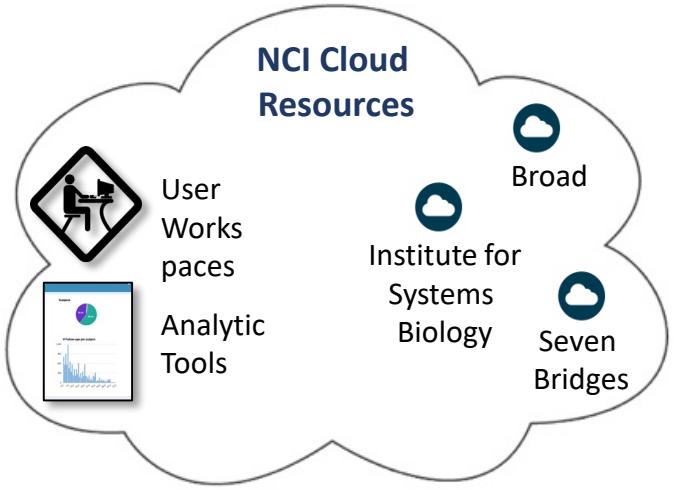
Modular Components

	Secure user authentication and authorization
	Metadata validation and tools
	Domain-specific, extensible data models and dictionaries
	API and container environment for tools and pipelines
	Access to computational workspaces for storing data, tools, and results

- The DCF will be used to stand up future Data Commons examples that the community can leverage to build their own commons
- University of Chicago building the DCF utilizing the Gen3 technology

NCI Cloud Resources

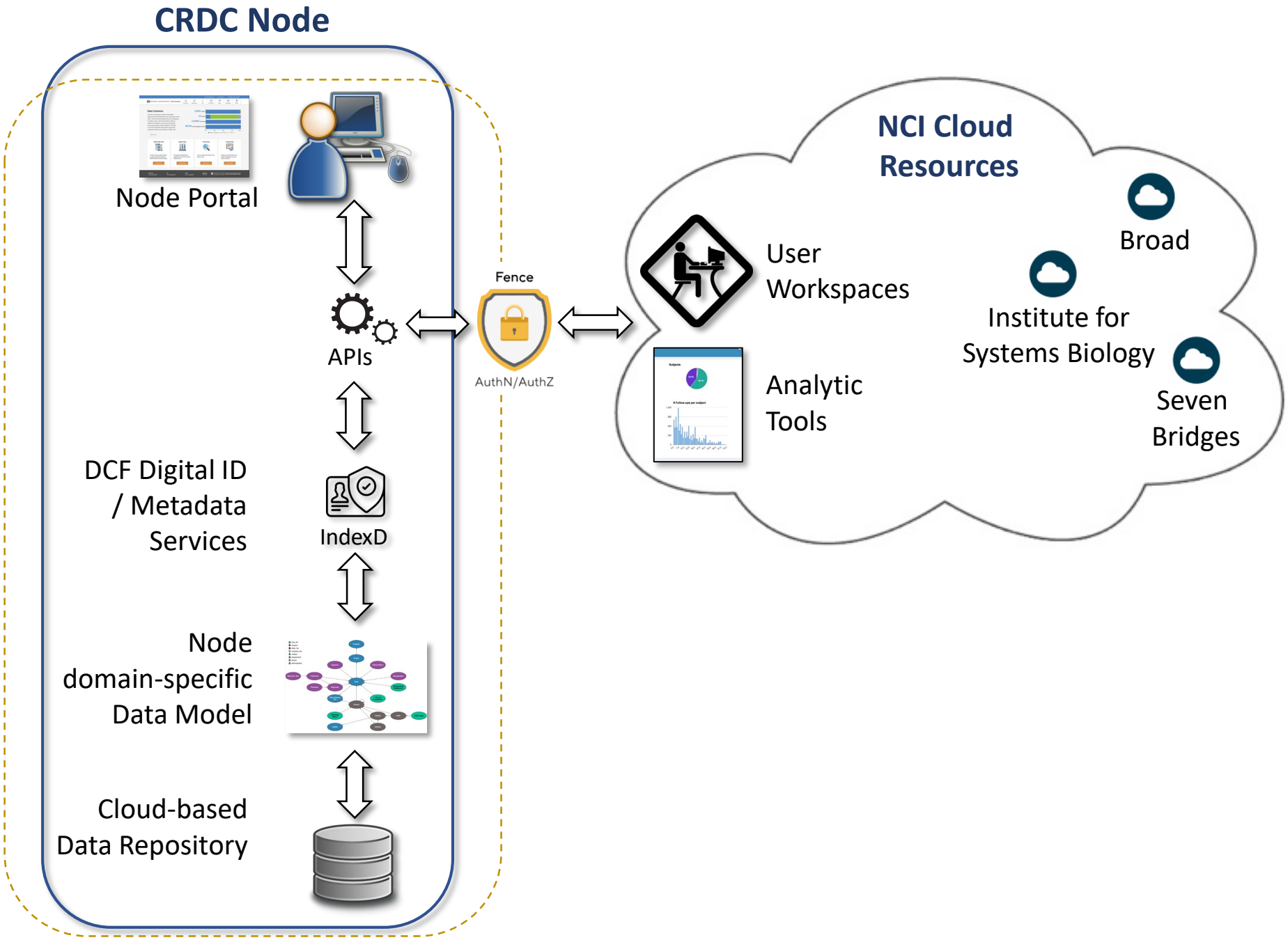
- Access to large data sets in the cloud without need to download
- Access numerous tools and pipelines within the Cloud Resources
- Bring your own tools and pipelines to the data
- Bring your own data and analyze in combination with existing data
- Workspaces, for researchers to save and share their data and results
- **Democratize access to cancer datasets and to create a cost-effective way to provide scalable computational capacity to the cancer research community**

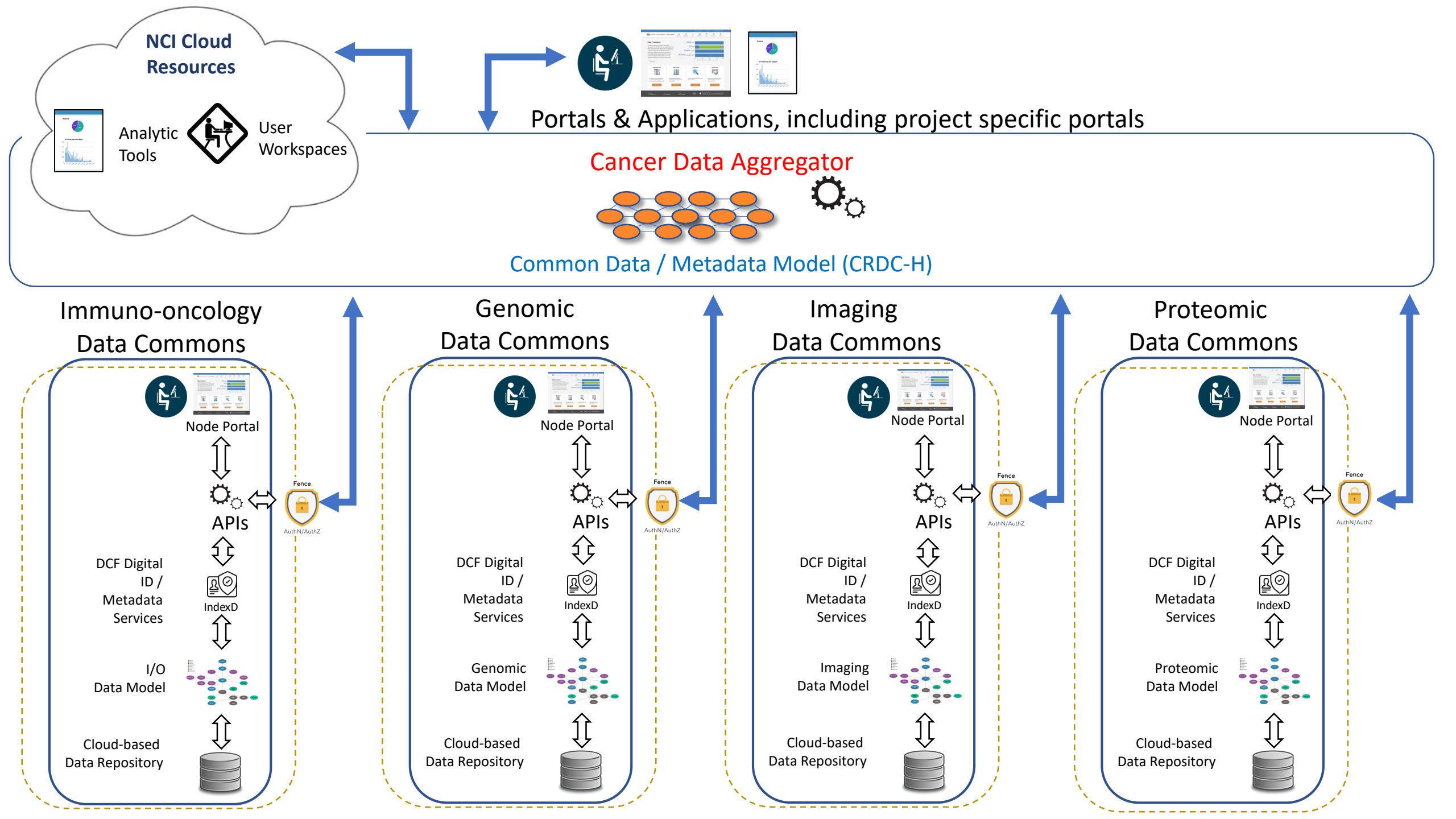


Data
- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

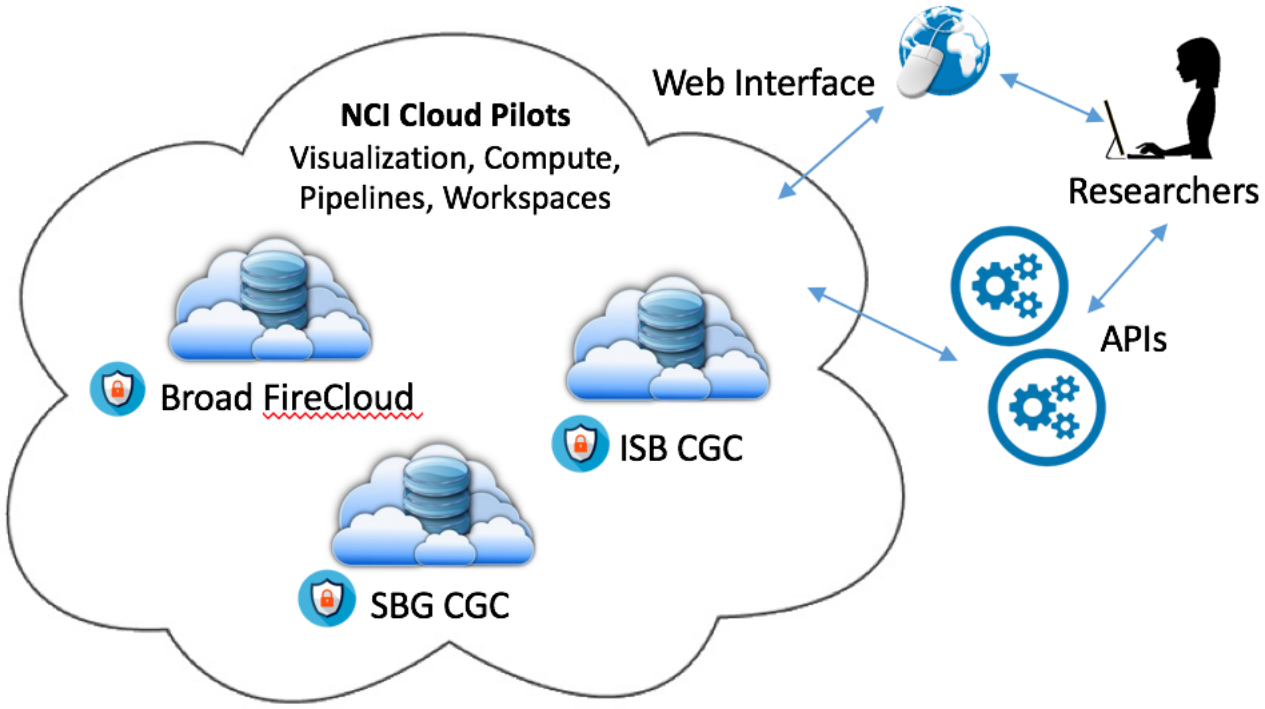
Compute
- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Security
- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines


CRDC Node
Node Portal
APIs
DCF Digital ID / Metadata Services
IndexD
Node domain-specific Data Model
Cloud-based Data Repository
Fence
AuthN/AuthZ
NCI Cloud Resources
User Workspaces
Analytic Tools
Broad
Institute for Systems Biology
Seven Bridges

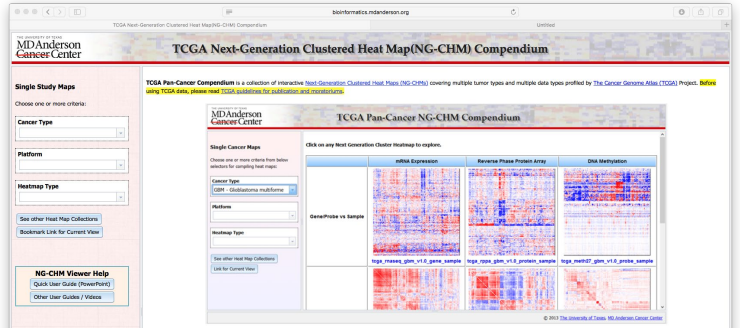

NCI Cloud Resources
Analytic Tools
User Workspaces
Portals & Applications, including project specific portals
Cancer Data Aggregator
Common Data / Metadata Model (CRDC-H)
Immuno-oncology Data Commons
Genomic Data Commons
Imaging Data Commons
Proteomic Data Commons
Node Portal
APIs
Fence
AuthN/AuthZ
DCF Digital ID / Metadata Services
IndexD
I/O Data Model
Genomic Data Model
Imaging Data Model
Proteomic Data Model
Cloud-based Data Repository

ITCR and the NCI Cloud Resources

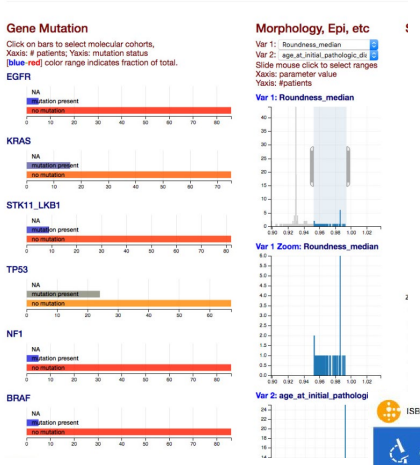


"Containerized" ITCR tools (or any containerized tools!) can be brought to the Cloud Resources

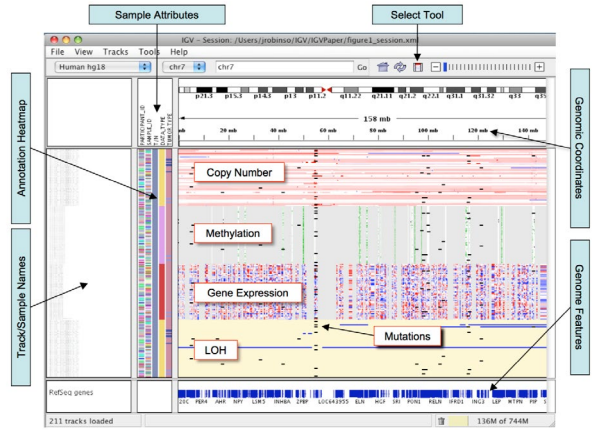
Next-Gen Clustered Heat Maps



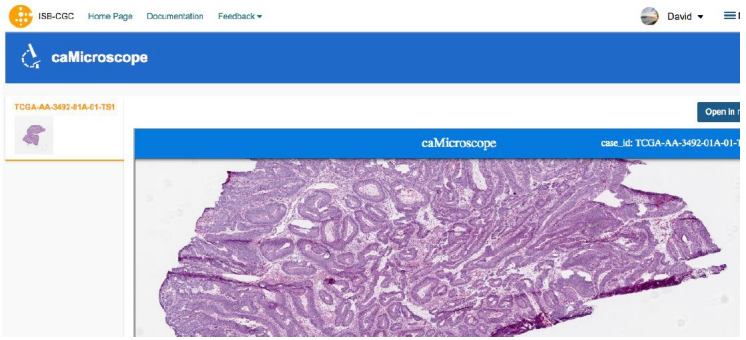
QuIP



IGV



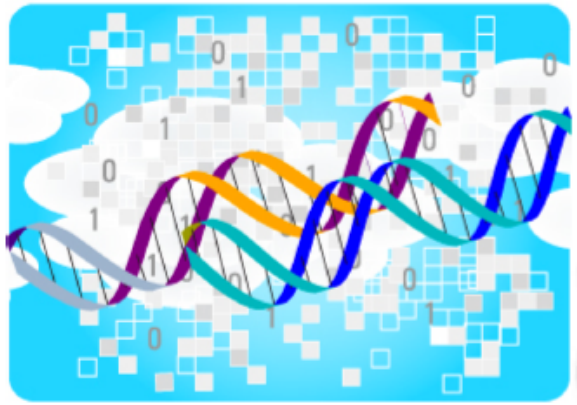
caMicroscope



Trinity Cancer Transcriptome Analysis Toolkit on FireCloud

FireCloud

Scalable Cancer Computing Solution for the NCI Cloud



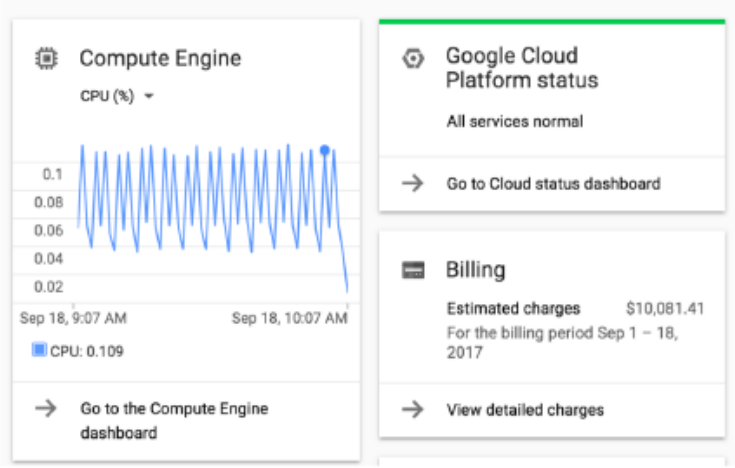
- Integration of Trinity CTAT into Docker and WDL workflows
- Process TCGA data
- Shareable workflows and data resources
- Data security

ITCR PI: Aviv Regev, Broad

Thank you NCI Cloud Credits Program!

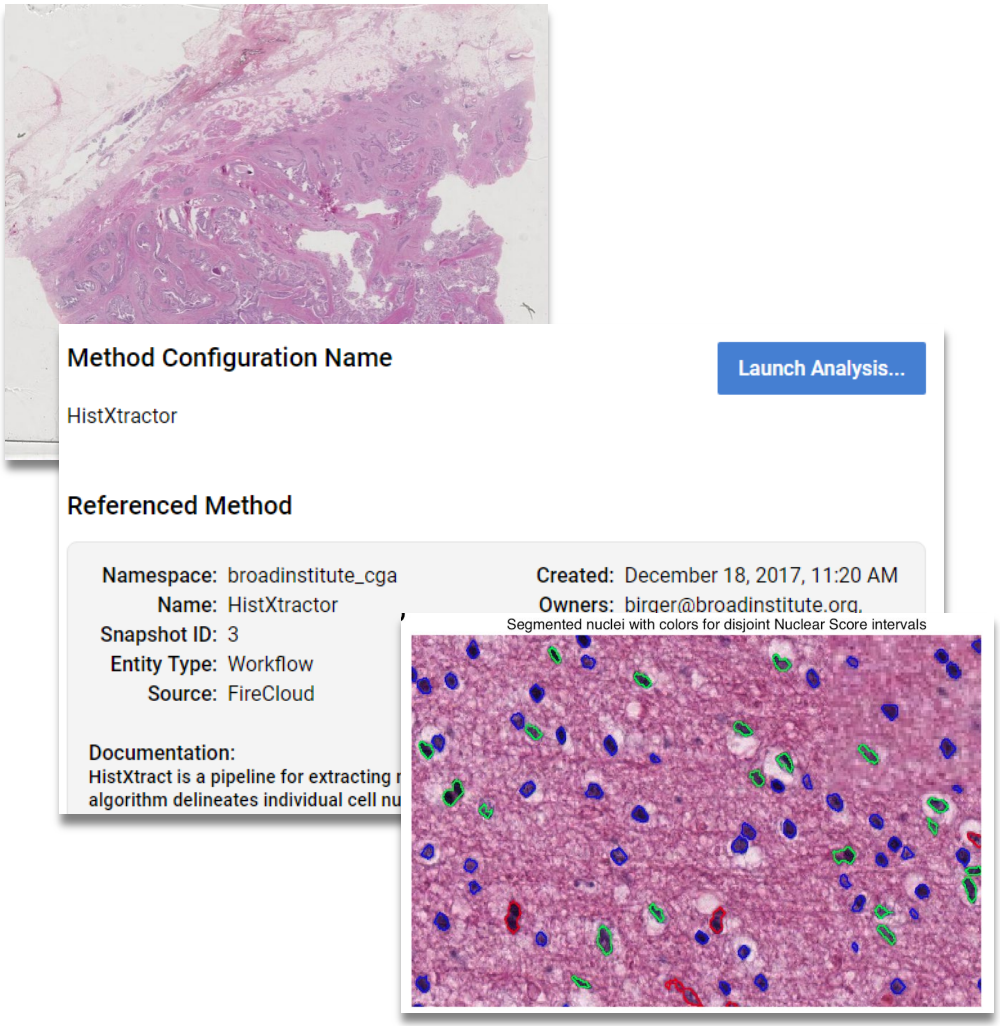
~5k samples processed in ~2 days on FireCloud

Total cost: ~$10k, so ~$2 / sample



Slide courtesy of Brain Haas, Broad Institute

Extracting nuclear morphometry features on FireCloud



Running HistXtract on TCGA diagnostic images in just a few clicks

HistXtract is a pipeline for extracting nuclear morphometry features from whole-slide images.

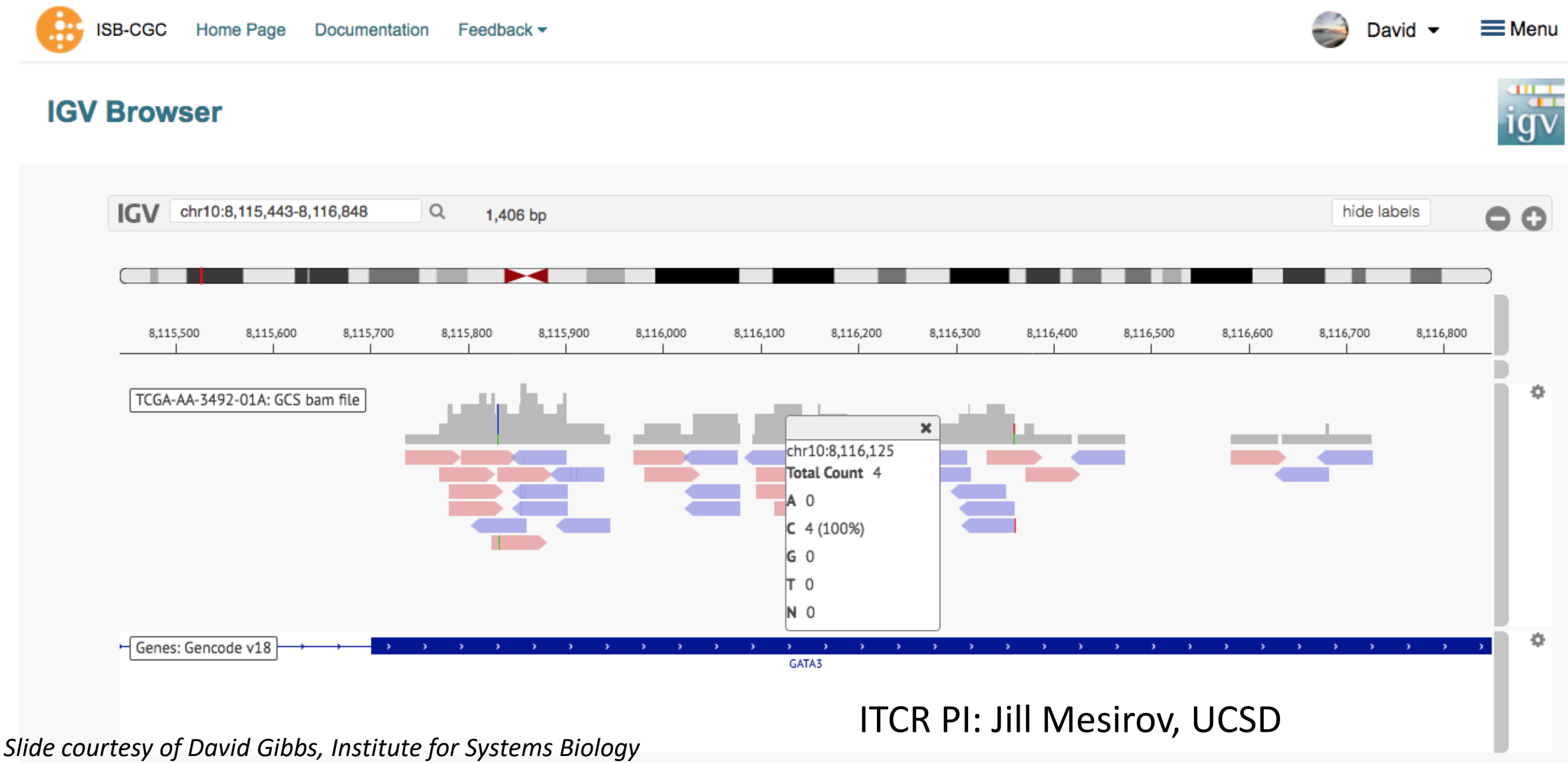
Members of the Getz Lab created an open-access FireCloud workspace preconfigured to download and analyze FFPE images for 9,600 participants across 32 types of cancer.

In just two steps, any FireCloud user can download the available images and run the HistXtract analysis workflow for some or all participants.

ITCR PI: Lee Cooper, Emory

Slide courtesy of David Siedzik, Broad Institute

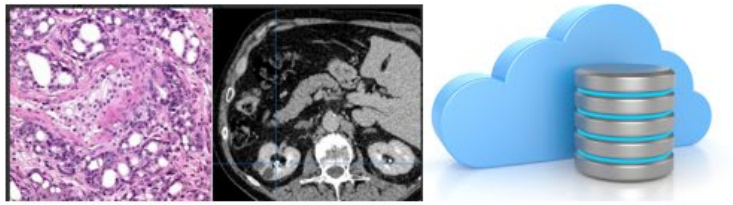
Accessing the Integrative Genomics Viewer on ISB-CGC



ITCR PI: Jill Mesirov, UCSD

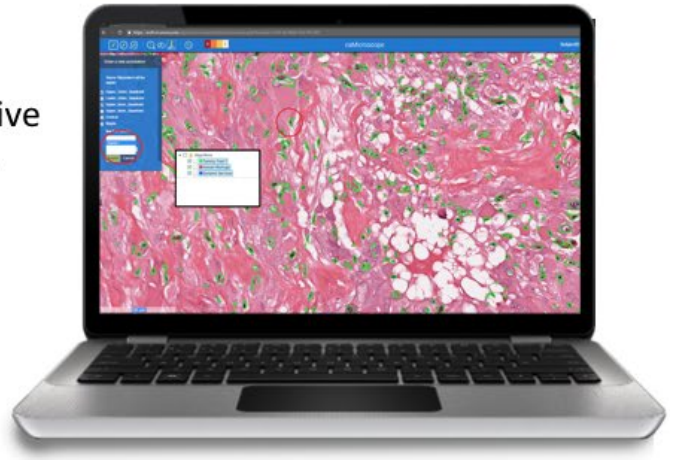
Slide courtesy of David Gibbs, Institute for Systems Biology

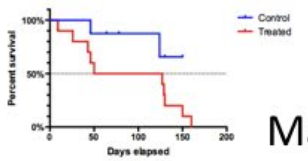
Generating Tumor Infiltrating Lymphocyte Maps on the ISB-CGC

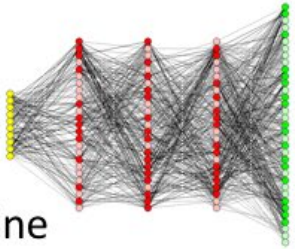


Pathology and Radiology images in Cloud Storage

Interactive Tools







Machine Learning & Integrative Analyses

Automated Segmentation & Feature Extraction

ITCR PIs: Joel Saltz, Ashish Sharma

Slide courtesy of David Gibbs, Institute for Systems Biology

Galaxy and Gen3/Data Commons Framework (DCF)

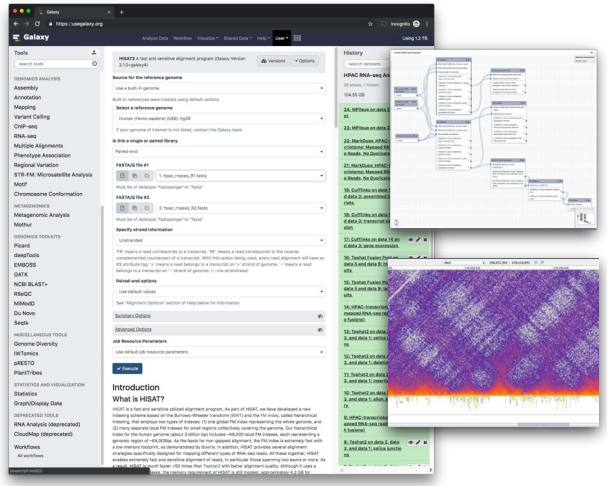
- Galaxy (https://galaxyproject.org) is a computational workbench for analysis of large biomedical datasets (e.g, genomics, proteomics)

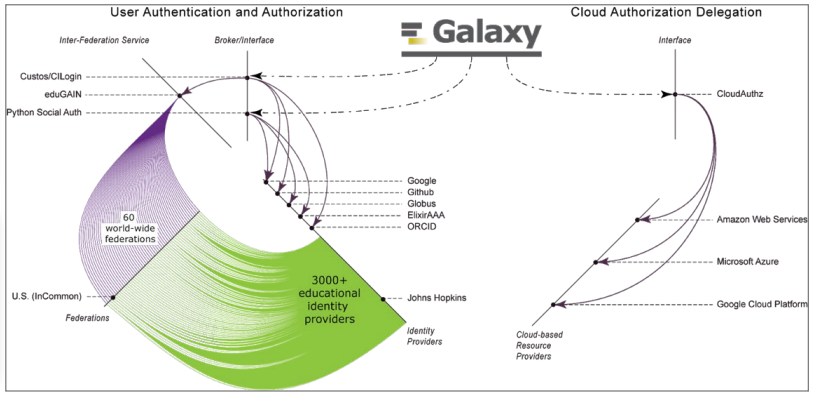
Goals:

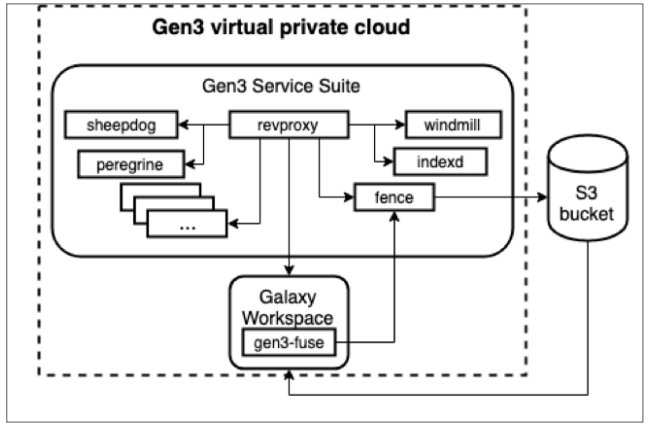
- Run Galaxy as a Gen3/DCF workspace so CRDC datasets can be analyzed securely in Galaxy
- Enable any tool/visualization integrated into Galaxy to analyze/visualize CRDC datasets in place (no copying)

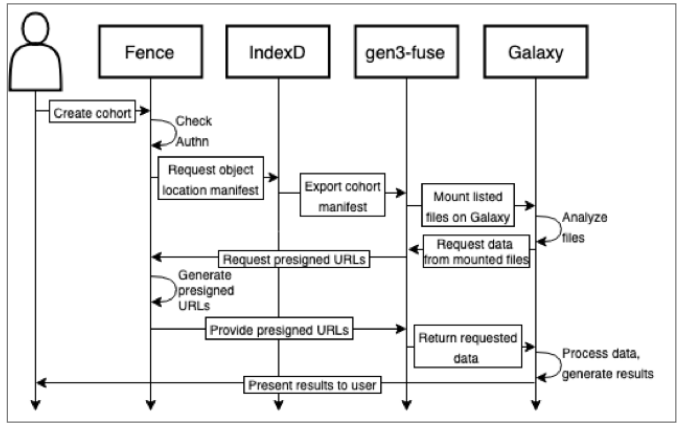
Progress:

- OIDC support in Galaxy for best-practice authentication and authorization with 3000+ entities, including Gen3
- Can use Gen3 credentials to log into Galaxy, and Galaxy stores Gen3 access tokens for future data requests
- Will use access tokens plus gen3-FUSE to access and analyze CRDC datasets in Galaxy on the cloud









Slide courtesy of Jeremy Goecks, Oregon Health and Science University

OHSU

How can you get involved?

- Check out our Cloud Resources and Data Commons Framework
  - Broad FireCloud
    - http://firecloud.org
  - Seven Bridges CGC
    - http://www.cancergenomicscloud.org
  - Institute for Systems Biology CGC
    - http://cgc.systemsbiology.net
  - University of Chicago DCF
    - http://dcf.gen3.org
- Contact me
  - tanja.davidsen@nih.gov

NIH

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol