The NCI Informatics Technology for Cancer Research (ITCR) Program and Imaging Data Commons

Stephen Jett, Ph.D.

AAAS Science & Technology Policy Fellow

NCI Center for Biomedical Informatics and Information Technology

NIH NATIONAL CANCER INSTITUTE

MICCAI
Sept 2018
Granada, España

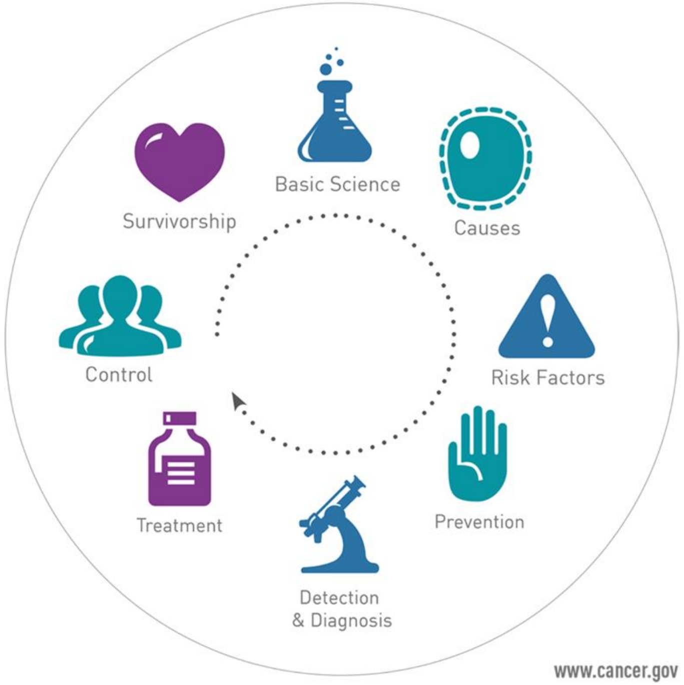
Disclosure Information
MICCAI 2018
Stephen Jett

- I have no financial relationships to disclose
- I will not discuss off label use and/or investigational use in my presentation.

The Informatics Technology for Cancer Research (ITCR) Program

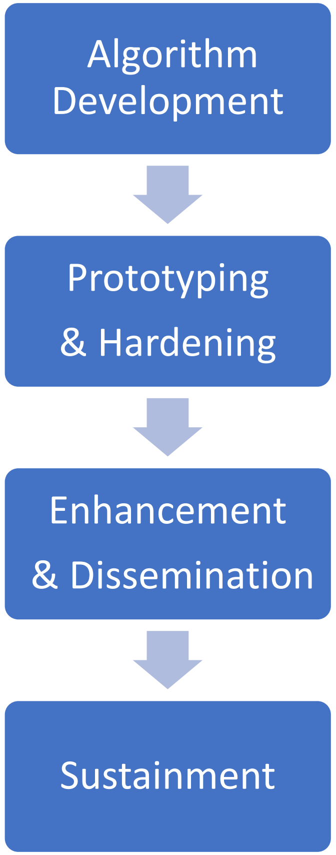
ITCR is a trans-NCI program to support investigator-initiated informatics technology development driven by critical needs in cancer research.

- Support informatics technology development driven by cancer research
- Develop open-source, interoperable software tools and resources
- Promote broad dissemination of user-friendly resources



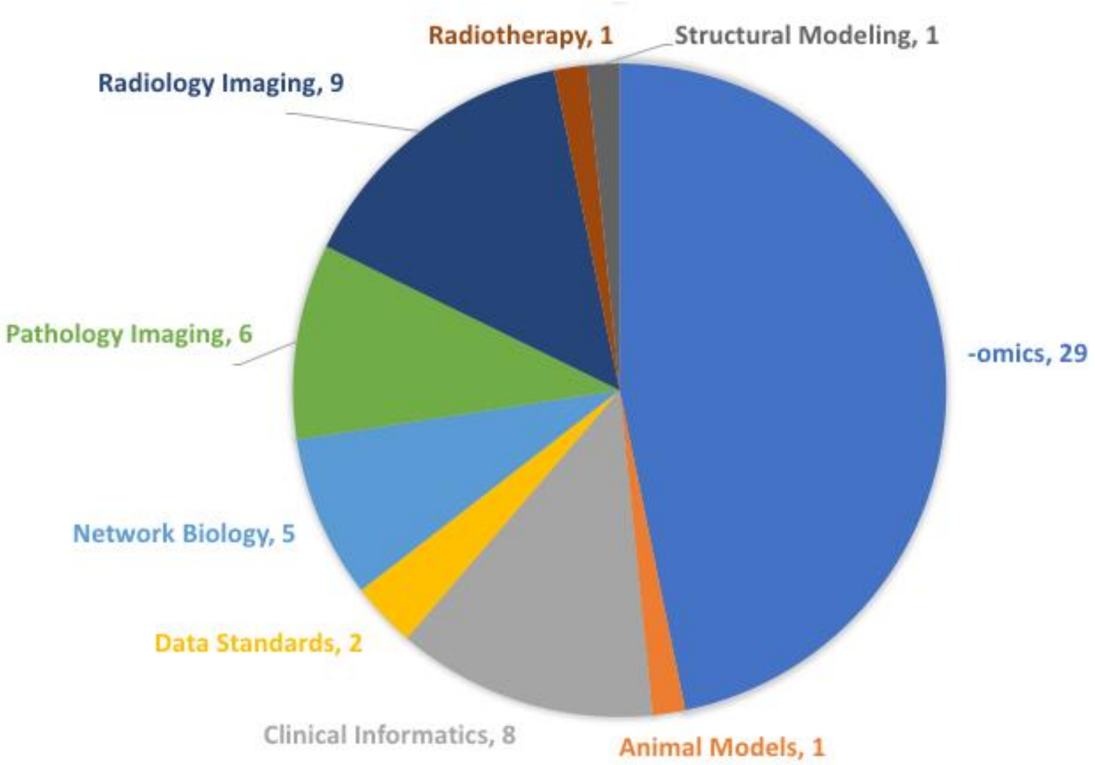
https://itcr.cancer.gov

ITCR supports the informatics technology development lifecycle

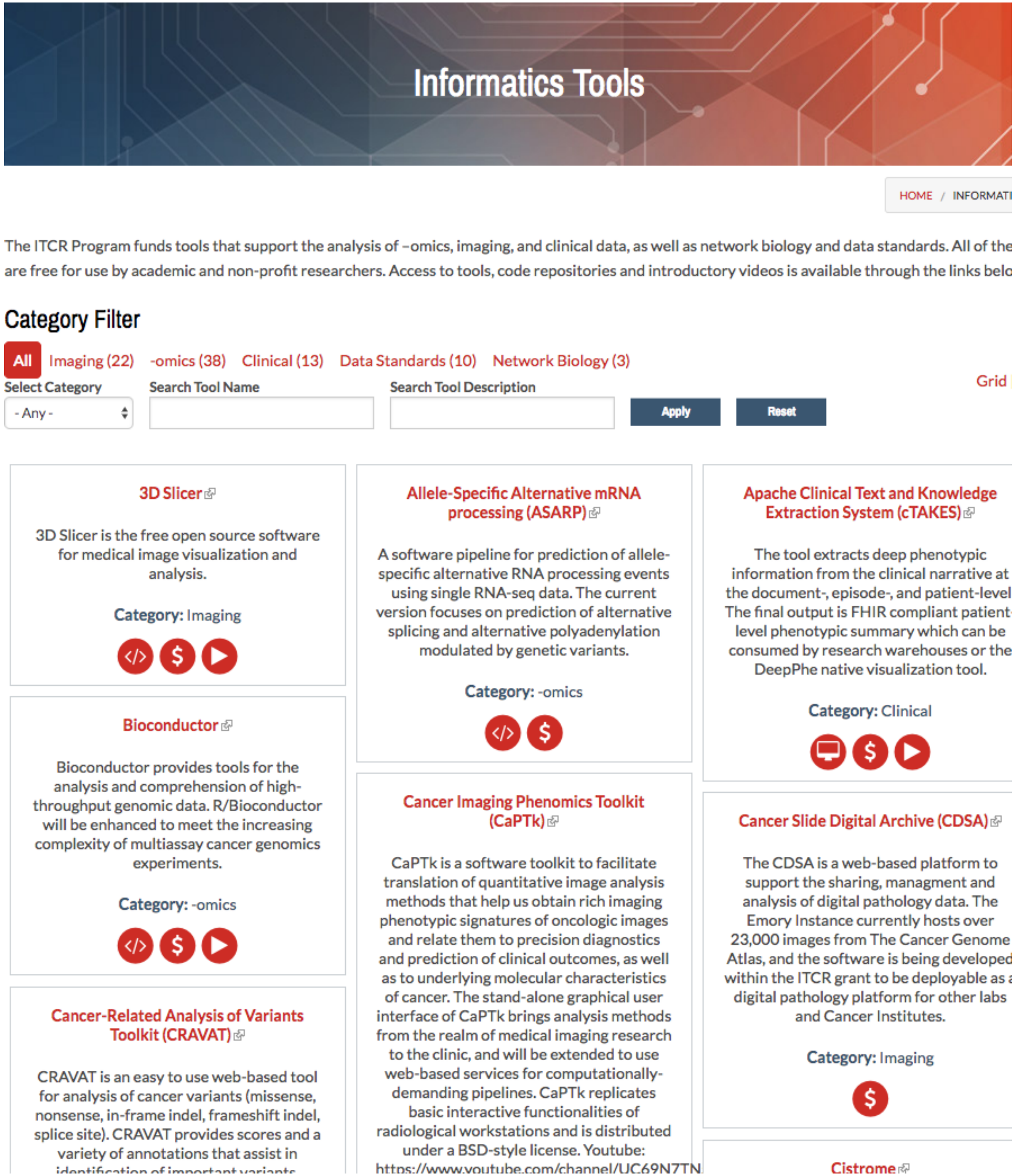


PAR-15-334	Algorithm Development	**R21** ***Innovative computational research*** Up to $275K DC for 2 years
PAR-15-332	Prototyping & Hardening	**U01** ***Early stage development*** Up to $300K DC/year for 3 years
PAR-15-331	Enhancement & Dissemination	**U24** ***Advanced stage development*** Up to $600K DC/year for 5 years
PAR-15-333	Sustainment	**U24** ***Sustain highly-accessed resources*** No budget ceiling; up to 5 years

Current ITCR Portfolio



All funded grants, by domain



Informatics Tools

HOME / INFORMATI

The ITCR Program funds tools that support the analysis of –omics, imaging, and clinical data, as well as network biology and data standards. All of the are free for use by academic and non-profit researchers. Access to tools, code repositories and introductory videos is available through the links belo

Category Filter

All Imaging (22) -omics (38) Clinical (13) Data Standards (10) Network Biology (3)

Grid

Select Category: - Any -
Search Tool Name:
Search Tool Description:

Apply Reset

3D Slicer

3D Slicer is the free open source software for medical image visualization and analysis.

Category: Imaging

Allele-Specific Alternative mRNA processing (ASARP)

A software pipeline for prediction of allele-specific alternative RNA processing events using single RNA-seq data. The current version focuses on prediction of alternative splicing and alternative polyadenylation modulated by genetic variants.

Category: -omics

Apache Clinical Text and Knowledge Extraction System (cTAKES)

The tool extracts deep phenotypic information from the clinical narrative at the document-, episode-, and patient-level The final output is FHIR compliant patient level phenotypic summary which can be consumed by research warehouses or the DeepPhe native visualization tool.

Category: Clinical

Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. R/Bioconductor will be enhanced to meet the increasing complexity of multiassay cancer genomics experiments.

Category: -omics

Cancer Imaging Phenomics Toolkit (CaPTk)

CaPTk is a software toolkit to facilitate translation of quantitative image analysis methods that help us obtain rich imaging phenotypic signatures of oncologic images and relate them to precision diagnostics and prediction of clinical outcomes, as well as to underlying molecular characteristics of cancer. The stand-alone graphical user interface of CaPTk brings analysis methods from the realm of medical imaging research to the clinic, and will be extended to use web-based services for computationally-demanding pipelines. CaPTk replicates basic interactive functionalities of radiological workstations and is distributed under a BSD-style license. Youtube: https://www.youtube.com/channel/UC69N7TN

Cancer Slide Digital Archive (CDSA)

The CDSA is a web-based platform to support the sharing, managment and analysis of digital pathology data. The Emory Instance currently hosts over 23,000 images from The Cancer Genome Atlas, and the software is being developed within the ITCR grant to be deployable as a digital pathology platform for other labs and Cancer Institutes.

Category: Imaging

Cancer-Related Analysis of Variants Toolkit (CRAVAT)

CRAVAT is an easy to use web-based tool for analysis of cancer variants (missense, nonsense, in-frame indel, frameshift indel, splice site). CRAVAT provides scores and a variety of annotations that assist in identification of important variants

Cistrome

ITCR Software is Free and Open Source

- The software is **freely available** to biomedical researchers and educators in the non-profit sector

- The terms of software availability should include the **ability of researchers to modify the source code**

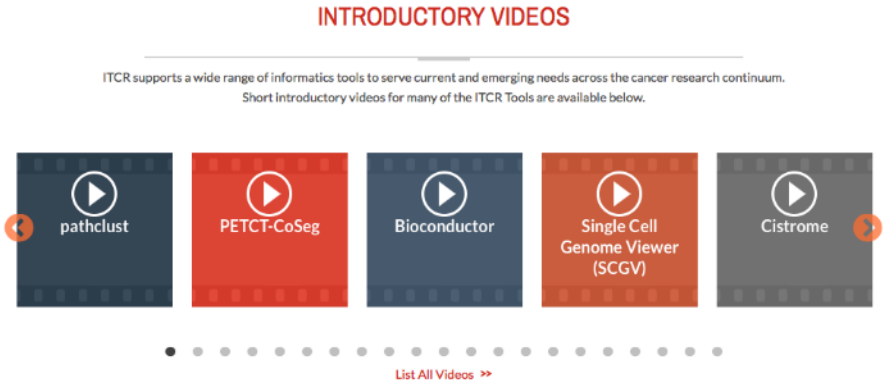
- The terms of software availability permit the **dissemination and commercialization** of enhanced or customized versions of the software

ITCR supports broad dissemination of the tool portfolio

- Conferences and workshops
- Social media - #nciitcr, @NCI_NCIP
- Introductory videos and tool catalog on the program website itcr.cancer.gov
- *Cancer Research* special issue on cancer informatics (published online Nov. 2017)

The Role of Academic Technology Development in Cancer Research

Workshop organized by the National Cancer Institute Informatics Technology for Cancer Research community at the CI4CC 2016 Spring Symposium



INTRODUCTORY VIDEOS

ITCR supports a wide range of informatics tools to serve current and emerging needs across the cancer research continuum. Short introductory videos for many of the ITCR Tools are available below.

Cancer Research

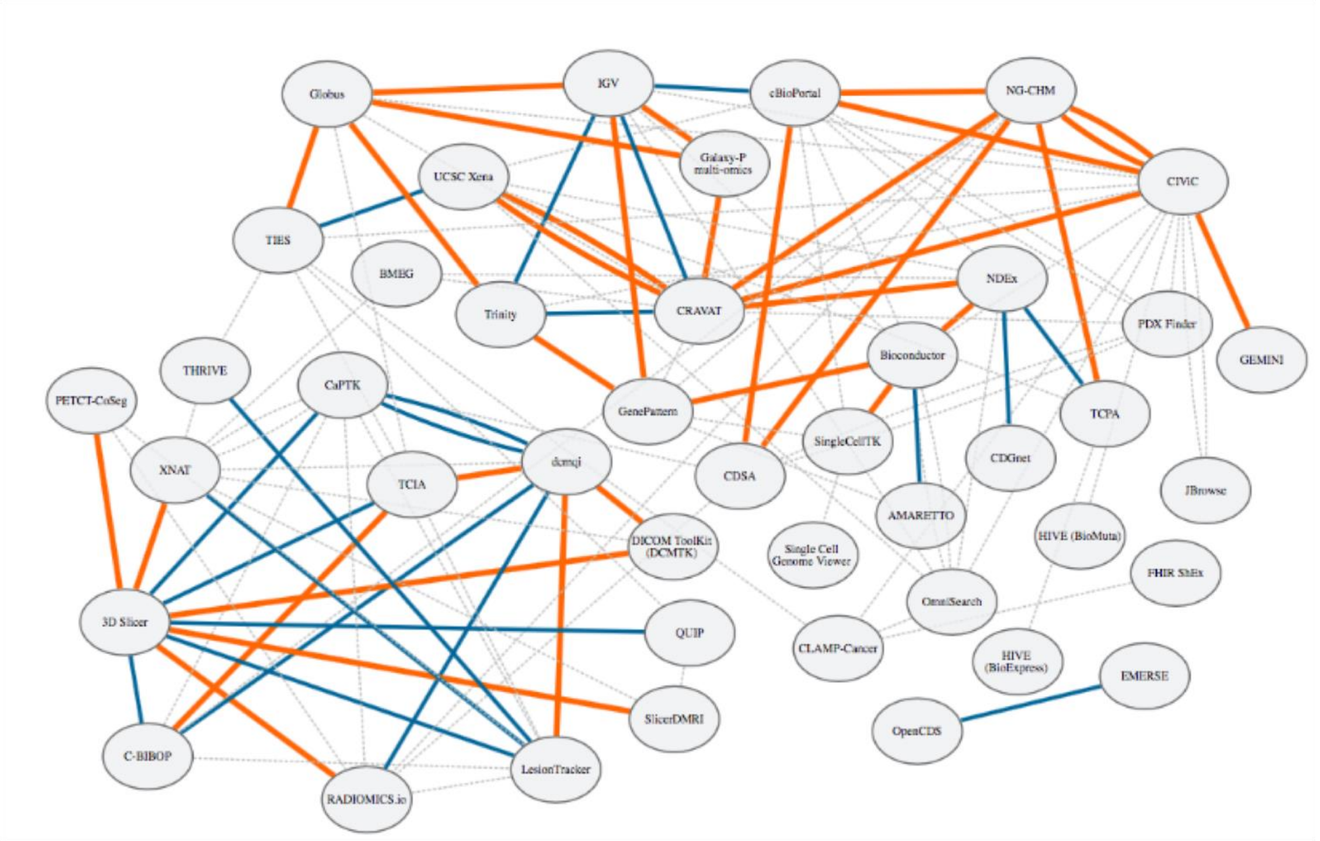
search
Advanced Search
Home
About
Articles
For Authors
Alerts

CANCER RESEARCH

Focus on Computer Resources

AACR
American Association for Cancer Research
Genomics | Proteomics | Animal Models | Imaging | Clinical

ITCR Promotes collaboration and interoperability



- Monthly PI conference calls
- Annual face-to-face meetings
- Investigator-led working groups
- Administrative supplements
- Collaborative set-asides
- Affiliated projects

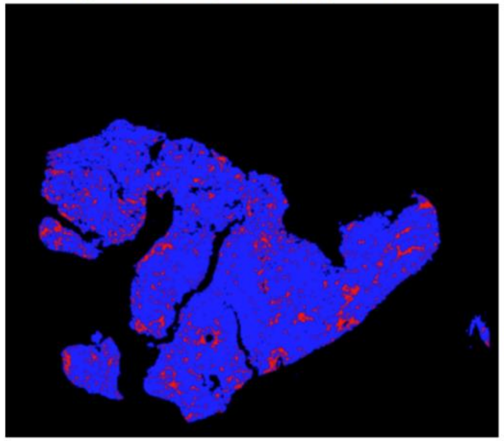
Tools to Analyze Morphology and Spatially Mapped Molecular Data

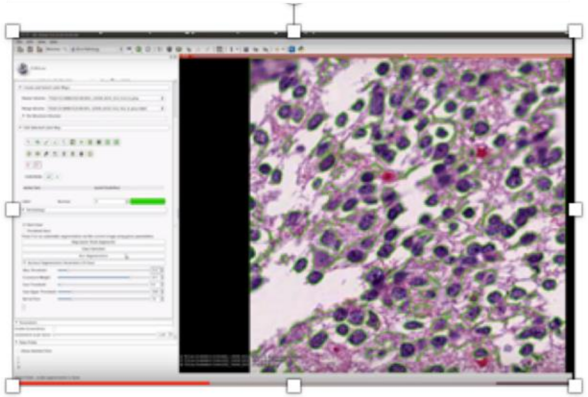
Joel Saltz, Stony Brook

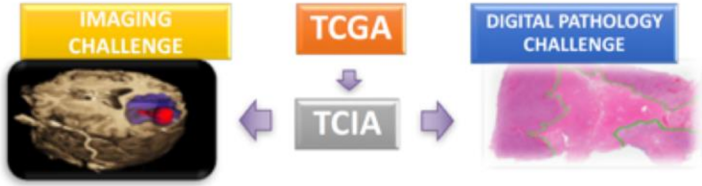
(U24, 3 of 5 years)

- Tools are being used to support several research collaborations:
  - Leading a TCGA Pan Cancer Atlas Immune group whole slide tissue image analysis effort
  - SEER pilot study on integrative whole slide tissue image data into the SEER repository
  - Working with a team at Emory to investigate the spatial and temporal coordination of cell boundary dynamics in NSCLC.
- Collaborating with several ITCR groups
  - QIICR: Added Pathology Analysis Extension to 3D Slicer
  - MGH team: MICCAI Digital Pathology challenges

Tumor Infiltrating
Lymphocytes
By CNN Deep Learning Method

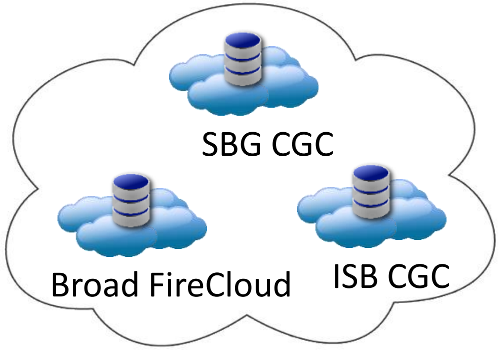






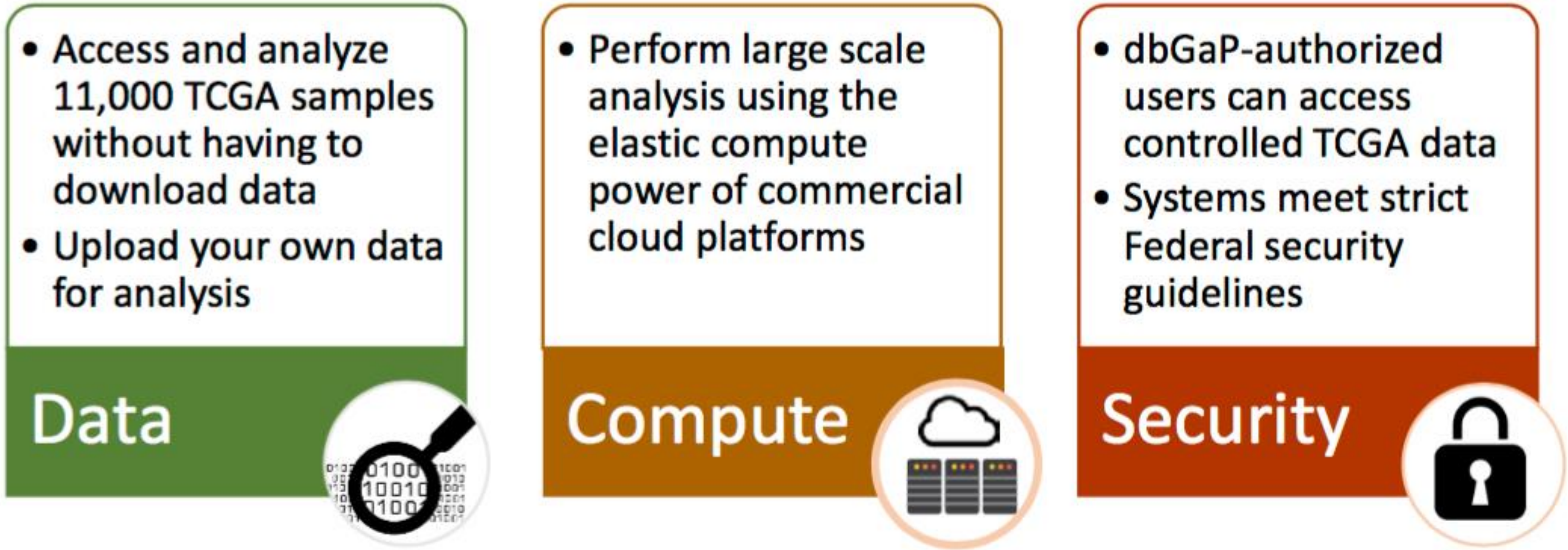
ITCR and the Cloud Resources

NCI Cloud Resources



Cloud Resources provide:

- Access to large genomic data sets without need to download
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses



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

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

#NCICloud

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

- What is a “container”?
  - A container is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run it….Containers will always run the same regardless of the environment.*
- Docker is the *de facto* standard software for creating containers.
- Dockstore is an open platform for sharing Docker-based tools and workflows, developed through GA4GH.

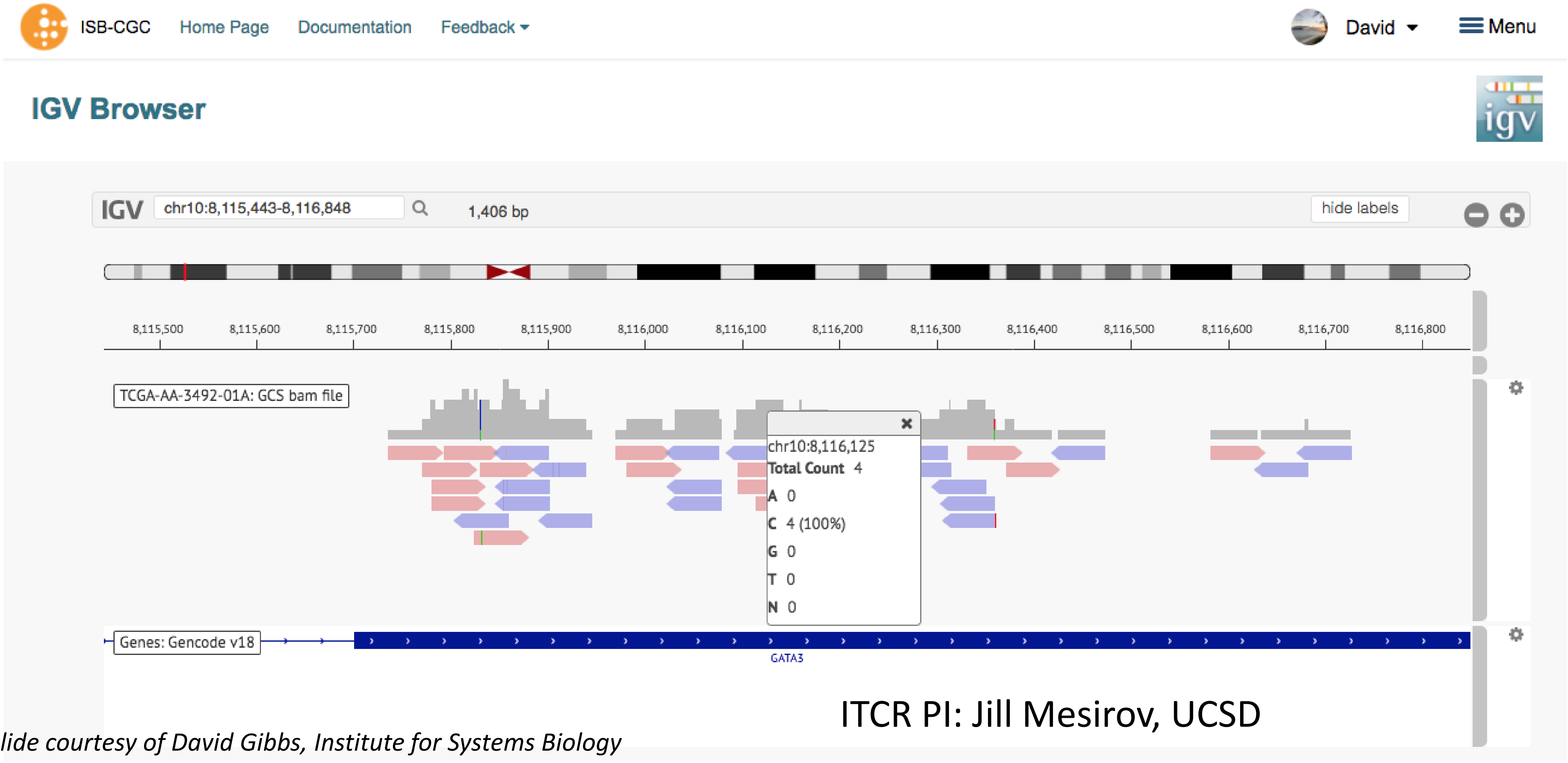
docker

Dockstore
Create, Share, Use

Global Alliance
for Genomics & Health

* https://www.docker.com/what-container

Accessing the Integrative Genomics Viewer on ISB-CGC



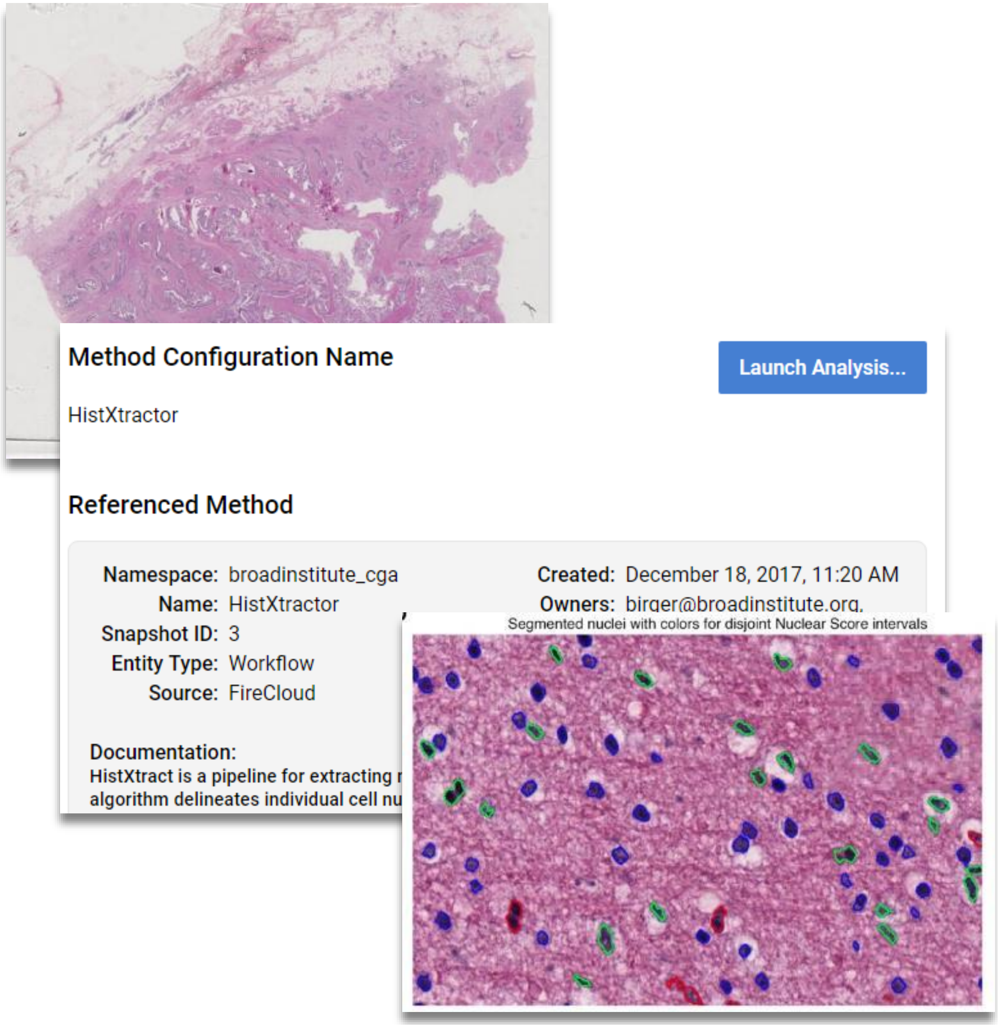
chr10:8,115,443-8,116,848 1,406 bp

chr10:8,116,125
Total Count 4
A 0
C 4 (100%)
G 0
T 0
N 0

ITCR PI: Jill Mesirov, UCSD

Slide courtesy of David Gibbs, Institute for Systems Biology

Extracting nuclear morphometry features on FireCloud



Method Configuration Name

HistXtractor

Referenced Method

Namespace: broadinstitute_cga
Name: HistXtractor
Snapshot ID: 3
Entity Type: Workflow
Source: FireCloud

Created: December 18, 2017, 11:20 AM
Owners: birger@broadinstitute.org,

Documentation:
HistXtract is a pipeline for extracting r
algorithm delineates individual cell nu

Segmented nuclei with colors for disjoint Nuclear Score intervals

Running HistXtract on TCGA diagnostic images in just a few clicks

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

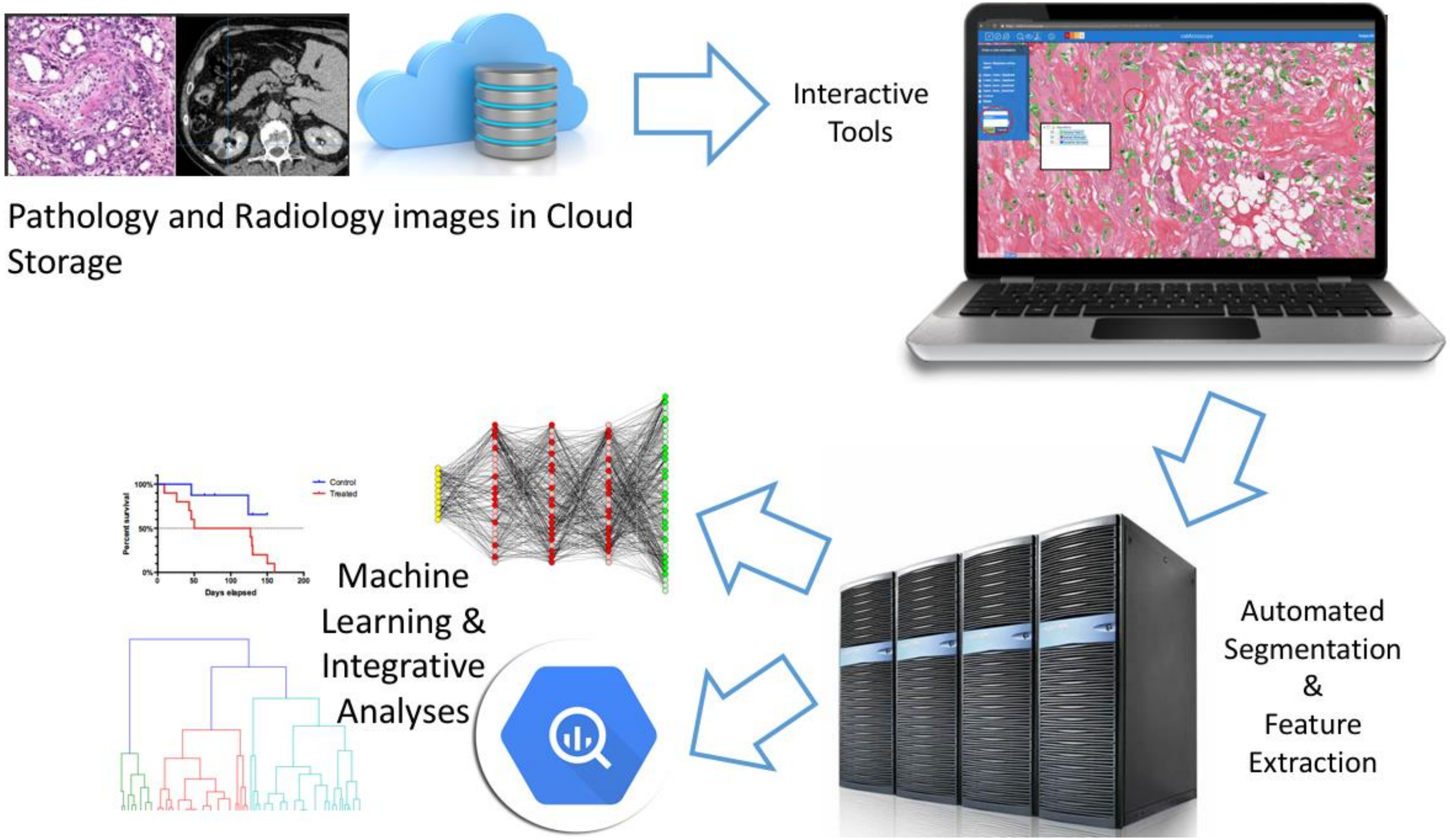
Members of the Getz Lab created an open-access FireCloud workspace preconfigured to download and analyze FFPE images for 9,600 participants across 32 types of cancer.

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

Generating Tumor Infiltrating Lymphocyte Maps on the ISB-CGC



Pathology and Radiology images in Cloud Storage

Interactive Tools

Automated Segmentation & Feature Extraction

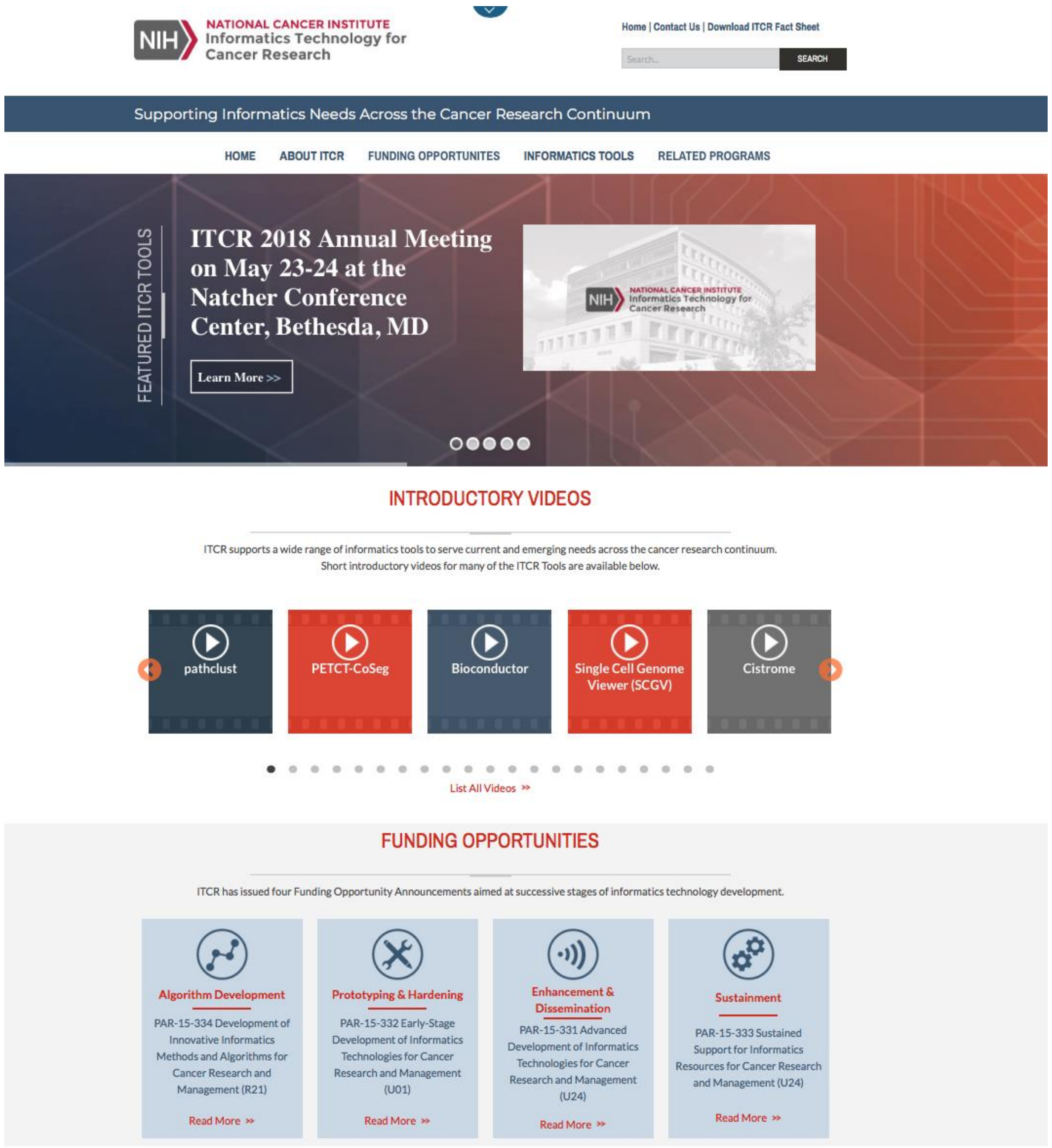
Machine Learning & Integrative Analyses

ITCR PIs: Joel Saltz, Ashish Sharma

Slide courtesy of David Gibbs, Institute for Systems Biology

Learn more!

- Information about tools, including introductory videos at https://itcr.cancer.gov
- Contact Juli Klemm: klemmj@mail.nih.gov
- Follow us on Twitter: #nciitcr, @NCI_NCIP
- Look at the *Cancer Research* Special Issue (Nov. 2017)



NIH NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

Home | Contact Us | Download ITCR Fact Sheet

Search... SEARCH

Supporting Informatics Needs Across the Cancer Research Continuum

HOME ABOUT ITCR FUNDING OPPORTUNITES INFORMATICS TOOLS RELATED PROGRAMS

FEATURED ITCR TOOLS

ITCR 2018 Annual Meeting on May 23-24 at the Natcher Conference Center, Bethesda, MD

Learn More >>

INTRODUCTORY VIDEOS

ITCR supports a wide range of informatics tools to serve current and emerging needs across the cancer research continuum.
Short introductory videos for many of the ITCR Tools are available below.

List All Videos »

FUNDING OPPORTUNITIES

ITCR has issued four Funding Opportunity Announcements aimed at successive stages of informatics technology development.

Algorithm Development
PAR-15-334 Development of Innovative Informatics Methods and Algorithms for Cancer Research and Management (R21)
Read More »

Prototyping & Hardening
PAR-15-332 Early-Stage Development of Informatics Technologies for Cancer Research and Management (U01)
Read More »

Enhancement & Dissemination
PAR-15-331 Advanced Development of Informatics Technologies for Cancer Research and Management (U24)
Read More »

Sustainment
PAR-15-333 Sustained Support for Informatics Resources for Cancer Research and Management (U24)
Read More »

The Imaging Data Commons

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.
  - Services that link disparate information, including clinical, image, and molecular data.
  - Essential underlying data science infrastructure, methods, and portals for the Cancer Data Ecosystem.
  - Establish sustainable data governance to ensure long-term health of the Ecosystem.
  - Develop standards and tools so that data are interoperable.

Enhanced Data Sharing Working Group Recommendation:

The Cancer Data Ecosystem

SBG CGC

Broad FireCloud

ISB CGC

Cancer Research Data Commons

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

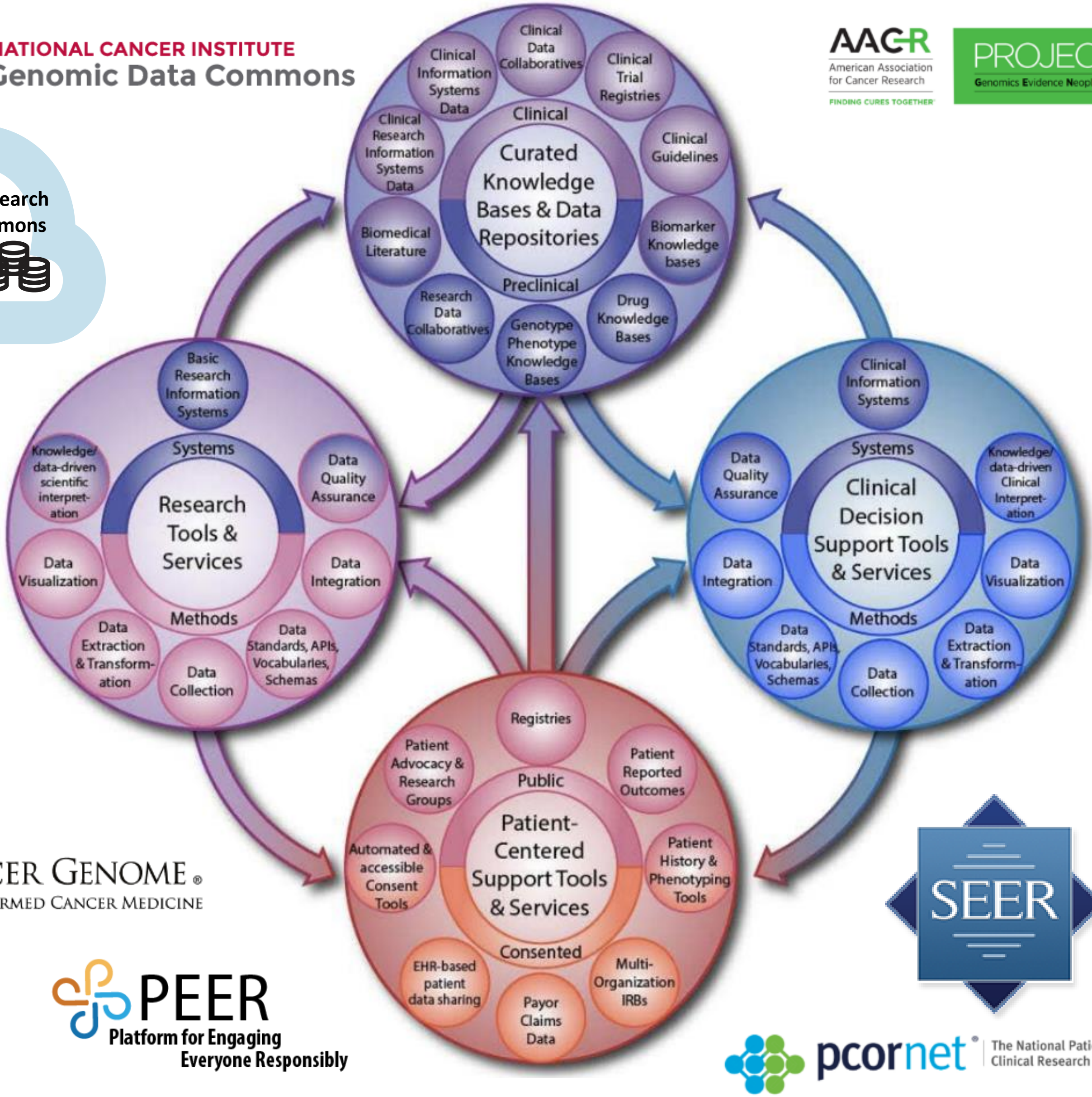
NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

monarch INITIATIVE

MY CANCER GENOME® GENETICALLY INFORMED CANCER MEDICINE

Sage BIONETWORKS

PEER Platform for Engaging Everyone Responsibly



Curated Knowledge Bases & Data Repositories

Clinical

Preclinical

Clinical Data Collaboratives

Clinical Trial Registries

Clinical Guidelines

Biomarker Knowledge bases

Drug Knowledge Bases

Genotype Phenotype Knowledge Bases

Research Data Collaboratives

Biomedical Literature

Clinical Research Information Systems Data

Clinical Information Systems Data

Research Tools & Services

Systems

Methods

Basic Research Information Systems

Data Quality Assurance

Data Integration

Data Standards, APIs, Vocabularies, Schemas

Data Collection

Data Extraction & Transform-ation

Data Visualization

Knowledge/ data-driven scientific interpret-ation

Clinical Decision Support Tools & Services

Systems

Methods

Clinical Information Systems

Knowledge/ data-driven Clinical Interpret-ation

Data Visualization

Data Extraction & Transform-ation

Data Collection

Data Standards, APIs, Vocabularies, Schemas

Data Integration

Data Quality Assurance

Patient-Centered Support Tools & Services

Public

Consented

Registries

Patient Reported Outcomes

Patient History & Phenotyping Tools

Multi-Organization IRBs

Payor Claims Data

EHR-based patient data sharing

Automated & accessible Consent Tools

Patient Advocacy & Research Groups

AACR American Association for Cancer Research FINDING CURES TOGETHER

PROJECTGENIE Genomics Evidence Neoplasia Information Exchange

ORIEN ONCOLOGY RESEARCH INFORMATION EXCHANGE NETWORK

ASCO CANCER•LINQ Learning Intelligence Network for Quality

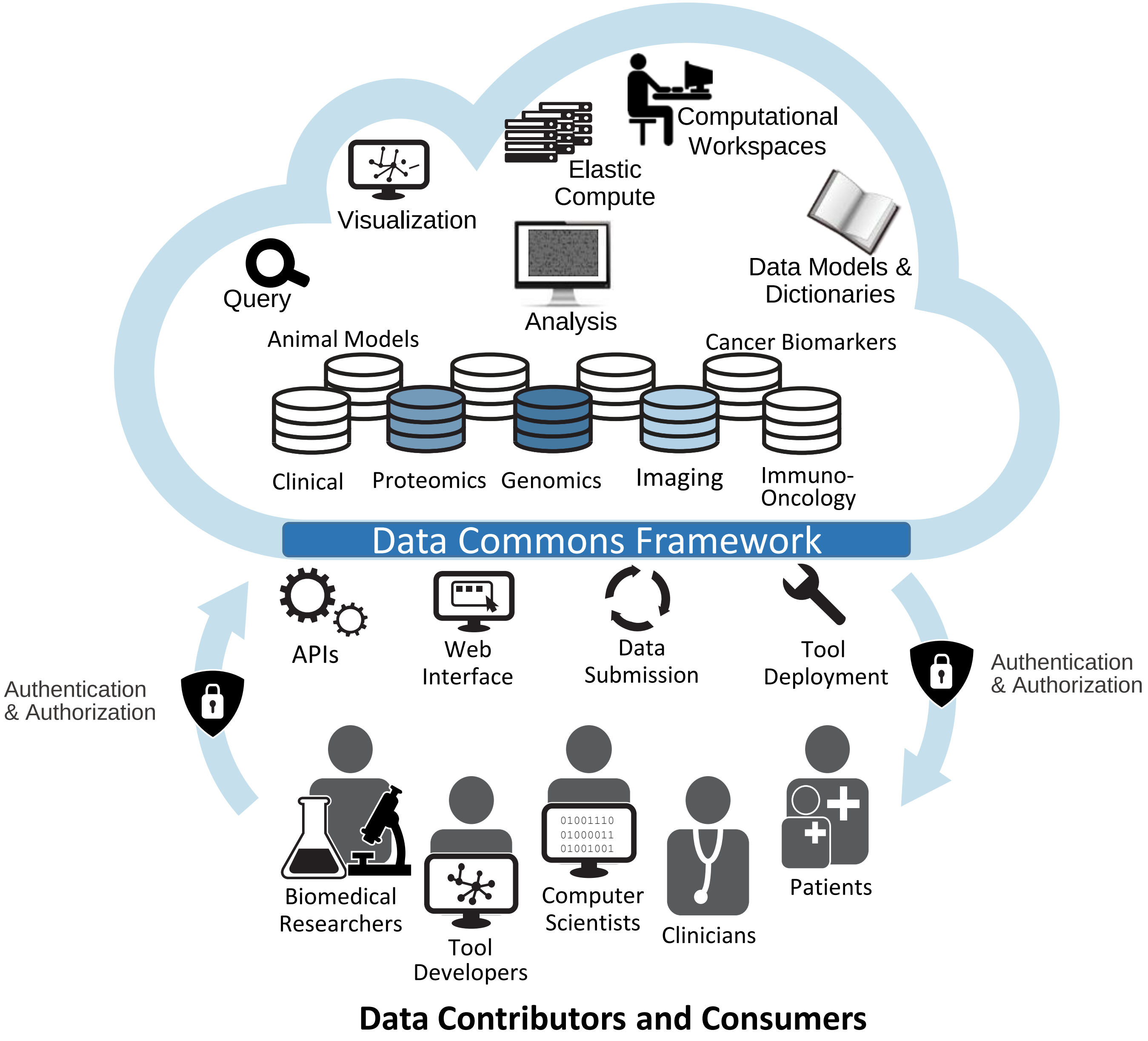
SEER

pcornet® The National Patient-Centered Clinical Research Network

NCI Cancer Research Data Commons (CRDC) - Concept

NCI Scope: *"Create a data science infrastructure necessary to connect repositories, analytical tools, and knowledge bases"*

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



*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016. Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.

Goals of the NCI CRDC

- Enable the cancer research community to share diverse data types across programs and institutions.

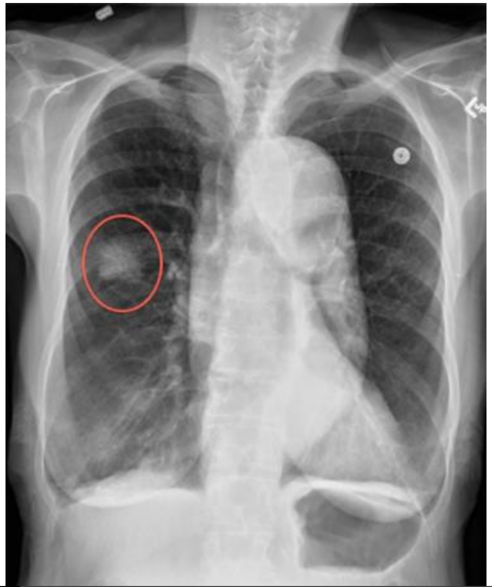
- Provide easy access to data, regardless of where they are stored.

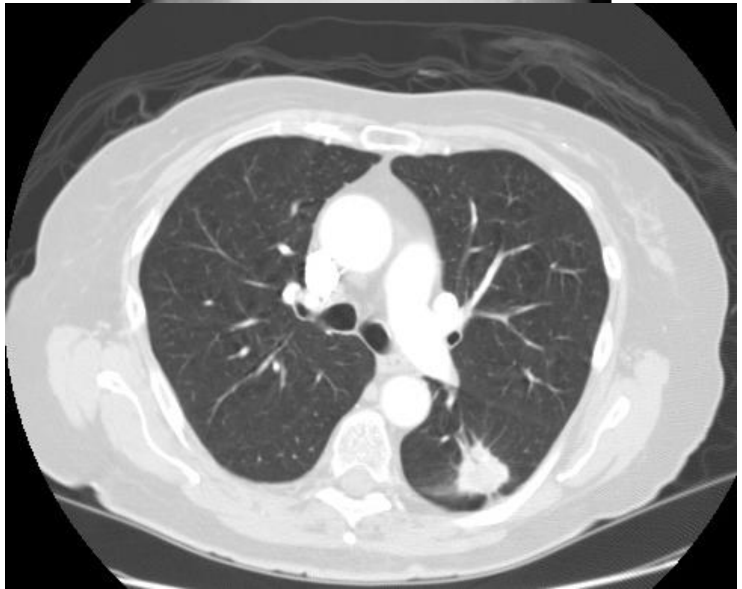
- Provide mechanisms for innovative tool discovery, access, and usage, e.g., ITCR tools.

- Help Data Coordinating Centers share their data publicly and provide longer term sustainability.

Imaging in Cancer is Comprised of a Variety of Image Types

- The Cancer Imaging Archive (TCIA)
  - NCI repository for radiology images (and now digital pathology)
  - Most images in DICOM standard
  - Currently ~20 TB of data, 31 million images from ~41,000 patients
- NCI projects generating image data
  - Human Tumor Atlas (HTA)
  - CPTAC (Cancer Proteomics Tumor Analysis Consortium)
  - APOLLO (Applied Proteogenomics Organizational Learning and Outcomes)

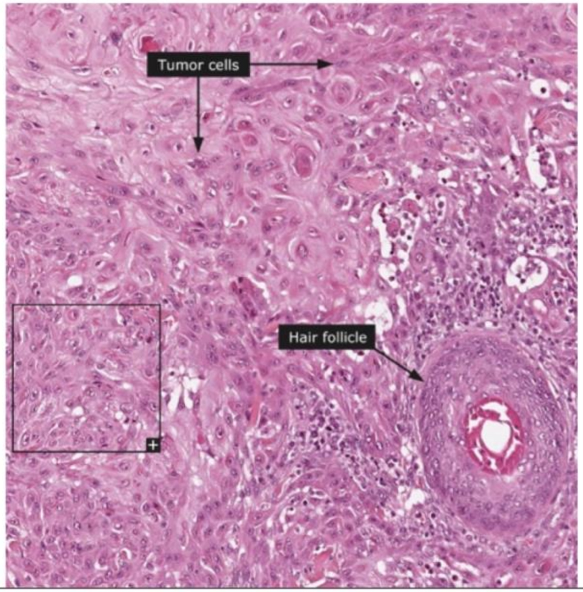


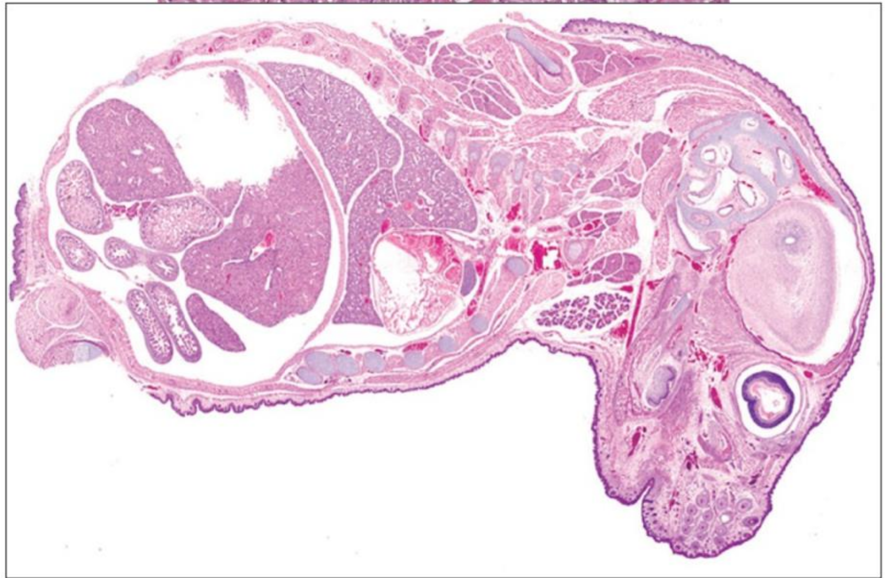


http://www.jpathinformatics.org/viewimage.asp?img=JPatholInform_2012_3_1_9_93891_f4.jpg
TCIA – https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

Imaging in Cancer is Comprised of a Variety of Image Types

- The Cancer Imaging Archive (TCIA)
  - NCI repository for radiology images (and now digital pathology)
  - Most images in DICOM standard
  - Currently ~20 TB of data, 31 million images from ~41,000 patients
- NCI projects generating image data
  - Human Tumor Atlas (HTA)
  - CPTAC (Cancer Proteomics Tumor Analysis Consortium)
  - APOLLO (Applied Proteogenomics OrganizationaL Learning and Outcomes)





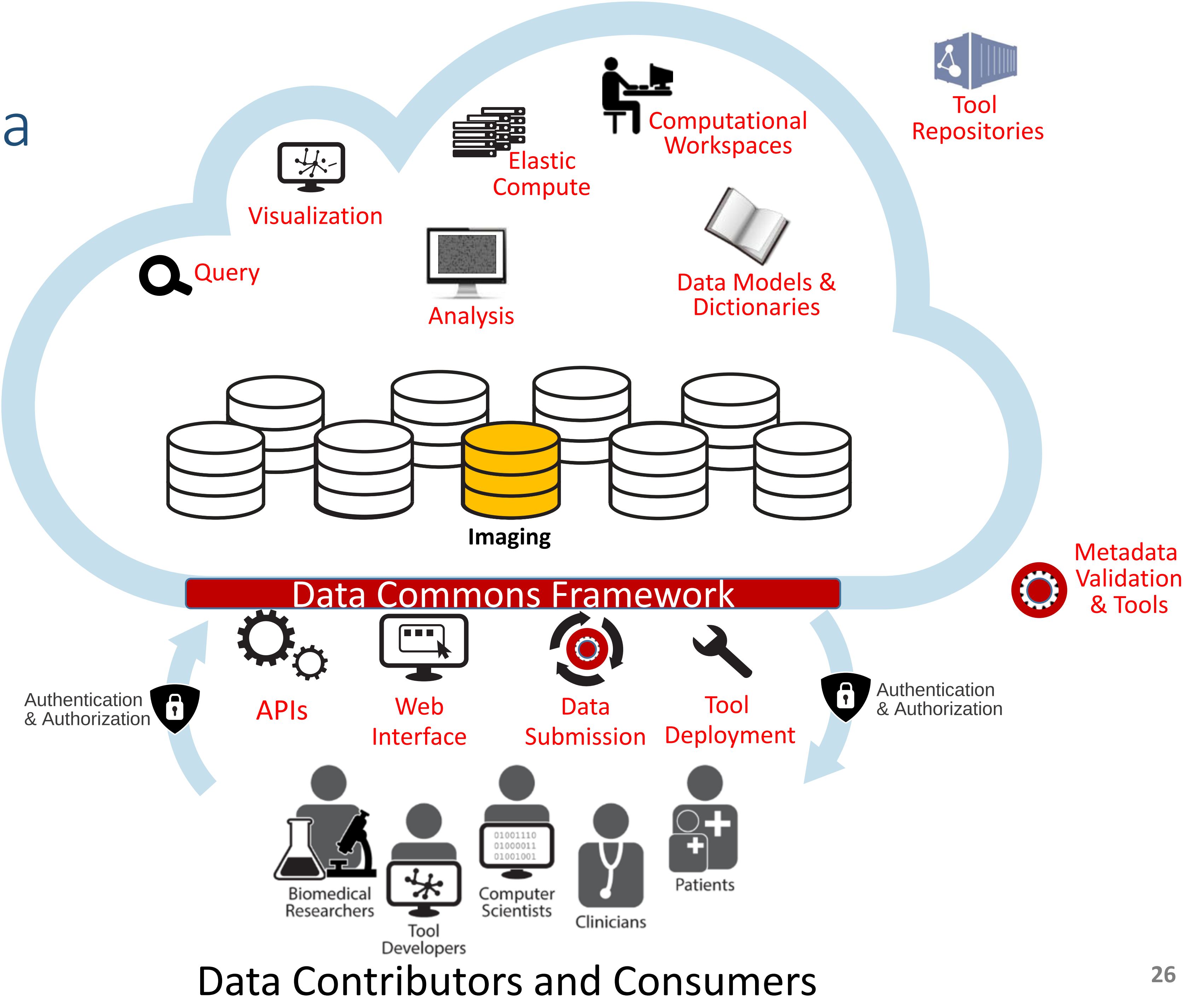
http://www.svuhradiology.ie/case-study/lung-cancer/.
https://www.proteinatlas.org/learn/dictionary/pathology/skin+cancer+3/detail+2.

Imaging Data Commons (IDC)

Goal: Develop a resource that provides access to and analysis of cancer-related imaging data.

- Along with the CRDC Resources, enable a secure environment for comparison and analysis of publicly available data with private data and enable both large and small scale collaborations
- Provide easy access to diverse imaging repositories visualization and analysis tools **(like those in the ITCR catalog)**
- Provide datasets for tool development and validation in multiple imaging disciplines
- Continuous community engagement to adapt to new projects and image types as needed to support ongoing integration of images with molecular and clinical data

The IDC will be a Cancer Research Data Commons (CRDC) Node



TCIA

*The Cancer Imaging Archive**

NIH Request for Information:
Input on Development of the NCI Imaging Data Commons

NOT-CA-18-060

The NCI is inviting comments and suggestions on the development of the NCI Imaging Data Commons (IDC), a node of the Cancer Research Data Commons. The IDC will provide:

- **access to image repositories**
- **analysis tools**
- **scalable computing resource**
- **a cloud-based, collaborative environment.**

To best serve the needs of the cancer imaging community, we are seeking input from potential users of the IDC to determine the best features to include in an IDC prototype. All stakeholders involved in cancer imaging are invited to respond to this Request.

More details about the RFI and how to respond can be found at

https://grants.nih.gov/grants/guide/notice-files/NOT-CA-18-060.html

The deadline for submission is May 4, 2018.

For any questions about this request, please contact
NCIIDCRFI@mail.nih.gov

NIH Request for Information:
Input on Development of the NCI Imaging Data Commons

NOT-CA-18-060

30 responses received, from one sentence replies to very thorough commentaries

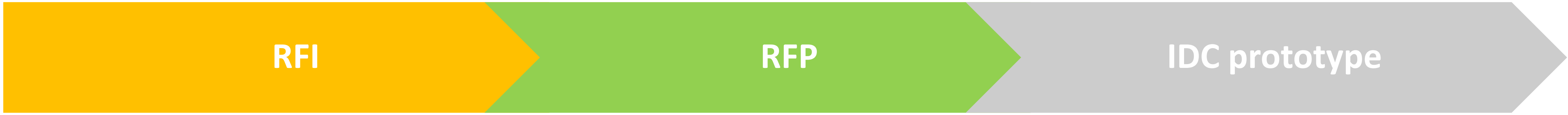
Lessons learned from the RFI responses:

- **The cancer imaging community is not a single community (no surprise), but can be roughly divided into medical imaging (including DP) and the microscopy community (not including DP)**
- **Standards – responses divide along the above classifications, with medical imagers strongly recommending DICOM, and microscopists not as cohesive**
- **Many suggested that the NCI act as the enforcer of standards**
- **Curated data sets are crucial to the software developers; the IDC should act as a repository for collections**
- **Not so much need currently for imaging intraoperability (basic microscopy – CT, for example); more interest in interoperability with other –omics data**

IDC Development Timeline

Timeline

- With the guidance of the **NCI IDC Advisory Committee,** perform landscape analysis via in person interviews (NCI) and issue an RFI to gain an understanding of the community's needs
- Issue and award of RFP for the development of an initial IDC and follow-on development
- Development of an IDC prototype



RFI

Generate and publish RFI; response window; data collation; RFP generation
3 months

RFP

Issuance of RFP; response window; awarding and negotiation of award
3 months

IDC prototype

Development and production of IDC prototype
6-9 months

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

NIH

NATIONAL
CANCER
INSTITUTE

www.cancer.gov

www.cancer.gov/espanol