“Train your colleague” remote training session (17MAR20)

The four stacks involved

- AnVIL (NHGRI), BioData Catalyst (NHLBI), Cancer Research Data Commons (NCI), Gabriella Miller Kids First - Data Resource Center (CF)

Framing

- The four platforms have similar goals and user bases. It benefits our users and the biomedical research community to make the data and tools on our platforms widely available by interconnecting our platforms.

Goals

- Understand the components of the four platforms and how users flow through the platforms.

Kids First Data Resource Center

Vision and DRC Participants

Gabriella Miller Kids First Vision: Alleviate suffering from childhood cancer *and* structural birth defects by fostering **collaborative research** to uncover the etiology of these diseases and supporting **data sharing** within the pediatric research community.

DRC Funder: NIH Common Fund's Gabriella Miller Kids First Pediatric Research Program; administered by NHLBI (PO: Charlene Schramm) overseen by NIH Kids First Working Group.

DRC PIs: Adam Resnick, Brandi Davis-Dusenbery, Vincent Ferretti, Robert Grossman, Allison Heath, Deanne Taylor, Sam Volchenboum

DRC Institutions: Children's Hospital of Philadelphia, CHU Sainte-Justine, University of Chicago, Seven Bridges Genomics
Go Live Date: September 10, 2018

Data Available

Kids First Released Datasets (7,223 participants):

- Congenital Heart Defects (2,133)
- Orofacial Cleft - European Ancestry (1,295)
- Ewing Sarcoma - Genetic Risk (1,047)
- Syndromic Cranial Dysinnervation (801)
- Orofacial Cleft - Latin American (804)
- Congenital Diaphragmatic Hernia (581)
- Disorders of Sex Development (300)
- Adolescent Idiopathic Scoliosis (262)

~24,000 additional participants approved for WGS sequencing as part of Kids First

Interoperable Datasets (1,677 participants):

- TARGET: Neuroblastoma (274)
- TARGET: Acute Myeloid Leukemia (321)
- Pediatric Brain Tumor Atlas: CBTTC (912)
- Pediatric Brain Tumor Atlas: PNOC (33)
- OpenDIPG: ICR London (137)

Cavatica enables "bring your own" data
NCI CDS and NCBI SDDP Interoperability coming

Kids First Data Resource Center

Interoperability Use Cases

We have a number of researchers that will be working across datasets in multiple platforms. In general, pediatrics and developmental biology spans organs, disease and data types - making effective and sustained interoperability a critical need.

Examples of overlapping cohorts:
INCLUDE: KFDRC and DataSTAGE
PCGC: KFDRC and DataSTAGE
CMG: KFDRC and AnVIL
CSER: KFDRC and AnVIL
NBL: KFDRC and NCI CRDC
AML: KFDRC and NCI CRDC
Brain Tumors: KFDRC and NCI CRDC

Early on, we realized we needed “intra”-operability for releases and developed a release coordination protocol. Perhaps an area for interoprability as well.

Clinical/Phenotypic Data Interop Pilot Opportunities

Clinical and phenotypic data has been identified as a limiting factor across supported studies. The wide range provides challenges, but also opportunities for interoperability. Areas under investigation include:

- FHIR as a core standard
- Terminologies for harmonized and computable data values
- Identifying virtual cohorts across the platforms based on clinical and phenotypic data
- Workflows for clinical/phenotypic data harmonization

Kids First Data Resource Center

Tools Available

Kids First DRC workflows

- Goal: functional equivalency to other large datasets/resources
- bwa-mem based alignment with GRCh38
- Germline:
  - Trio/family-based GATK germline best practices, GATK genotype refinement
- Somatic:
  - Strelka2, Mutect2, Lancet, VarDict, Manta, Control-FREEC
- RNA-Seq:
  - STAR 2-pass, RSEM, Kallisto, STAR-fusion, Arriba

Cavatica

- Over 300 public apps, interactive analysis with Jupyter Notebooks and RStudio
- Over 200,000 workflows run on Cavatica since KFDRC go live

Integration with PedCBioPortal

- Open access visualization and analysis of somatic mutations, expression, proteomics

Variant warehouse development underway

Authentication, Authorization and Indexing

Authentication:

- **Controlled access:** eRA Commons
- **Registered access:** Google, Facebook and ORCiD (OAuth2)

Authorization:

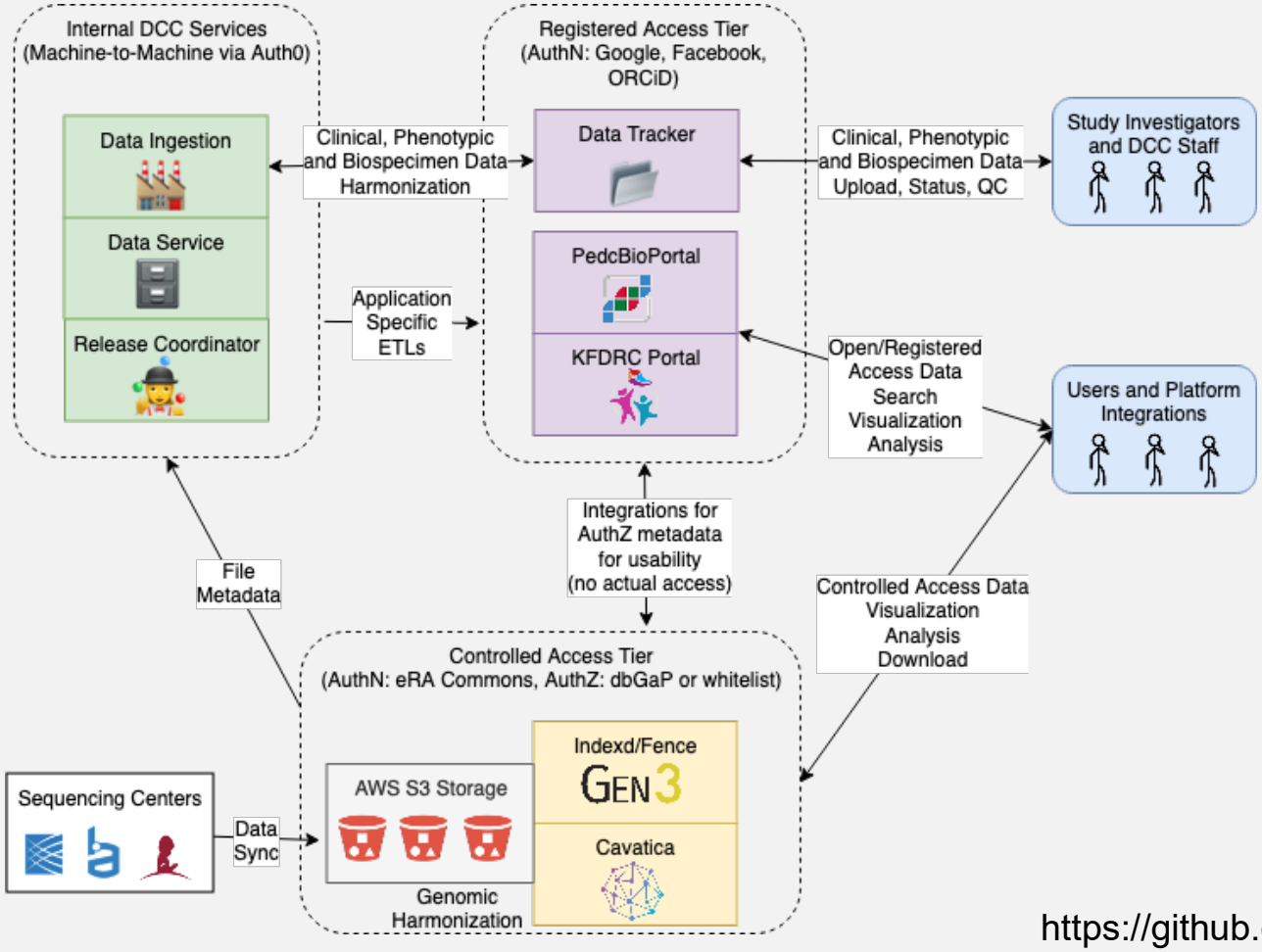
- **NIH Datasets:**
  - Telemetry reports from dbGaP either to Kids First Framework Services or NCI CRDC Framework Services
- **Consortium Datasets:**
  - Whitelist via Gen3 (also eRA Commons AuthN)

Indexing:

- The Kids First Data Service integrates with Gen3 indexd to index and associate relevant data for querying to the files. Cavatica also recently released GA4GH DRS endpoints for all files.

Kids First Data Resource Center

Architectural Diagram - Technical



https://github.com/kids-first/

NHLBI BioData Catalyst

Project Description

The NHLBI DataSTAGE project (Storage, Toolspace, Access and analytics for biG data Empowerment) aims to develop a cloud-based platform using FAIR principles to advance Heart, Lung, Blood, and Sleep research, building tools and workflows that enable investigators engaged in mining large, high-value datasets. The goal is to democratize data and compute access, while ensuring appropriate data security, to accelerate efficient biomedical research and maximize community engagement, productivity, and discovery.

Initial Go Live Date: January 2020

Participants

Funder: National Heart, Lung, and Blood Institute, Gary Gibbons (Director), Alastair Thomson (CIO), Jon Kaltman (Program Officer)

PIs: Ahalt, Avillach, Boyles, Bradford, Davis-Dusenbery, Krishnamurthy, Grossman, Paten, Philippakis, Thessen

Institutions: The Broad Institute, University of Chicago, University of California, Santa Cruz, Harvard Medical School, Berkeley Lab, Oregon State University, RTI International, University of New Mexico, UNC-CH/RENCI, Seven Bridges Genomics, Elsevier, Repositive, Veterans Affairs

NHLBI BioData Catalyst

Data available

Current/near-term:
- 145,753 whole genome sequences on platform
- 56,379 whole genome sequences available to users under controlled access
- supporting and phenotypic data from 51 pre-existing studies
- > 15,000 CT chest scans from COPDGene

For more info: bit.ly/TOPMedFreeze5b

In progress:
- 140,306 WGS under controlled access
- supporting and phenotypic data from >70 pre-existing studies

Future:
- Continued growth

Tools available

DataSTAGE is being designed to support **workflows** for batch data analysis, **notebooks** for interactive analysis, and **apps/services** for web apps. Users can bring their own workflows and notebooks.

Workflows
DataSTAGE supports workflows written in CWL and WDL. Highlights include workflows from the DCC and TOPMed alignment and variant calling.

Notebooks
RStudio and Jupyter Notebooks are supported with examples leveraging DataSTAGE for image visualization, machine learning, and GPU acceleration.

Apps/Services
DataSTAGE will support services such as the PIC-SURE API and web apps like i2b2/tranSMART and the Bravo browser.

NHLBI BioData Catalyst

Authentication

eRA Commons IDs are used for controlled access data through the Data Commons Framework Services.

In the coming year, evaluation of ORCID or other authentication methods will be completed.

Authorization

DCFS' dbGaP integration is used to streamline access for those with completed dbGaP applications.

Users with access to data through other agreements can be directly added to the whitelist.

Indexing

Data objects are assigned permanent globally unique IDs (GUIDs) to allow for access across tools, without requiring copies be created and transferred.

Datasets are identified through text-based and faceted search.

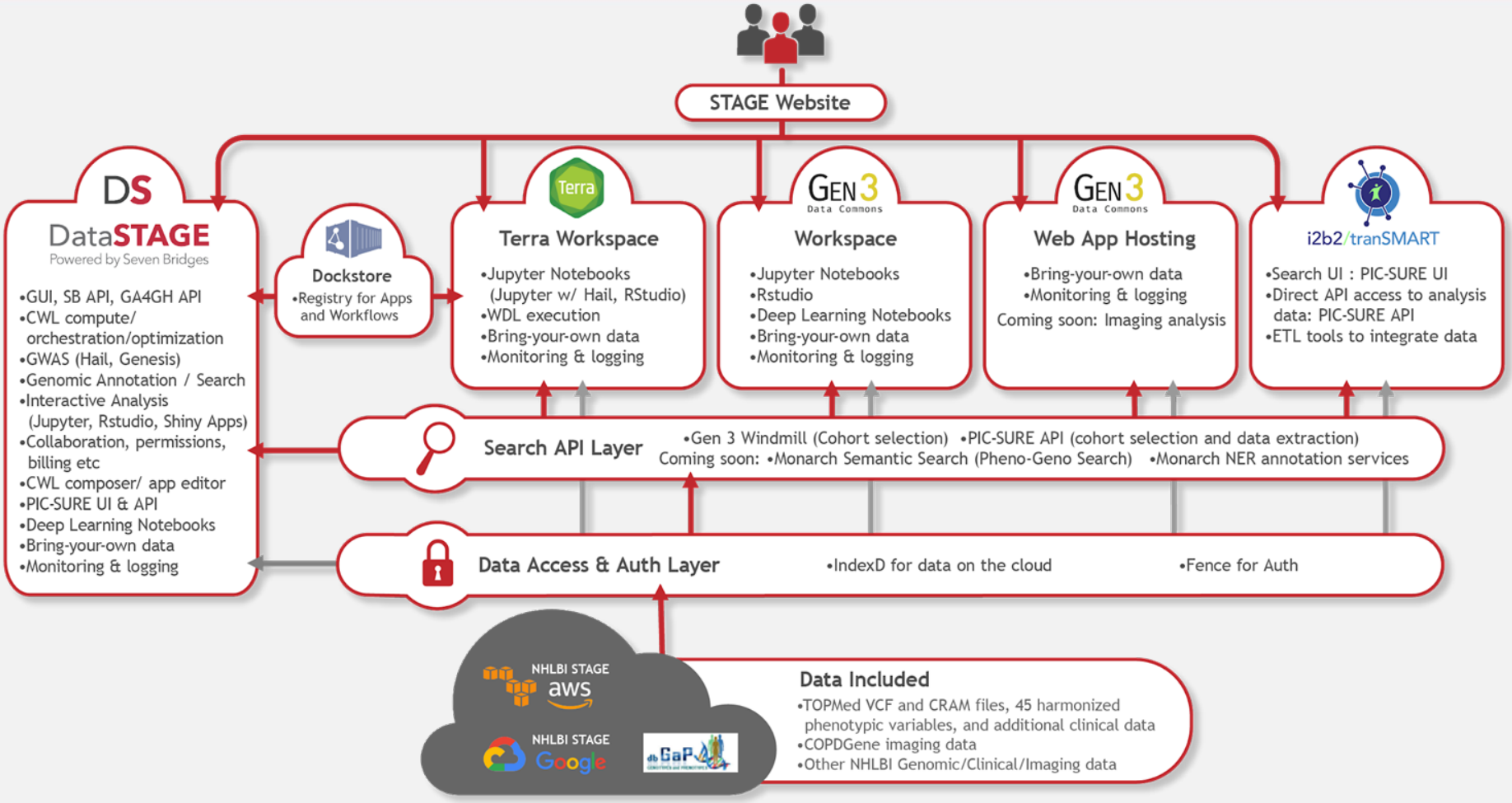
Data Modeling

Data interoperability using standards from GA4tGH and CD2H with FHIR and BioLink as meta models.

Monarch Initiative provides onto-informed knowledge graph for data discovery and integration.

NHLBI BioData Catalyst

Architectural Diagram



Cancer Research Data Commons

Project Description

CRDC Vision: To enable all participants across the cancer research and care continuum, including providers and patients, to contribute, access, and analyze diverse data that will enable new discoveries and lower the burden of cancer.

CRDC Mission: To create a virtual, expandable infrastructure that provides secure access to multi-modal data, allowing users to analyze, share, and store results, leveraging the storage and elastic compute of the cloud.

CRDC Approach: To provide interoperable resources through federation, data harmonization, standards, and tools and services that can be reused across the research community and to enable enhanced data sharing.

Live since 2015, new nodes annually

Participants

Funder: NCI

PIs: Brandi Davis-Dusenbery, Anthony Philippakis, Bill Longabaugh, David Pot, Bob Grossman, Anand Basu, Ron Kikinis

Institutions:
Seven Bridges
The Broad Institute
Institute for Systems Biology
University of Chicago
General Dynamics Information Technology
Enterprise Science and Computing (ESAC)
Frederick National Labs
Brigham and Women's Hospital

Cancer Research Data Commons

Data available

Genomics: BEATML, CCL, CGCI, CPTAC, CTSP, FM, HCMI, MMRF, NCICCR, Organoid, TARGET, TCGA, VAREPOP, COSMIC, HCA, GECCO, LCCC 1108, PPTC

Proteomics: CPTAC2, CPTAC3, CPTAC-TCGA, PBTA

Imaging: TCIA

23+ datasets covering genomics, proteomics, imaging, and more. Some data are present in both AWS and GCP, some present only on one cloud infrastructure.

~300 Reference and analyzed datasets available in BigQuery

Bring your own data is available.

New data being added through both existing and new data nodes on a continual basis.

Full list at bit.ly/CRDCdatasets

Tools available

Wide range of tools available across CRDC for all stages of analysis, types of users, with more added continuously by both new nodes and resources.

Bring your own tools is available.

SB: 427 publicly available tools and workflows in Common Workflow Language, + Dockstore, Rstudio, Jupyter notebooks, collaborative genome browser

Broad: >700 publically available workflows and tools in Workflow Development Language, Integrated Genome Viewer, Docktore, Jupyter notebooks, BigQuery, ML, pipelines

ISB-CGC: Google: VMs, BigQuery, AI, ML, Pipelines, Cohorts, Image Viewers, Notebooks, Plotting, Dockstore

GDC: Data Analysis Visualization Exploration (DAVE) tools allow users to interact intuitively with GDC data and promote the development of a true cancer genomics knowledge base.

PDC: Pepquery, Morpheus, Genome Browser, DDA & DIA common data analysis pipeline (coming soon)

Cancer Research Data Commons

Authentication

- eRA Commons IDs (controlled data)
- NCI Data Commons Framework Services (DCFS) by Gen3
- Individual, OIDC platform authentication

Authorization

- dbGaP access
- DCFS by Gen3
- Authorization enabled by Trusted Partnerships with NIH

Indexing

Permanent globally unique IDs (GUIDs) for data in Google & Amazon locations

GUIDs are cloud agnostic, promoting access and providing a mechanism for versioning data

Data Models & More

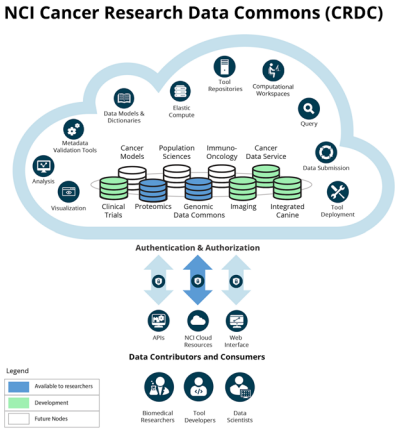
There are many data models across the CRDC, including ICDC, CTN, PDC, and GDC

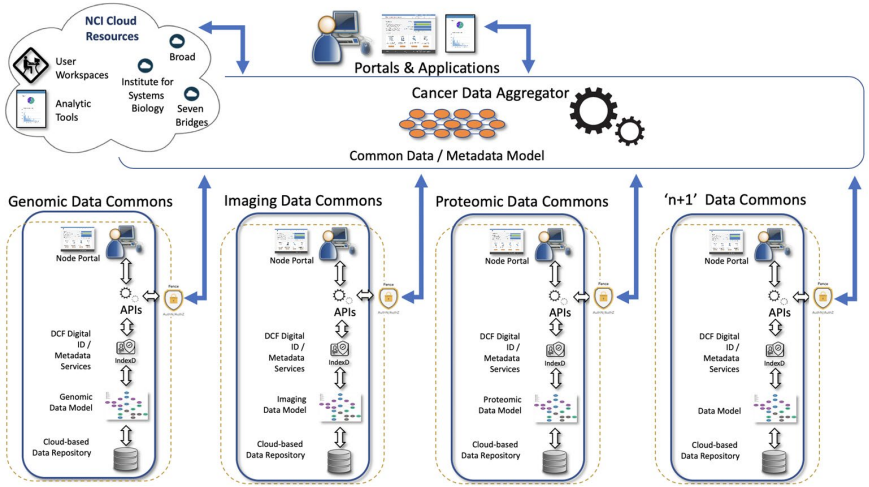
Bridging efforts are underway

We also participate in GA4GH efforts

Cancer Research Data Commons

Architectural Diagram





User Perspective

System perspective

NHGRI AnVIL

Project Description

The NHGRI AnVIL is a cloud based environment that hosts high value data sets and commonly used bioinformatics tools in a secure environment that can scale to meet the computational needs of researchers.
The AnVIL will provide automated data access with the DUOS platform. AnVIL users will have access to extensive training materials spanning from general genomics to advanced analysis methods using modular and open access Massively Open Online Courses (MOOCs).

Go Live: July 2019

Participants

Funder: National Human Genome Research Institute, Eric Green (director), Valentina Di Francesco (program officer), Ken Wiley (program officer)

PIs: Philippakis, Taylor, Grossman, Morgan, Paten, Nekrutenko, Carroll, Goecks, Hall, Carey, Afgan, Leek, Hansen, Schatz, Ellrott, Waldron, MacArthur

Institutions: The Broad Institute, Johns Hopkins University, University of Chicago, Penn State University, University of California, Santa Cruz, Oregon HEalth and Sciences University, Harvard Medical School, Vanderbilt University Medical Center, Roswell Park Comprehensive Cancer Center, Washington University, City University of New York

NHGRI AnVIL

Data available

AnVIL hosts CCDG, CMG, GTEx, and 1000 Genomes. As data sets are added to these groups they will become available on AnVIL. EMERGE will become available on AnVIL mid 2020.

Source	Subject	Sample	Project
CCDG	54302	54302	51
CMG	6224	6228	28
1000 Genomes	2504	2504	1
GTEx (v8)	979	17382	1
eMERGE*			

Tools available

AnVIL uses the Terra platform to launch and run tools on Google Cloud Platform within a FISMA secure boundary.
Users can currently run batch analysis with WDL and interactive analysis with Jupyter Notebooks supporting Python and R. The Dockstore workflow repository is integrated with Terra, providing access to hundreds of published workflows.
Access to Bioconductor through Jupyter Notebooks and R Studio is available and Galaxy will be available late summer 2020.
AnVIL will enable users to bring their own tools to the platform.

NHGRI AnVIL

Authentication

Both Google emails and eRA Commons IDs are used as an authentication mechanism for controlled access data

Authorization

Consortium and developer whitelists are maintained to provide access to data

General users submit DARs through dbGAP

Indexing

Data objects are assigned permanent globally unique IDs (GUIDs)to allow for access across tools, without requiring copies be created and transferred

Datasets are identified through faceted search over phenotypic data

Data Models

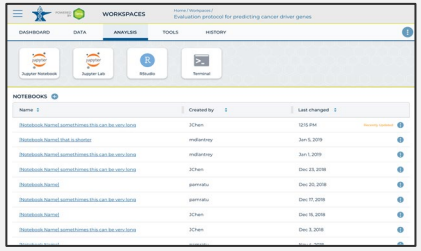
Functionality to support multiple, diverse data models is underway

These models are, where possible, being linked to external ontologies to improve consistency while maintaining diversity of datasets

NHGRI AnVIL

Architectural Diagram

All data use and analysis in a FISMA moderate environment

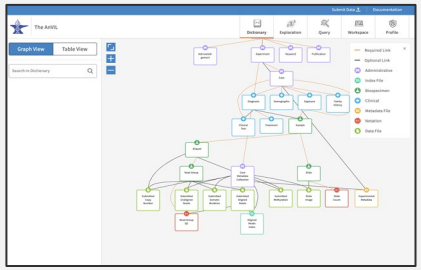


Terra

AnVIL / Terra: analysis workspaces and batch workflows

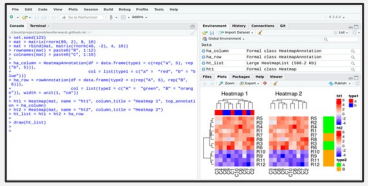
Dockstore
Create, Share, Use

AnVIL / Dockstore: sharing containerized tools and workflows



GEN3 Data Commons

AnVIL / Gen3: Data models, indexing, querying



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Galaxy

AnVIL / Analysis Environments: Jupyter Notebooks, RStudio, Galaxy, ...

FISMA
FEDERAL INFORMATION SECURITY MANAGEMENT ACT

FISMA Moderate
2 ATOs
Pursuing FedRAMP

Implemented on Google Cloud Platform