Opportunities for NIH cloud interoperability approaches to improve outcomes of pediatric diseases

Middleware and Grid Interagency Coordination Team (MAGIC)

April 7, 2021

Alisa Knodle Manning, PhD
amanning@broadinstitute.org

BROAD INSTITUTE

Presentation outline

1. Cloud analysis and researcher journey
2. NIH/NHLBI BioData Catalyst
3. NIH Cloud Platform Interoperability Effort (NCPI)
4. Important data governance lessons
5. Use Case: Pilot analysis to address scientific question

BROAD INSTITUTE

Cloud Analysis and Researcher Journey

BROAD INSTITUTE

We use the cloud for complex trait genetics analysis

Manning Lab
Massachusetts General Hospital
Broad Institute

2017 - 2018: First researchers to perform a GWAS using FireCloud

NIH/NIDDK K01
"Integrating diabetes pathophysiology from genotype to phenotype in whole genome sequence association studies of glycemic traits"

NIH/NHLBI's TOPMed Whole Genome Sequencing Program
- Diabetes Working Group
- Analysis Committee

TOPMed Cloud Computing Pilots on the FireCloud platform from the Broad Institute

2018 - 2019: Collaborative Development of Cloud-based Workflows

NIH/NHLBI R01
Large-scale Gene-environment Interaction
- Open-source statistical software tools
- Workflows
- Docker, github

User resources: GWAS in the cloud
- Featured Workspace in the Terra platform at the Broad Institute
- Workshop at American Society of Human Genetics meeting

2020: Collaborative analysis in NHLBI's BioData Catalyst

NIH/NHLBI's BioData Catalyst Program
- Principle Investigator (Broad Institute)
- Co-Chair of User Experience Working Group

BioData Catalyst - Fellows Cohort 1, 3; Public Cohort
- Gene-environment Interaction studies
- TOPMed Diabetes working group data harmonization efforts
- TOPMed 'omics analysis

Cloud IC Interoperability (CICI) Pilot Project
- Pilot and debug the process for cross-IC cloud platform analysis

BROAD INSTITUTE

Problem: User with a research question and an analysis plan

Pre-interoperability effort

Find Data

- Data Access Request on dbGAP, NIH's controlled access data portal

Set up place to do analysis

Authorization to use data

- Institution manages training, policies
- Principle Investigator manages research team

BROAD INSTITUTE

Problem: User with a research question and an analysis plan

Pre-interoperability effort

Find Data

- Data Access Request on dbGAP, NIH's controlled access data portal
- **Manual dbGap data download, decrypt and organize files for analysis**

Set up place to do analysis

- **Local and Institutional compute (computer, high performance cluster)**

Authorization to use data

- Institution manages training, policies
- Principle Investigator manages research team

BROAD INSTITUTE

Problem: User with a research question and an analysis plan

Pre-interoperability effort

Find Data

- dbGAP project proposal
- **Manual dbGap data download, decrypt and organize files for analysis**

Set up place to do analysis

- **Local and Institutional compute (computer, high performance cluster)**

Authorization to use data

- Institution manages training, policies
- Principle Investigator manages research team

Current paradigms

Find Data

- dbGAP project proposal
- **Web portal in NHLBI's BioData Catalyst**
  - **Indexes NHLBI's BioData Catalyst data**
  - **Organizes files associated with approved dbGap projects**
  - **Allows user to export links to data to an analysis workspace**

Set up place to do analysis

- **Analysis workspace in NHLBI's BioData Catalyst**
  - **Tools, Apps, Workflows to enable analysis**

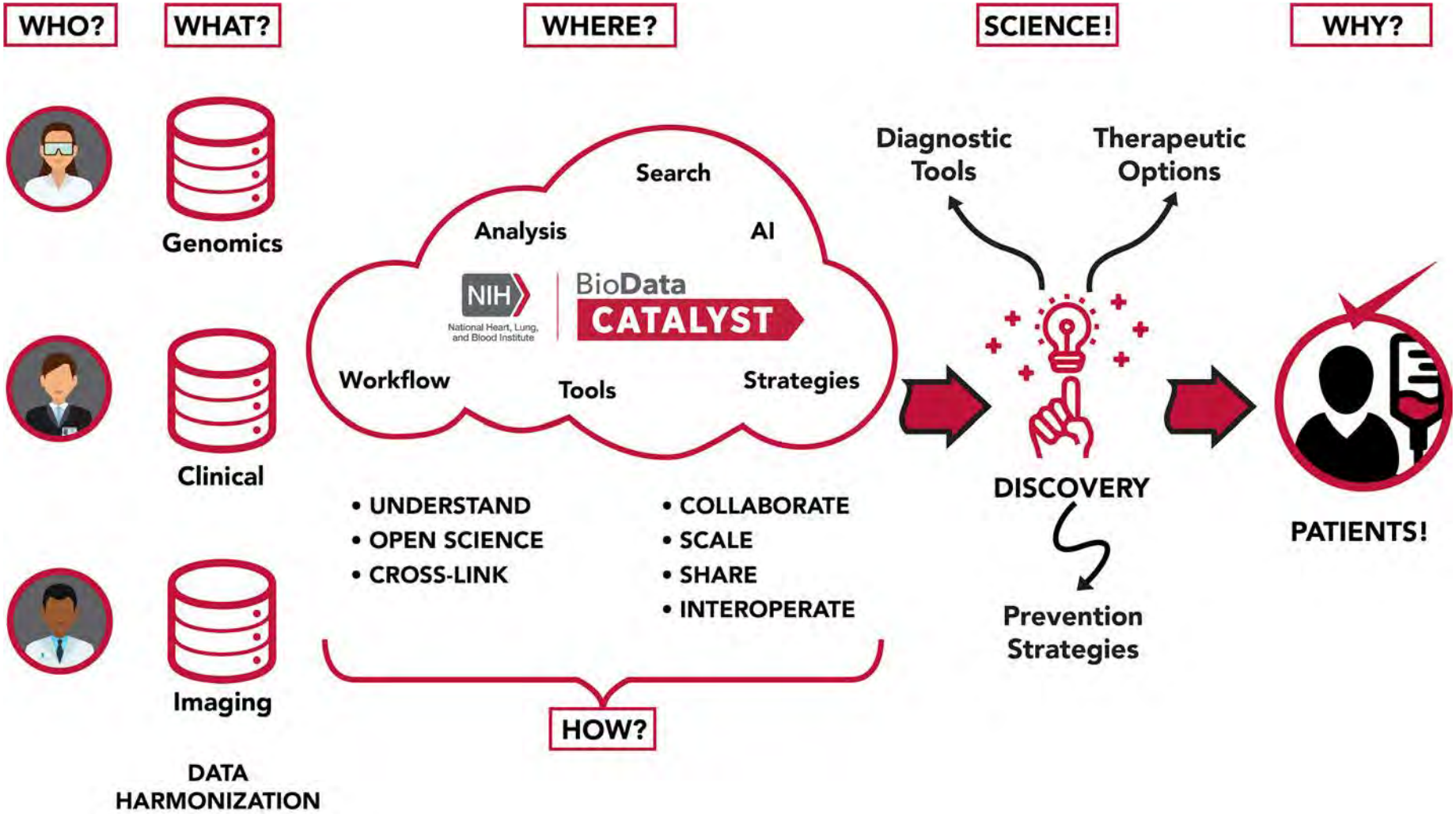
Authorization to use data

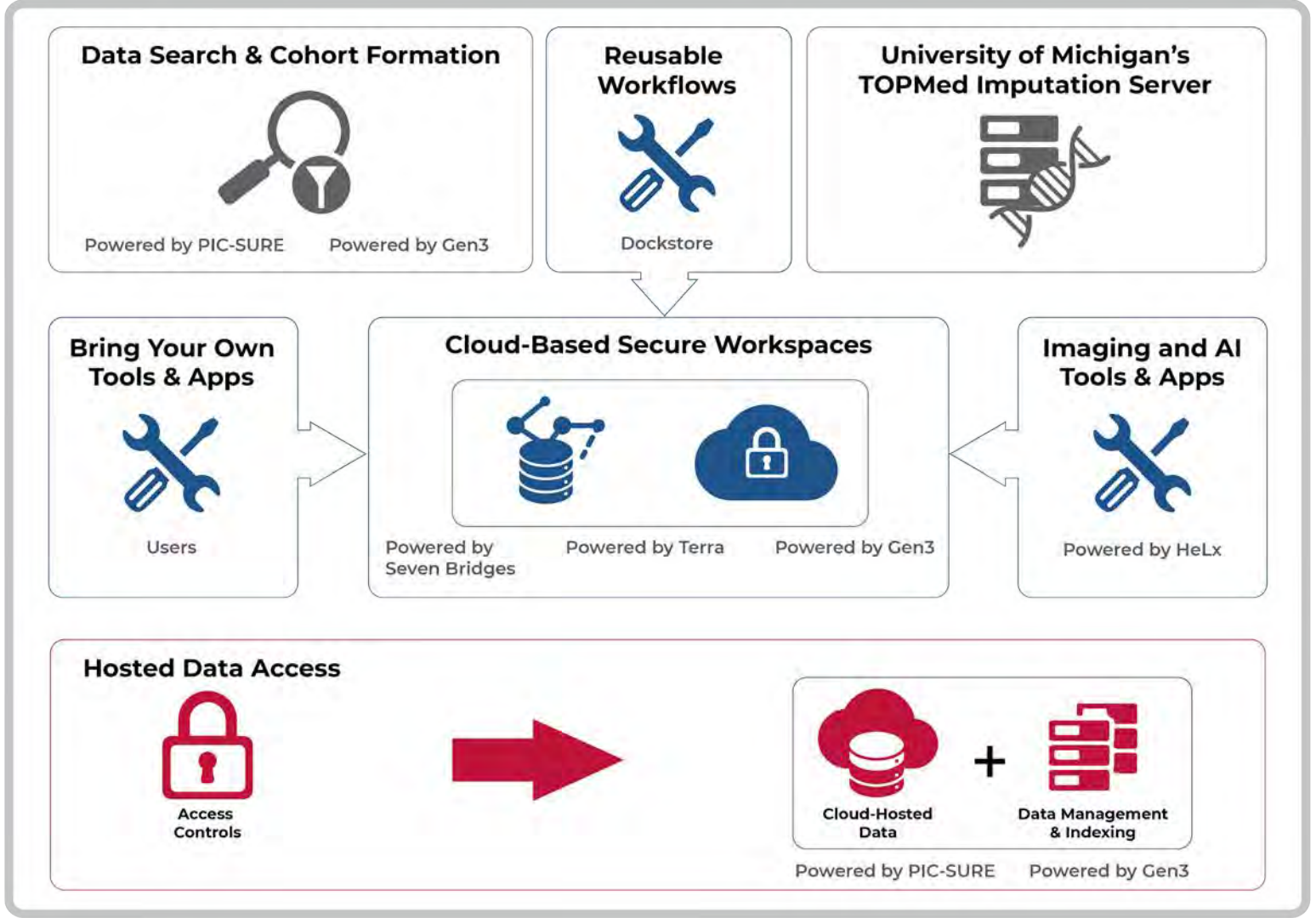
- Institution manages training, policies
- Principle Investigator manages research team
- **Automatic check for permission to use data when accessed from analysis workspace**

BROAD INSTITUTE

BioData Catalyst

BROAD INSTITUTE


WHO?
WHAT?
WHERE?
SCIENCE!
WHY?
Genomics
Clinical
Imaging
DATA HARMONIZATION
Search
Analysis
AI
NIH
National Heart, Lung, and Blood Institute
BioData CATALYST
Workflow
Tools
Strategies
• UNDERSTAND
• OPEN SCIENCE
• CROSS-LINK
• COLLABORATE
• SCALE
• SHARE
• INTEROPERATE
HOW?
Diagnostic Tools
Therapeutic Options
DISCOVERY
Prevention Strategies
PATIENTS!


Data Search & Cohort Formation
Powered by PIC-SURE
Powered by Gen3
Reusable Workflows
Dockstore
University of Michigan's TOPMed Imputation Server
Bring Your Own Tools & Apps
Users
Cloud-Based Secure Workspaces
Powered by Seven Bridges
Powered by Terra
Powered by Gen3
Imaging and AI Tools & Apps
Powered by HeLx
Hosted Data Access
Access Controls
Cloud-Hosted Data
Data Management & Indexing
Powered by PIC-SURE
Powered by Gen3

NHLBI BioData Catalyst

National Heart, Lung, and Blood Institute

Providing strategic leadership and funding the researchers and other professionals developing the ecosystem.

Director: Gary Gibbons
CIO: Alastair Thomson
Program Officer: Jon Kaltman

Steering Committee

Providing strategic decision-making and achieving consensus for the Consortium.

Ingrid Borecki (Chair), Principal Investigators, NHLBI Working Group

External Expert Panel

Independently informing and advising the work of the Consortium.

Donna Arnett
Mark Craven
Jason Williams
David Mendelson
Warren Kibbe

Coordinating Center

Coordinating project management, communications, project reporting, and collaboration standards.

Ahalt, Boyles

Data Stewards

Partnering with the Consortium on data accessibility and interoperability.

TOPMed, COPDGene

The Broad Institute, University of Chicago, University of California, Santa Cruz, Vanderbilt University Medical Center

Grossman, Manning, Paten, Philippakis

Providing authorized access and faceted search of harmonized data across studies, genomic analysis and visualization in virtual workspace, and high-quality Docker-based research tools.

Harvard Medical School

Avillach

Exploring data with interactive search and visualizations for feasibility assessment and providing data science tools to access and analyze clinical and genomic data.

RTI International, UNC-CH/RENCI

Bradford, Cox, Krishnamurthy

Developing tools and apps for machine learning; deep learning models; semantic search; and visualizing, annotating and analyzing biomedical images. Developing methods for tool and app creation to enhance the ecosystem.

Seven Bridges Genomics

Davis-Dusenbery

Finding, accessing, analyzing TOPMed genomics data at scale; bringing your workflows or choosing from hosted CWL tools; performing association studies with tooling for variant aggregation.

BROAD INSTITUTE

NIH Cloud Platform Interoperability Effort (NCPI)

BROAD INSTITUTE

NIH Cloud Platform Interoperability Effort (NCPI)

Establish and implement guidelines and technical standards to empower end-user analyses across participating platforms and facilitate the realization of a trans-NIH, federated data ecosystem.

The ***Systems Interoperation Working Group*** will spearhead technical improvements to the NCPI participating platforms that enable improved interoperability.

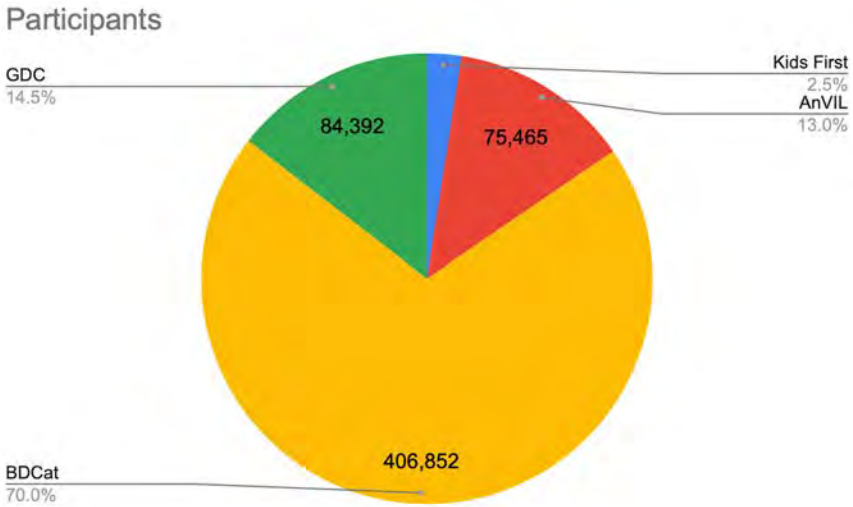
NCPI's additional participating platforms:

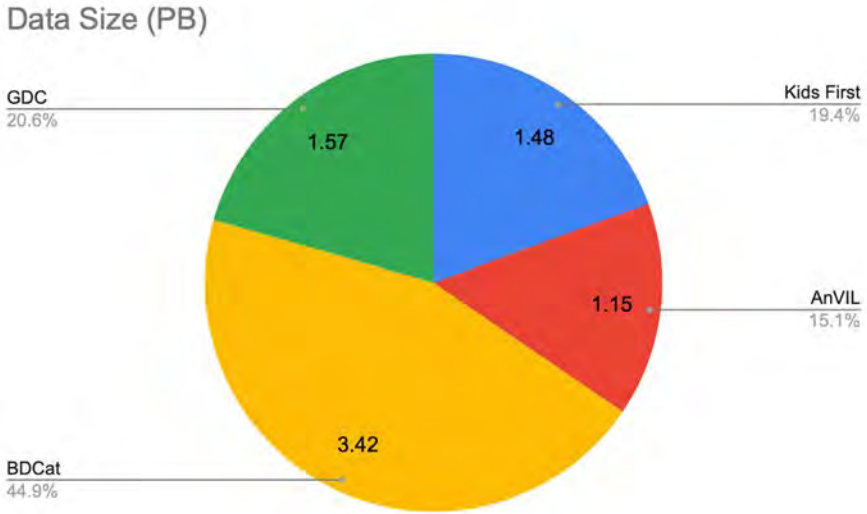
- NHGRI's Genomic Analysis, Visualization, and Informatics Lab-space (AnVIL)
- The NIH Common Fund's Gabriella Miller Kids First Pediatric Research Program ("Kids First")
- National Cancer Institute's Cancer Research Data Commons (CRDC)

BROAD INSTITUTE

https://anvilproject.org/ncpi

Motivation: Researchers want to access data across ICs/stacks.





Aggregation of data across these IC stacks is huge ~7.6PB

BROAD INSTITUTE

Slide courtesy of Brian O'Connor, Broad Institute

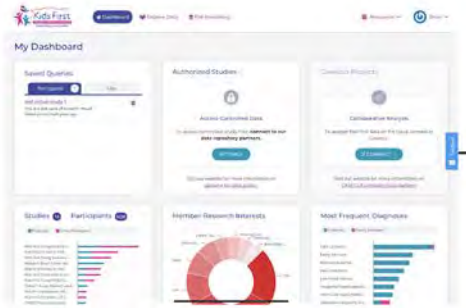
Interoperability between NIH Stacks in early 2020

Data portals connect (**intra-IC**) with analysis systems (workspaces)

Cloud Resources

BioData CATALYST
Powered by Gen3

BioData CATALYST
Powered by Terra



BROAD INSTITUTE

Slide courtesy of Brian O'Connor, Broad Institute

Standards-based Interoperability Features

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

https://www.ga4gh.org/genomic-data-toolkit/

GA4GH Passports v1

A GA4GH-approved Standard The GA4GH Passport specification aims to support data access policies within current and evolving data access governance systems. This specification defines Passports and Passport Visas as the standard way of communicating the data access authorizations that a user has based on either their role (e.g. researcher), affiliation, or access status.

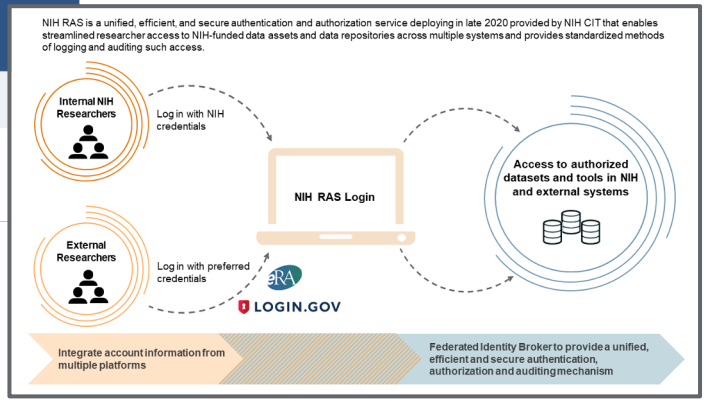
Data Repository Service v1

A GA4GH-approved Standard The Data Repository Service (DRS) API, a standard for building data repositories and adapting access tools to work with those repositories, works with other approved APIs from the GA4GH Cloud Work Stream to allow researchers to discover algorithms across different cloud environments and send them to datasets they wish to analyze. The API allows data consumers to access datasets regardless of the repository in which they are stored or managed.

NIH National Institutes of Health
Turning Discovery Into Health

NIH LOGIN SERVICES (Formerly ITRUST)

- **Researcher Auth Service (RAS) integration**. RAS utilizes OAuth2/OIDC for GA4GH compliant integrations of applications for researcher access to NIH data repositories and systems.



NIH Researcher Auth Service 1.0: Conceptual Overview

https://datascience.nih.gov/researcher-auth-service-initiative

BROAD INSTITUTE

Portable Format for Bioinformatics (PFB) Developed by the Gen3 Data Commons

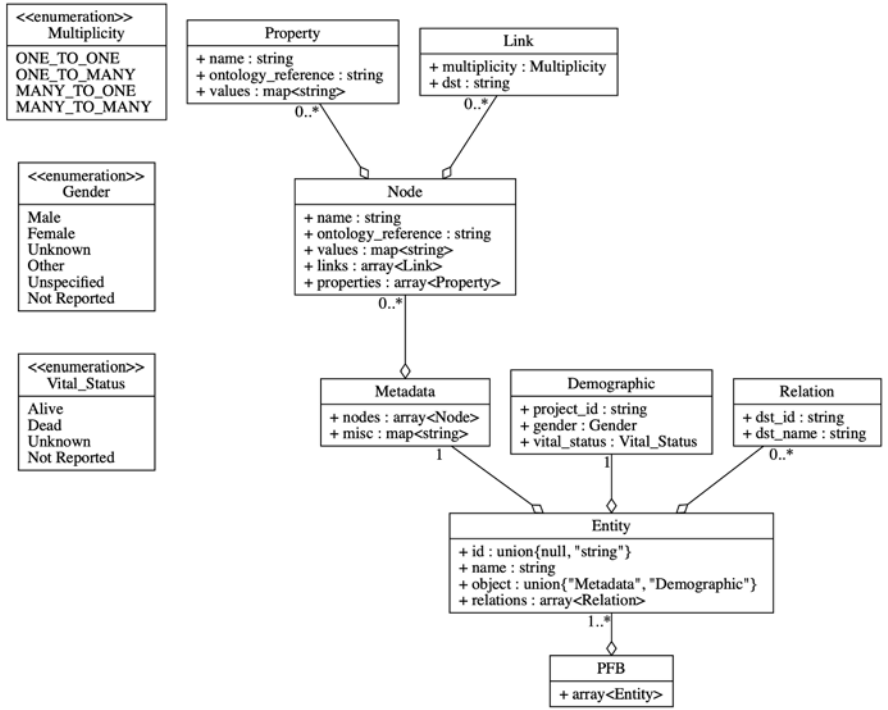
GEN3 DATA COMMONS

What was PFB designed for?

- A self-describing data model + serialization format

What does PFB transport?

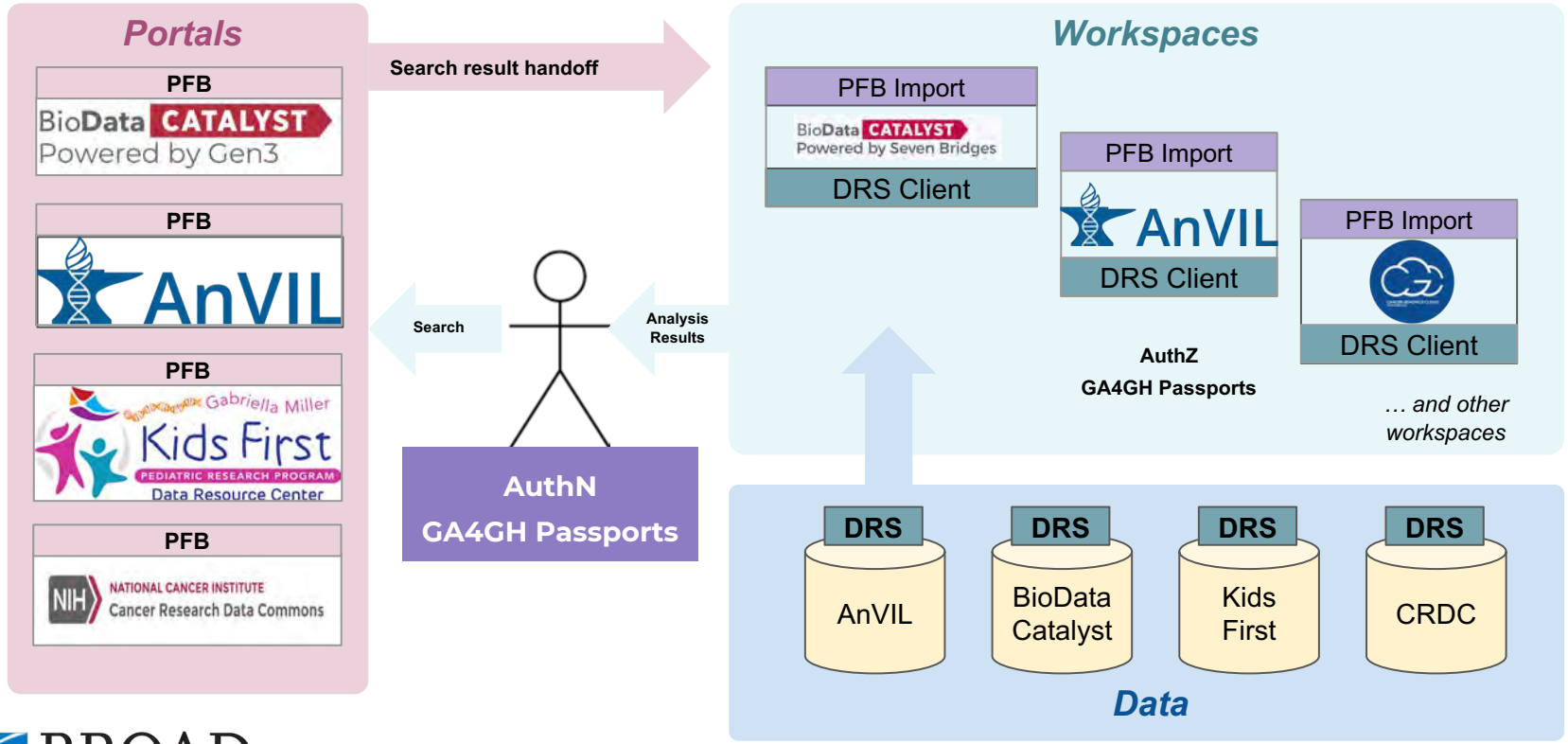
- The schema of the structure of the data
- The data itself



BROAD INSTITUTE

Slide courtesy of Brian O'Connor, Broad Institute

BioData Catalyst NCPI Interoperability Model (2020)



BROAD INSTITUTE

Slide courtesy of Brian O'Connor, Broad Institute

NCPI Systems Interoperation Working Group - 2020 Accomplishments

*Achieved improved interoperability in 2020 across multiple systems through **PFB**, **GA4GH DRS**, and **GA4GH Passports**.*

- **Search Result Handoff:** PFB

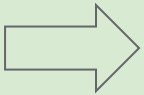
  2 portals
  ~417K subjects accessible

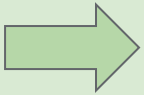
- **Data Access:** DRS 1.1

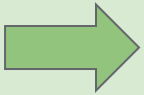
  4 DRS Servers
  ~7.6PB of data

- **Auth**: RAS for AuthN

  RAS







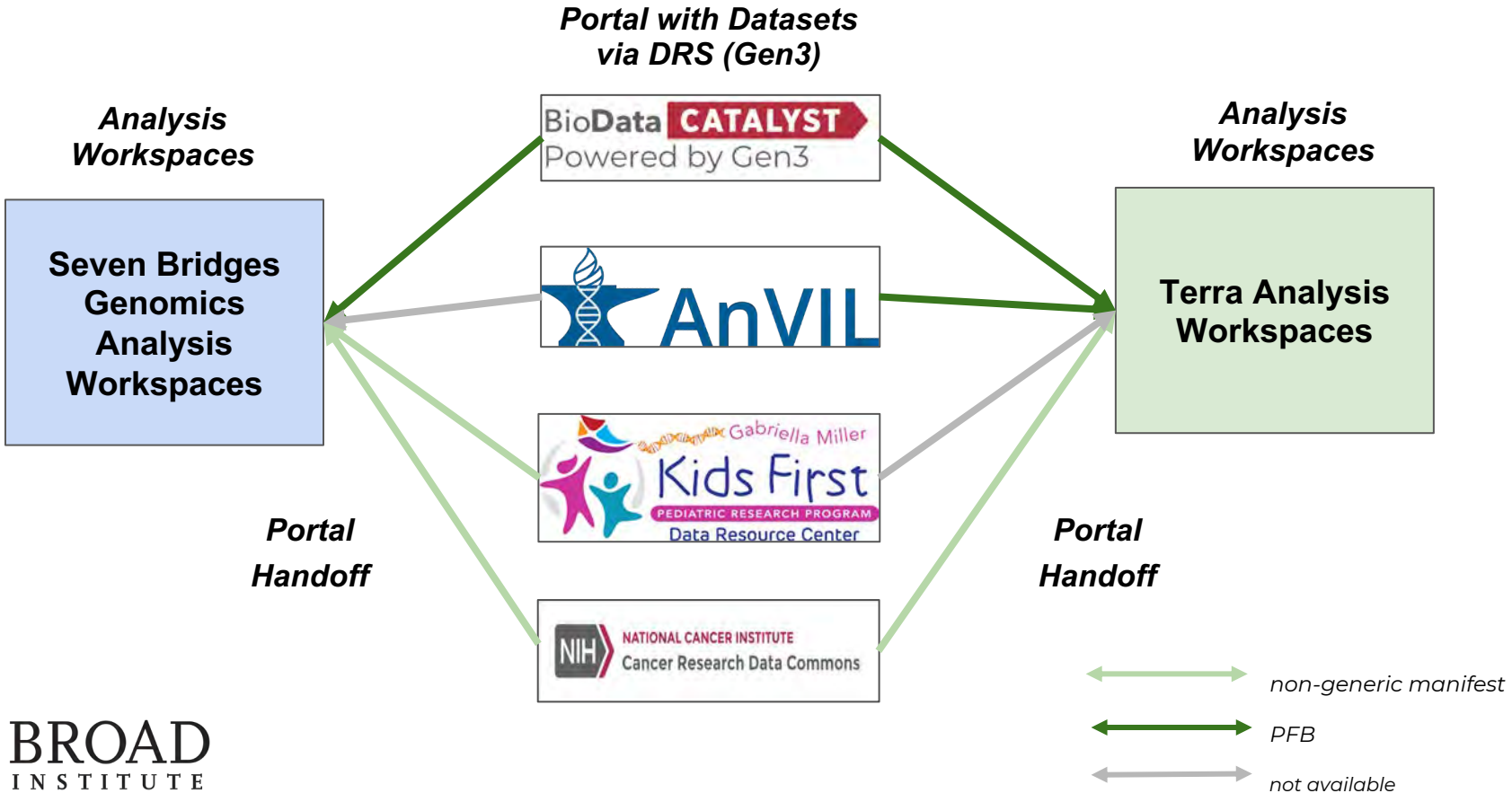
Supported Platforms

- The **NHGRI AnVIL** and **NHGRI BioData Catalyst** portals both support handoff of search results to **workspaces** (Terra, Gen3, SBG)
- Data accessible on **AnVIL, BDCat, CRDC,** and **Kids First** via **DRS 1.1** support
- **GA4GH Passports** are in use by **RAS** and support visas from dbGaP made accessible by Gen3.

BROAD INSTITUTE

Slide courtesy of Brian O'Connor, Broad Institute

NCPI Systems Interoperation Working Group - Early 2021 - Finish Connections



BROAD INSTITUTE

Important data governance lessons

BROAD INSTITUTE

Bring together the Pediatric Cardiac Genetics Consortium Study data for the first time in the cloud for researchers

Pediatric Cardiac Genetics Consortium Study (PCGC; phs00119)

- Observational study of participants with congenital heart defects (CHD).

Gabriella Miller Kids First Pediatric Research Program (GMKF)

- Stores a subset of the PCGC project, representing whole genome sequences from over 2000 participants (phs001138).

The NHLBI's TOPMed program

- Includes an additional subset of the PCGC project, representing up to 3230 participants with whole genome sequence data (phs001735).

Platform	Datasets	dbGaP	Sample
AnVIL	GTEx	phs000424.v8.p2	980
Kids First	PCGC	phs001138.v3.p2	699
BioData Catalyst	TOPMed PCGC	phs001735, phs001194.v2.p2	1,901
	FHS	phs000974.v4.p3, phs000007.v30.p11	4,155
	JHS	phs000964.v4.p1	2,777

BROAD INSTITUTE

Rebecca Boyles (RTI), Alisa Manning (Broad), Steve Cox (RENCI), Alison Heath (CHOP)

Challenge of enabling this cross platform data access is in maintaining each program's data governance

Location of Data

- Kids First / Cavatica: Amazon Web Services (AWS)
- TOPMed / BioData Catalyst: Google Cloud Platform (GCP)

Requirement is to maintain:

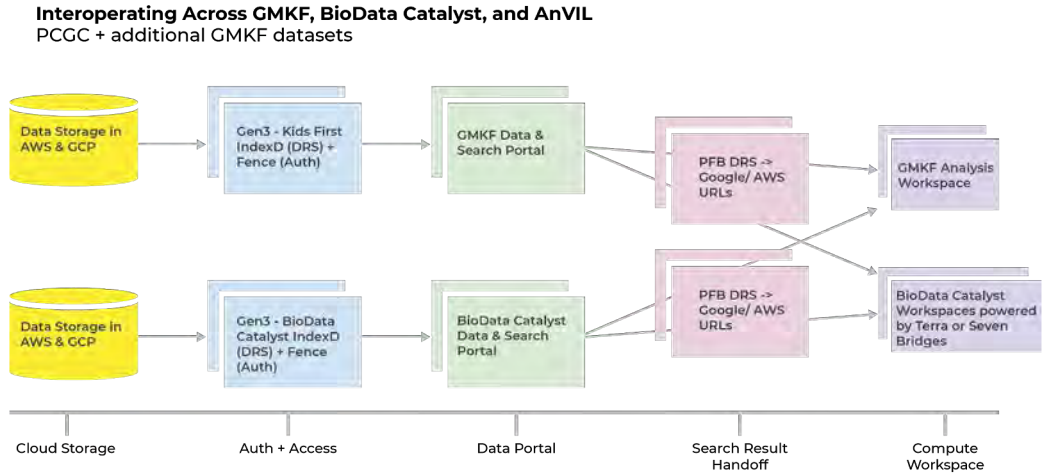
- Roles of each data steward
- Data steward's control of Authentication and Authorization Infrastructure (AAI)
- Data steward's management of data access through index service (DRS)

BROAD INSTITUTE

Rebecca Boyles (RTI), Alisa Manning (Broad), Steve Cox (RENCI), Alison Heath (CHOP)

Interim solution for data access for analysis in GMKF Data Resource or BioData Catalyst workspace

- TOPMed PCGC (phs001735) and GMKF PCGC (phs001138) data reside in respective cloud buckets
- Each program maintains own index and authorization tool
- Users will see both TOPMed PCGC (phs001735) and GMKF PCGC (phs001138) in either portal interface
    - Data can be searched and exported with the PFB convention
    - GA4GH DRS will be used to access the data



Draft graphic

BROAD INSTITUTE

Rebecca Boyles (RTI), Alisa Manning (Broad), Steve Cox (RENCI), Alison Heath (CHOP)

Interim solution for data access for analysis in GMKF Data Resource or BioData Catalyst workspace

Our goal is to enhance the interoperability capability within BioData Catalyst

- Remove the need to have data mirrored in GCP and AWS clouds.
- BioData Catalyst will provide ***signed URLs*** to all data to authorized researchers.
- Avoids data egress charges since GMKF currently provides free egress from AWS while providing access to all “Kids First” data to BioData Catalyst

We are working on a paper describing our novel approach

BROAD INSTITUTE

Rebecca Boyles (RTI), Alisa Manning (Broad), Steve Cox (RENCI), Alison Heath (CHOP)

Acknowledgements

Brian O'Connor

Asia Mieczkowska

Becky Boyles

Patrick Patton

Steven Cox

Michael Baumann

Andrew Rula

Alex Baumann

Allison Heath

David Higgins

Maia Nguyen

- Gabriella Miller Kids First Pediatric Research Program
- Pediatric Cardiac Genomics Consortium (PCGC)
- Genotype-Tissue Expression (GTEx) project
- TOPMed's PCGC's Congenital Heart Disease Biobank
- Framingham Heart Study
- Jackson Heart Study

Funding:

National Institutes of Health, National Heart, Lung, and Blood Institute, through the BioData Catalyst program (award 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01).

Cloud IC Interoperability (CICI)

BROAD INSTITUTE

Use Case:

Pilot analysis to address scientific question

BROAD INSTITUTE

SCIENTIFIC QUESTION:

Investigate genetic factors related to congenital heart defects in a study design that uses healthy controls from two NHLBI cohorts. Perform pooled analysis on AnVIL powered by Terra.

BROAD INSTITUTE

Find Data

Pediatric Cardiac Genomics Consortium (PCGC)

- Observational study of participants with congenital heart defects
- Started in 2009 by the NHLBI to learn more about why children are born with heart disease
- https://benchtobassinet.com/

Framingham Heart Study (FHS)

- longitudinal population cohort of participants and their offspring who had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke and who have been followed over many years
- https://framinghamheartstudy.org/

Jackson Heart Study (JHS)

- “Largest single-site, community-based epidemiologic investigation of environmental and genetic factors associated with cardiovascular disease among African Americans ever undertaken”
- https://www.jacksonheartstudy.org/

The Genotype-Tissue Expression (GTEx) project

- “Ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation”
- https://www.gtexportal.org/

SCIENTIFIC QUESTION:

Investigate genetic factors related to congenital heart defects in a study design that uses healthy controls from two NHLBI cohorts. Perform pooled analysis on AnVIL powered by Terra.

Set up place to do analysis

- NHLBI: BioData Catalyst
- NIH Common Fund: Gabriella Miller Kids First Data Resource Center (“Kids First”)
- NHGRI: Genomic Data Science The Genomic Analysis, Visualization, and Informatics Lab-space (AnVIL)

Authorization to use data

- Approved dbGap application

Data availability

Pediatric Cardiac Genomics Consortium (PCGC)

- Whole genome sequencing
    - NHLBI’s TOPMed Program
    - NIH Common Fund Gabriella Miller Kids First Pediatric Research Program

Framingham Heart Study (FHS)

- Whole genome sequencing by NHLBI’s TOPMed Program

Jackson Heart Study (JHS)

- Whole genome sequencing by NHLBI’s TOPMed Program

The Genotype-Tissue Expression (GTEx) project

- Transciptome sequencing in samples collected from 54 non-diseased tissue sites across nearly 1000 individuals

BROAD INSTITUTE

Genetics of congenital heart defects: improving outcomes of pediatric diseases

Study aims:

1. Identify, access, and summarize available genetic and phenotypic data on native cloud platforms
2. Leverage individual-level data from multiple cloud platforms to assess rare variants contributing to congenital heart defect risk

Method: Proxy External Controls Association Test (ProxECAT)

Compare ratio of rare, synonymous and nonsynonymous variants per gene between cases and controls

Approach:

Look for genes with ProxECAT signal using Whole Genome Sequence data sets

- Internal cases (PCGC)
- External controls (FHS/JHS)

Explore gene expression profiles of most interesting genes

- Gene expression from many tissues in a small cohort (GTEx)

Platform	Datasets	dbGaP	Sample
AnVIL	GTEx	phs000424.v8.p2	980
Kids First	PCGC	phs001138.v3.p2	699
BioData Catalyst	TOPMed PCGC	phs001735, phs001194.v2.p2	1,901
	TOPMed FHS	phs000974.v4.p3, phs000007.v30.p11	4,155
	TOPMed JHS	phs000964.v4.p1	2,777

BROAD INSTITUTE

ProxECAT: *Hendricks et al (2018). doi: 10.1371/journal.pgen.1007591*

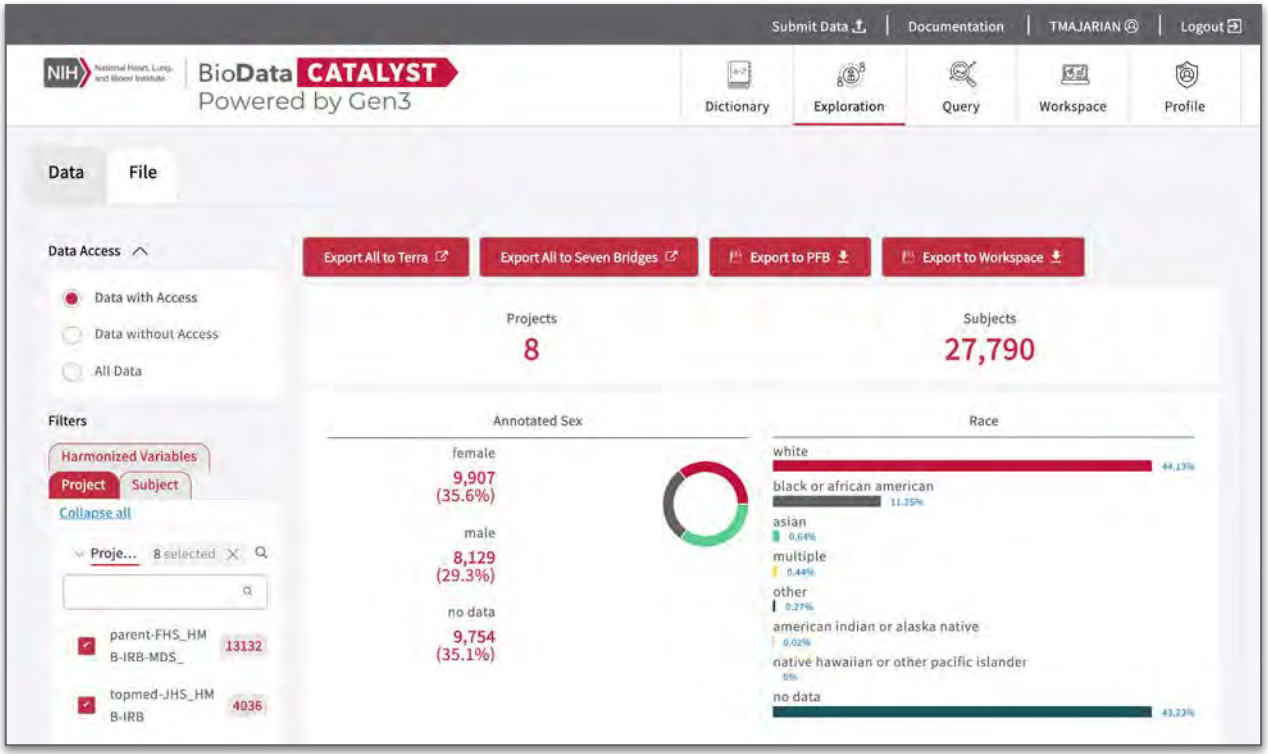
Use Case: NHGRI AnVIL + Kids First DRC + NHLBI BioData Catalyst

Researchers: Alisa Manning, Brian O'Connor, Timothy Majarian
Institution: Broad Institute
Platforms: NHLBI's BioData Catalyst, Kids First and NHGRI's AnVIL

Platform	Program	Target Datasets	dbGap Accession(s)	Data Use Regulations (Consent) groups	Sample Size
AnVIL	GTEx	GTEx	phs000424.v8.p2	GRU (General Research Use)	980
Kids First		PCGC	phs001138.v3.p2	HMB (Health / Medical / Biomedical)	202
BioData Catalyst	TOPMed	PCGC_CHD	phs001735, phs001194.v2.p2	HMB	1,901
		Framingham Heart Study (FHS)	phs000974.v4.p3, phs000007.v30.p11	HMB-IRB (Institutional Review Board)-MDS	3,555
				HMB-IRB-NPU(Non-profit Use)-MDS	600
		Jackson Heart Study (JHS)	phs000964.v4.p1	HMB-IRB	2,018
				HMB-IRB-NPU	759

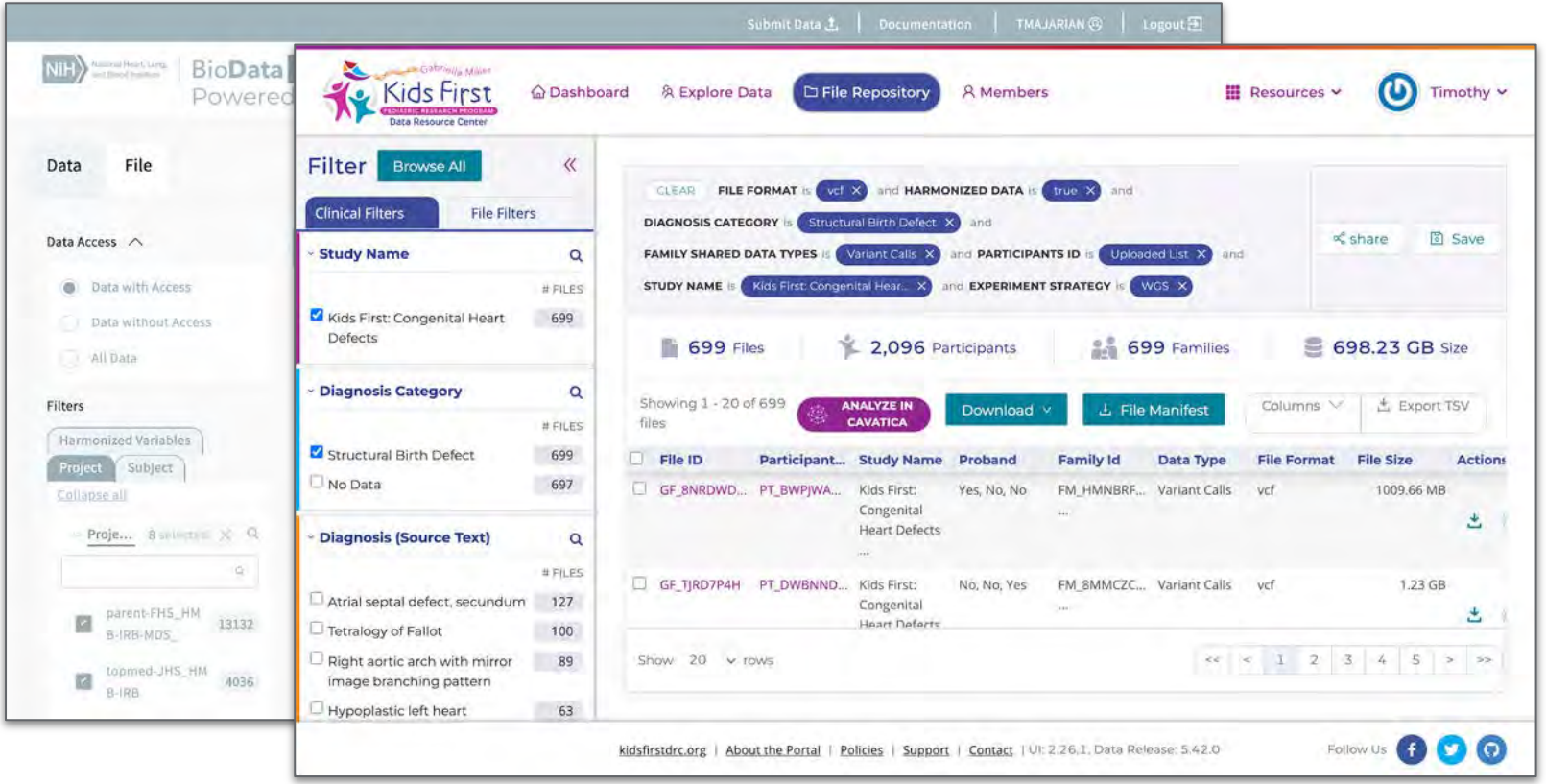
BROAD INSTITUTE

Export to native cloud platforms



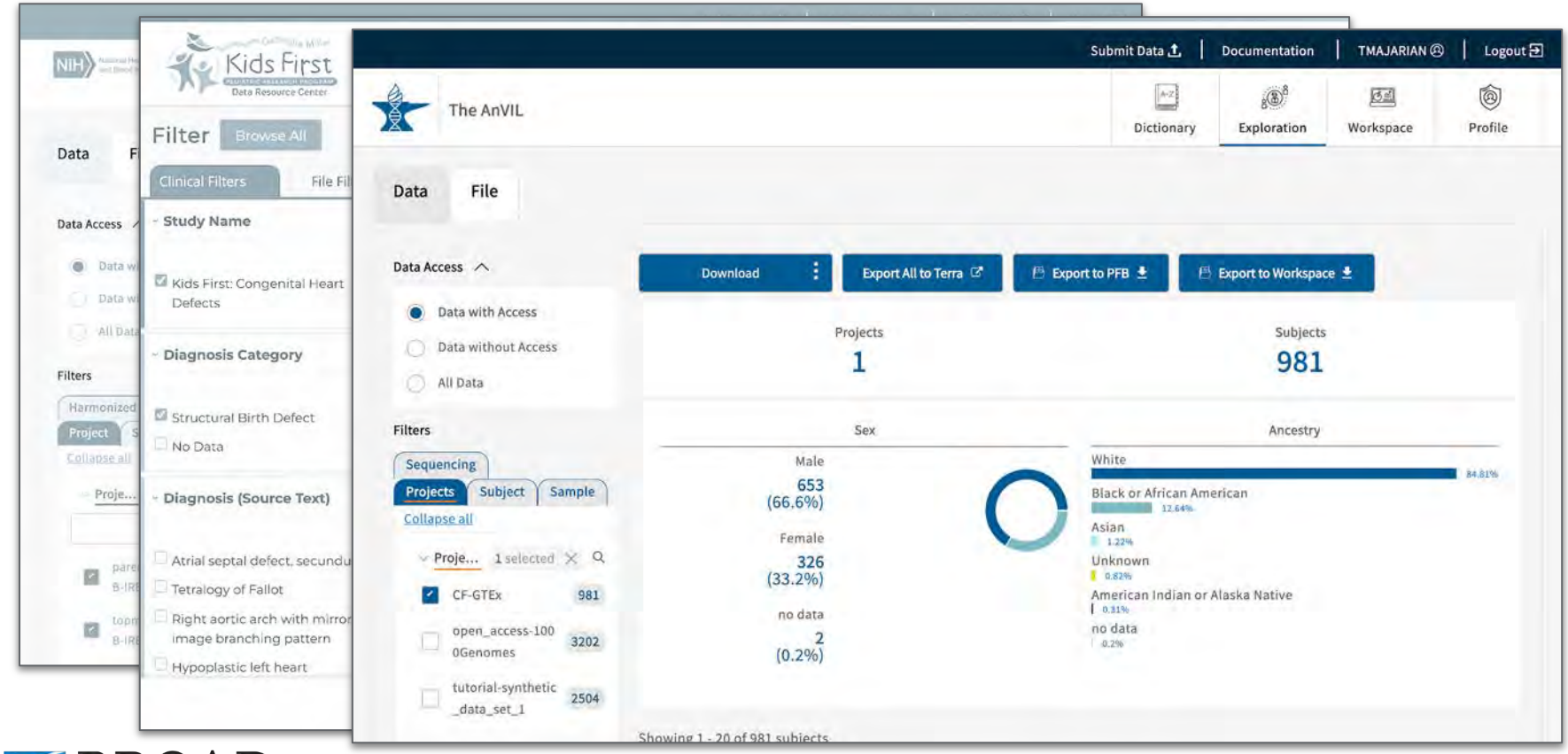
BROAD INSTITUTE

Export to native cloud platforms



BROAD INSTITUTE

Export to native cloud platforms



BROAD INSTITUTE

Preparation of genetic data for association analysis

All preparation steps were performed within separate ecosystems

1. Kids First -> Cavatica
2. BioData Catalyst -> Terra
3. AnVIL -> Terra

Variants included in analysis:

- MAF < 1%
- Protein coding exonic

Variant annotation - Synonymous and non-synonymous

- ANN field in VCF files for Kids First
- DBSNFP for JHS and FHS

For each protein coding gene

- Count synonymous and non-synonymous variants
- Separated by cases (Kids First) and controls (JHS and FHS)

ANN: *annotation* field

- Predicted variant effect on gene expression or protein function

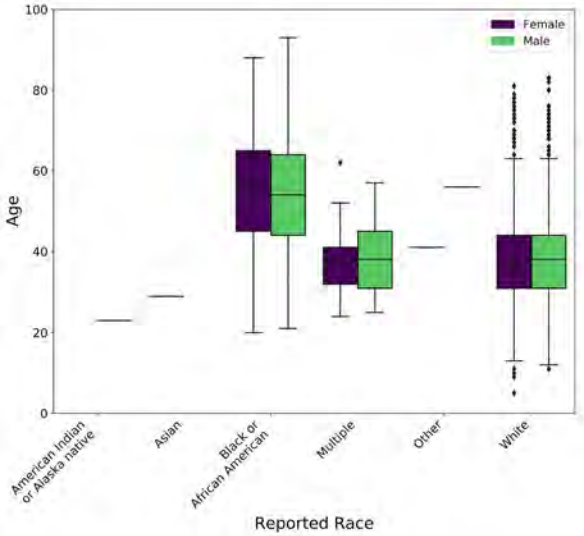
DBSNFP:

- Database of functional predictions for all coding variants
- Includes same variant effect predictions as ANN field

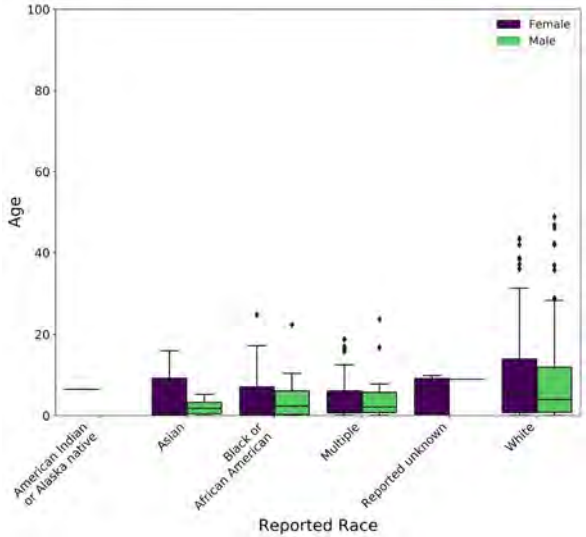
BROAD INSTITUTE

Platform-specific summaries

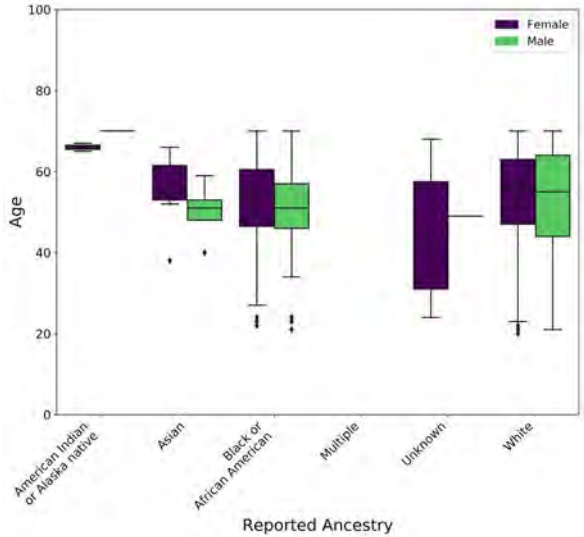
NIH National Heart, Lung, and Blood Institute | BioData CATALYST Powered by Terra



Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM | CAVATICA



AnVIL GTEx



BROAD INSTITUTE

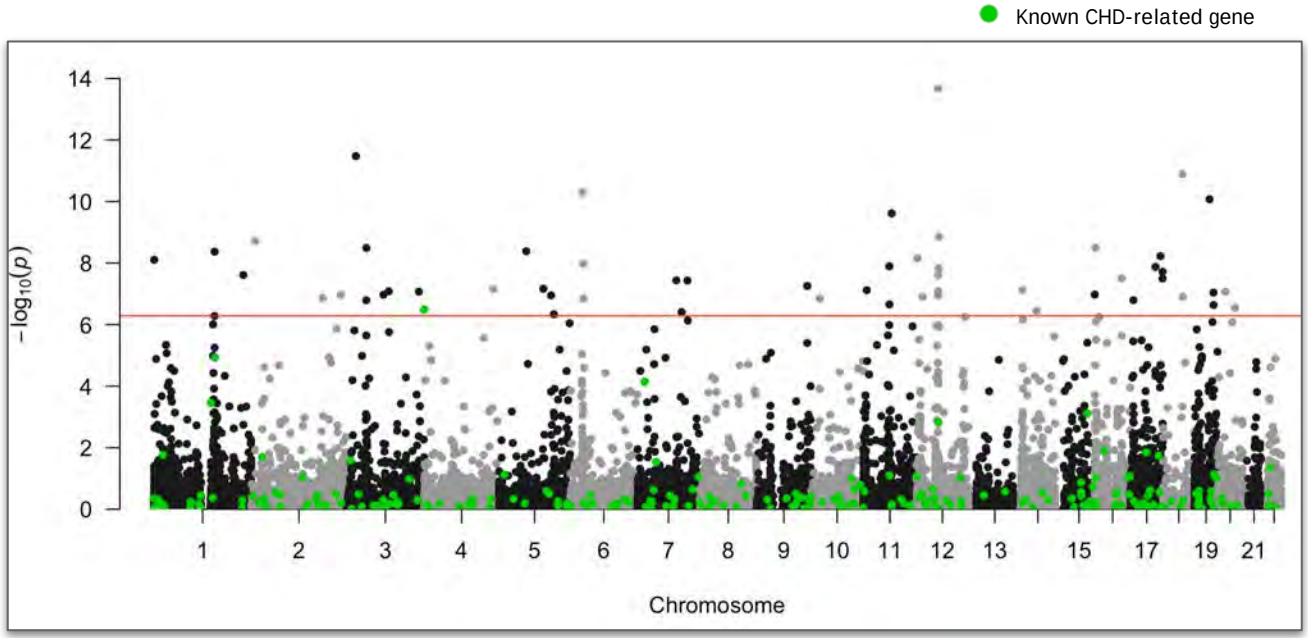
ProxECAT results

NIH National Heart, Lung, and Blood Institute | BioData CATALYST Powered by Terra

Association analyses were performed within the BioData Catalyst ecosystem

Kids First data was manually downloaded and uploaded to a BioData Catalyst workspace

- 17,285 genes tested
- 55 genes with $P<5x10^{-7}$
- 1 known Congential Heart Defects gene with $P<5x10^{-7}$



BROAD INSTITUTE

Then vs Now vs Future

Pre-interoperability effort

Data authorization
- Obtain dbGaP access
- Log into dbGaP
- Create download request

Access and localization to cloud platform
- Start GCS VM
- Download data via Aspera
- Upload data to GCS bucket
- Access through Terra workspace

Data preprocessing & Final analysis
- Single Terra workspace

Current paradigms

Data authorization
- Obtain dbGaP access

Access and localization to cloud platform
- ERA credentials through Gen3 or KFDR
- Export data links (DRS) within a individual ecosystems

Data preprocessing
- Separate workspaces within individual ecosystems

Final analysis
- Single BDC workspace
- Download & upload KFDR data for analysis

Future

Data authorization
- Obtain dbGaP access

Access and localization to cloud platform
- Single sign in within a BDC ecosystem

Data preprocessing
- One BDC workspace for all data

Final analysis
- One BDC workspace
- No download and upload

BROAD INSTITUTE

Stumbles and roadblocks

Data availability across platforms - KFDR (Cavatica) to BDC (Terra)

PFB import to Terra - TOPMed PCGC (BDC) [**SOLVED**]

DRS links - GTEx (AnVIL) [**SOLVED**]

Workflow compatibility - CWL (Cavatica) vs. WDL (Terra)

Data documentation: Data are easy to access but finding exactly how the data were generated remains difficult

Ex: Why is the ANN field missing in the TOPMed cohort-level VCFs?

Ex: What fields are included in genetics data and what do they mean?

Ex: What methods were used for genotype calling? (KFDR vs. TOPMed)

BROAD INSTITUTE

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov