FHIR WG

Robert Carroll (Vanderbilt University Medical Center)
Allison Heath (Children's Hospital of Philadelphia)

Overview

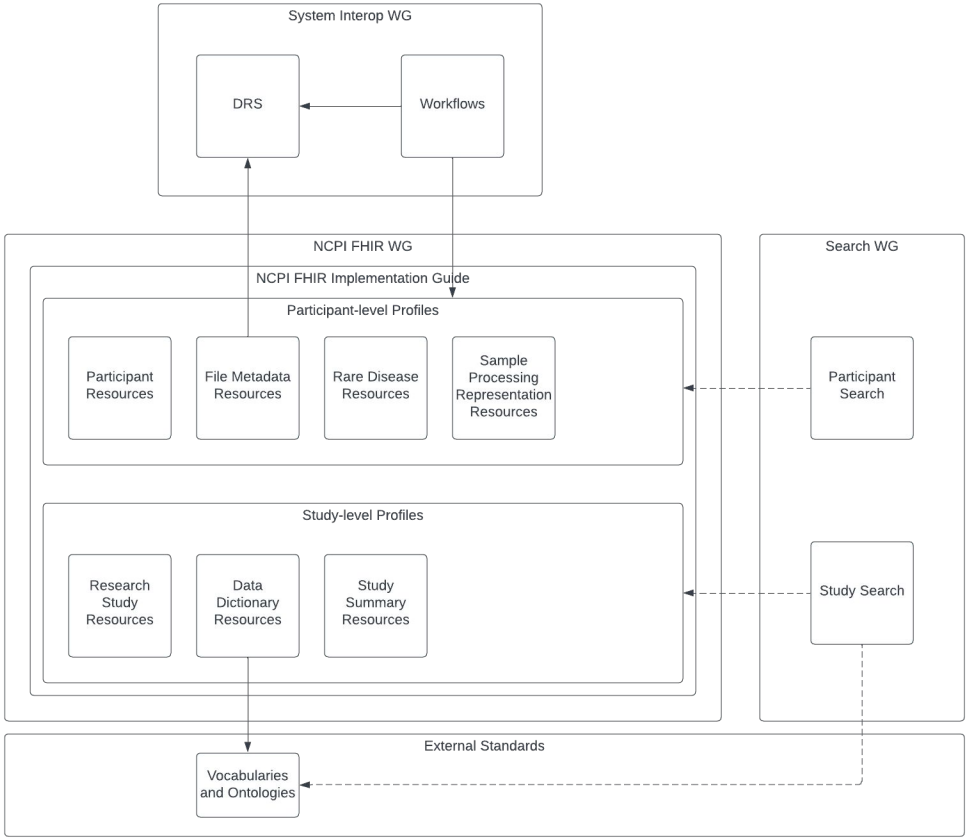
- Objectives for FHIR
- FHIR Service Deployment
- FHIR Implementation Guide v0.1 Complete
- Refactoring our approach- IG v0.2
- FHIR Code-a-thon next week!

Objectives of FHIR

1. To provide an API to allow access to study and participant level data.
2. To provide standardized structures for study and participant data.
3. To enable structured semantics for data where available.

While there are solutions to some of these problems across NCPI, FHIR is an international standard with broad support across academics and vendors (including cloud providers) that provides methods to address all of them.

Objectives of FHIR

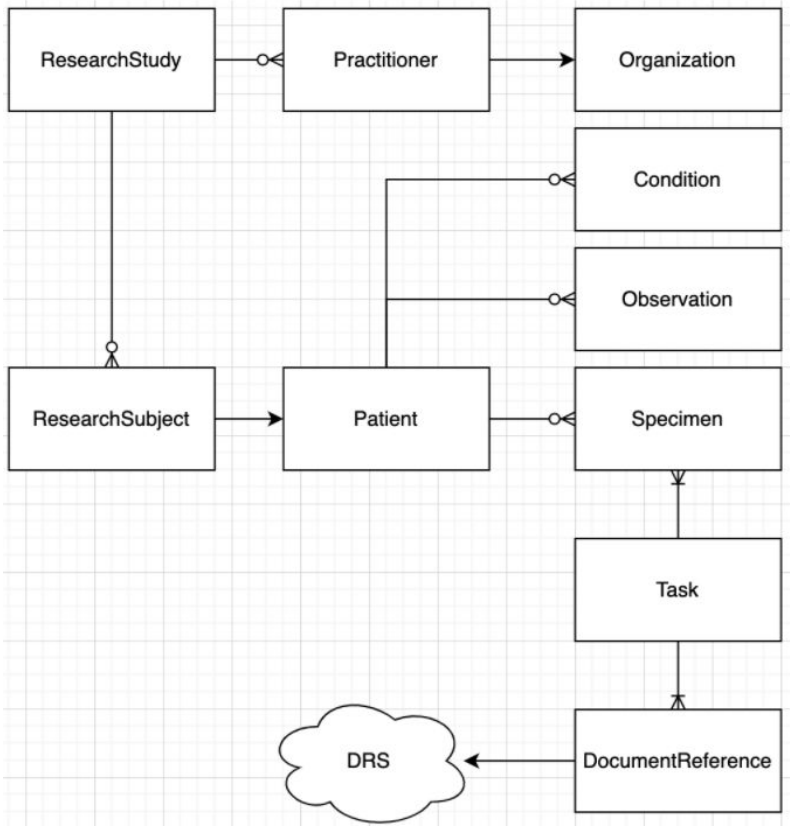


FHIR Service Deployment

- Formal NCPI Teams
  - Kids First: Production FHIR Services deployed
    - https://kf-api-fhir-service.kidsfirstdrc.org/
    - Open access data, requires login to KF Portal
  - dbGaP: Public data services deployed
    - https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/
    - Study level data only
    - Work in progress on controlled access data, pilot implementations complete
  - AnVIL: Non-production service pilots
    - Test deployment indexing AnVIL data across Terra
    - Pilot study specific ETL
- Highlighted community groups
  - ImmPort: Developed IG and have deployed services, includes dev service: https://fhir.dev.immport.org/
  - INCLUDE DCC: Production FHIR service with registered user data access: https://include-api-fhir-service.includedcc.org/

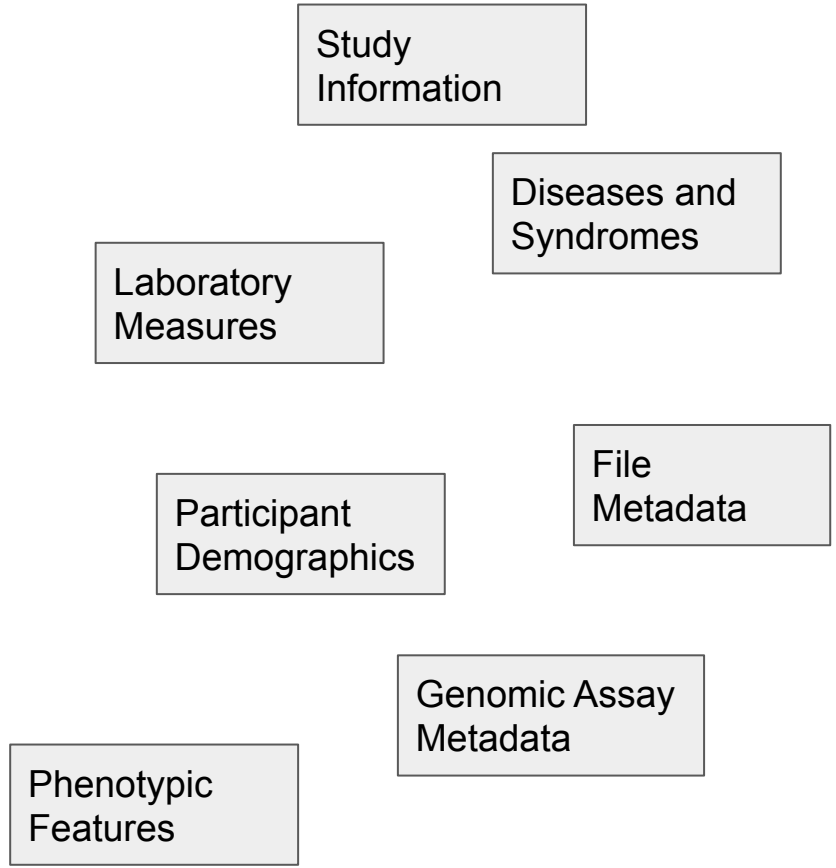
Implementation Guide v0.1

- Github: https://github.com/NIH-NCPI/ncpi-fhir-ig
- Pages: https://nih-ncpi.github.io/ncpi-fhir-ig/
- Originally published in 2021, focused on rare disease modeling for genomic research
- Live deployments have generated valuable feedback
  - Broader use cases
  - Refining approach to asserting semantics



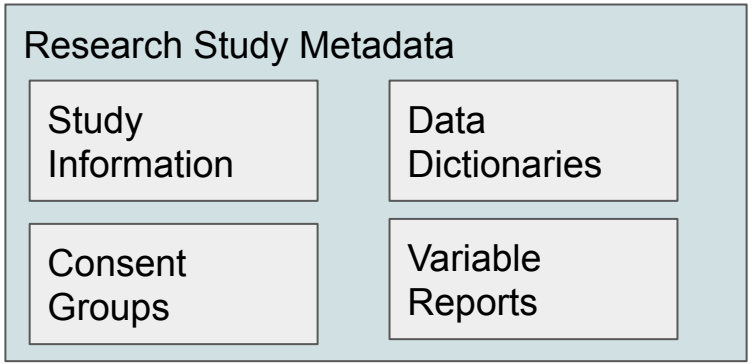
Interoperable Data Services

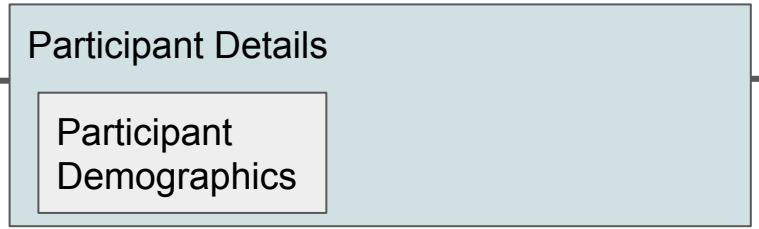
- The vision for FHIR across NCPI is to provide a set of services for the data and metadata to empower researchers.
- Not all services apply to all datasets nor platforms, but many are common!

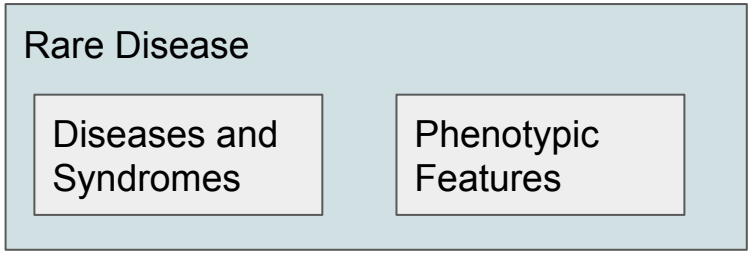


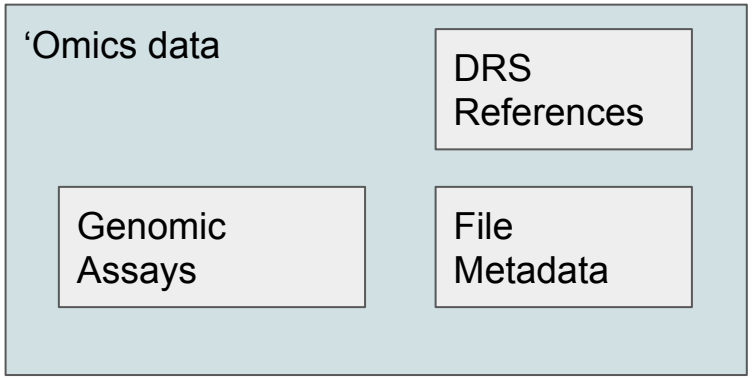
Interoperable Data Services

- We are re-organizing into a set of modules or services that help make clear what is being provided.
- This slide has a rough sense of some use cases.









IG v0.2

- This reorganization will make the underlying objective of the IG more clear
- Additionally, documentation will be more accessible to implementers and users of the NCPI FHIR services
- Use cases will be better integrated as well, with guides to users to help understand what services may be offered and how that may impact their analyses.

FHIR Code-a-thon

- Last summer, support from the ODSS enabled us to host a general purpose FHIR training for the NCPI community.
- Next week, 27 and 28 June 2022, we are hosting another event!
- We will implement an end-to-end analysis using a suite of NCPI-supported standards and tools, including FHIR and DRS.
- We will analyze RNASeq-derived Gene Expression data, with the primary target of clustering samples by gene expression.
- We hope to show the power of the work many of you have contributed!

FHIR Code-a-thon

- Event Overview: NCPI FHIR Code-a-thon 27-28 June 2022
- Registration Link
- Github Repository for managing shared code
- Github Project for tracking event status

- There are opportunities to contribute across technical, scientific, and documentation domains; please drop in if you are able.
- If you can't make it this week, the code and access information may help you get started in the future!

NCPI Outreach WG

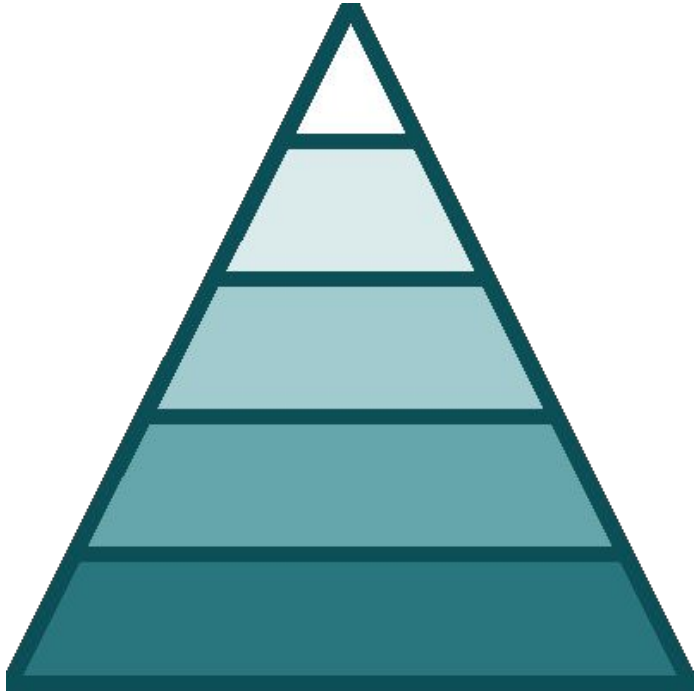
Stephen Mosher (Johns Hopkins University)

NCPI Outreach WG Mission

To prevent the development of siloed platforms by providing unified access to key information and training resources associated with each NCPI platform.

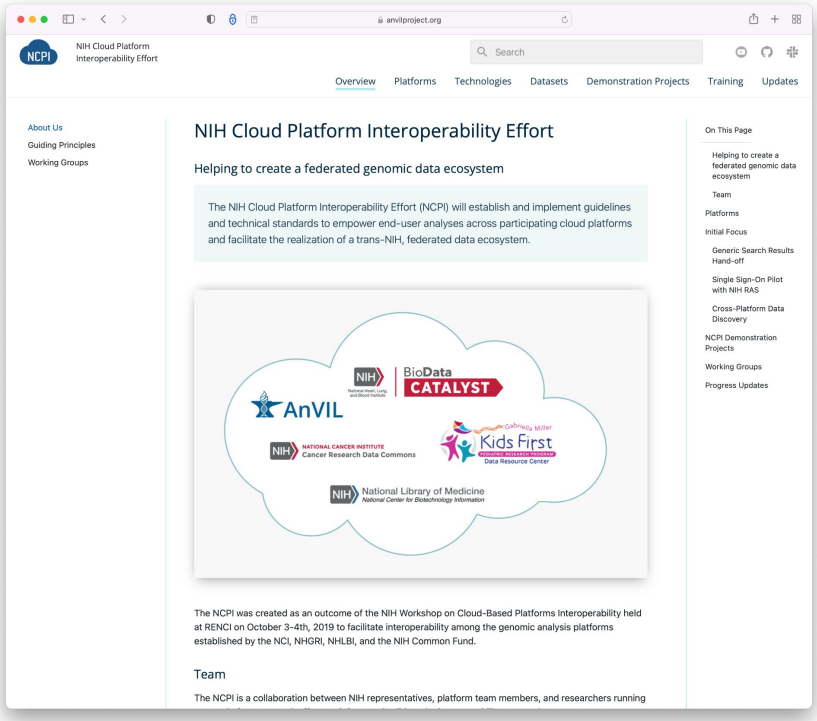
Goals

- Develop and maintain NCPI Portal
- Aggregation of platform-related outreach and training materials
- Document commonly used resources
- Maintain a catalogue of NCPI datasets
- Support NCPI Workshops

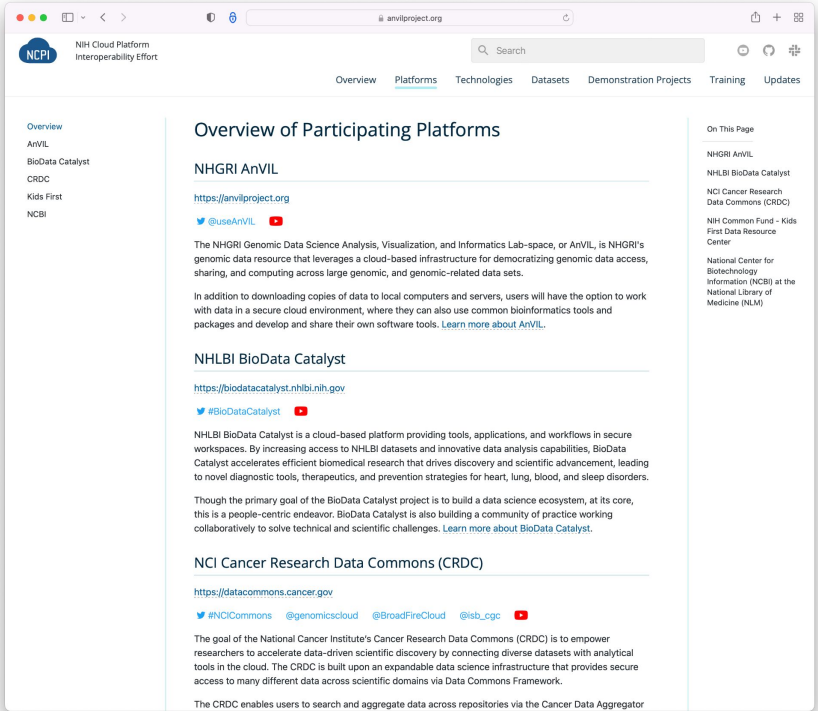


NCPI Portal

https://anvilproject.org/ncpi

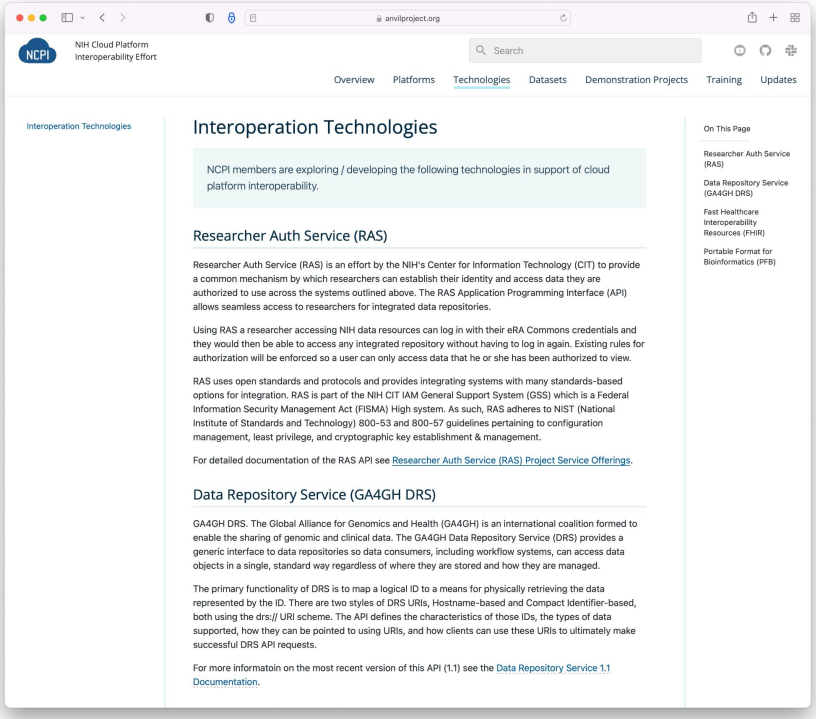


Participating platforms

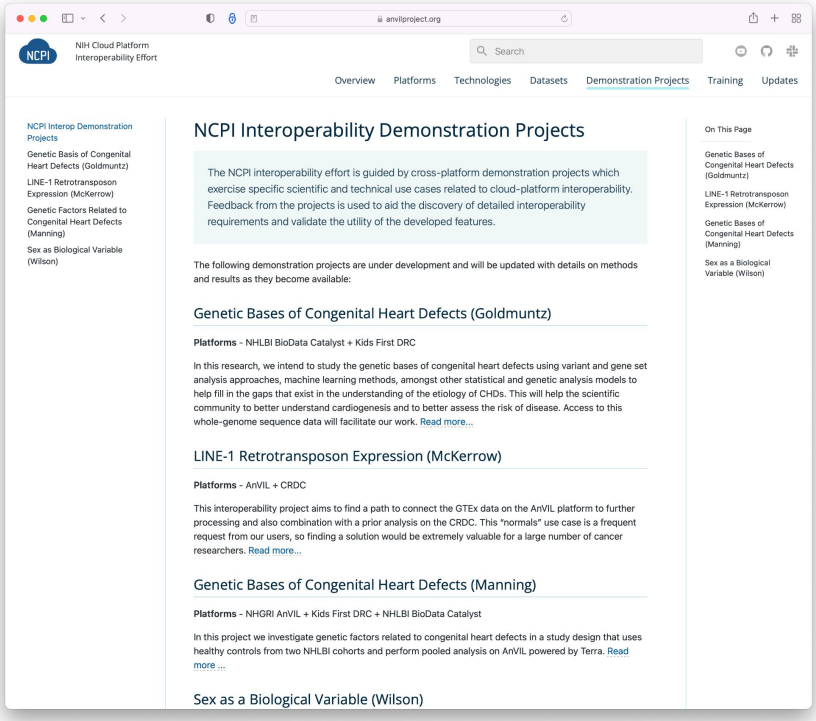


NCPI Portal

Technologies enabling science

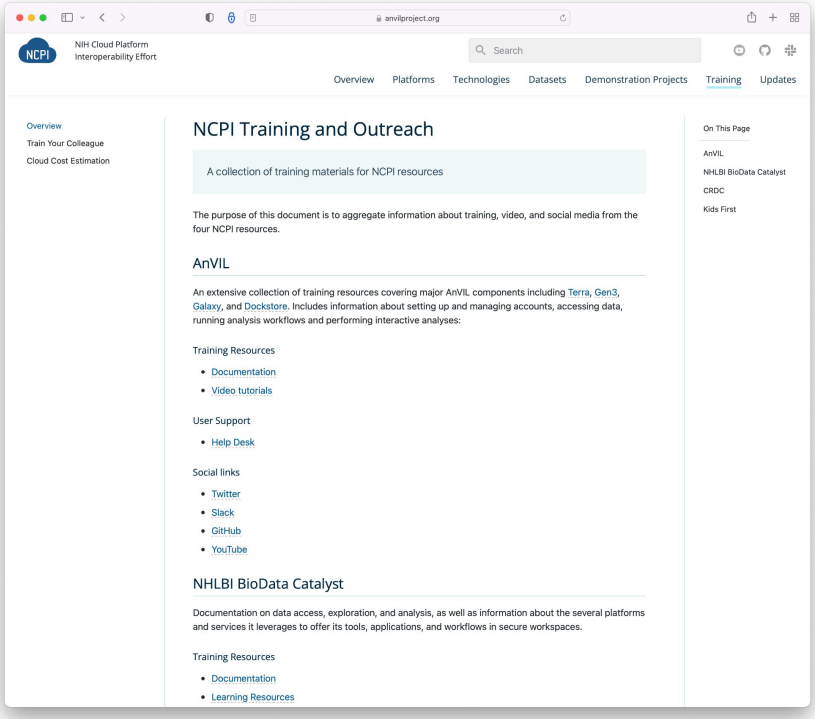


The science driving the tech

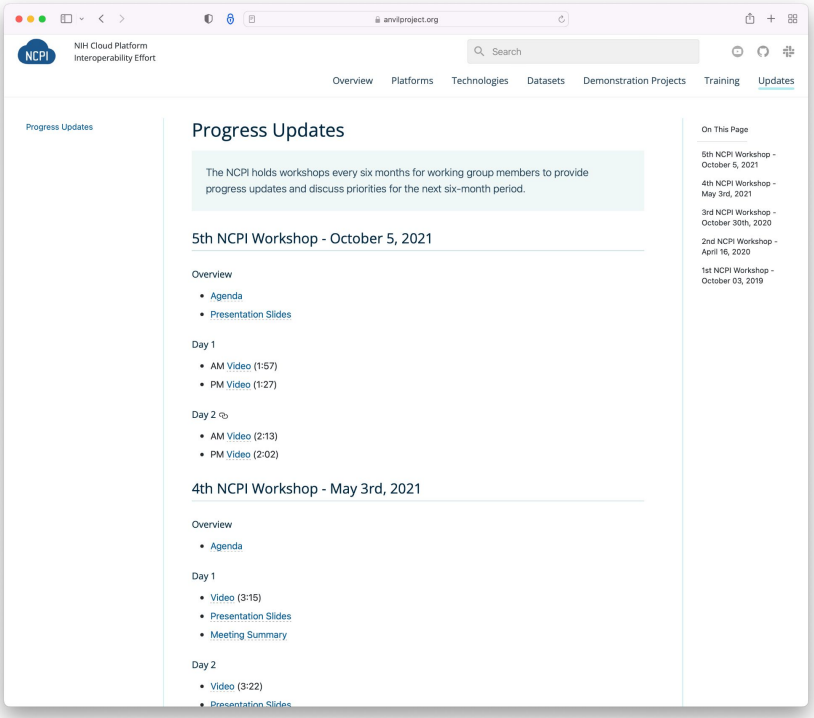


NCPI Portal

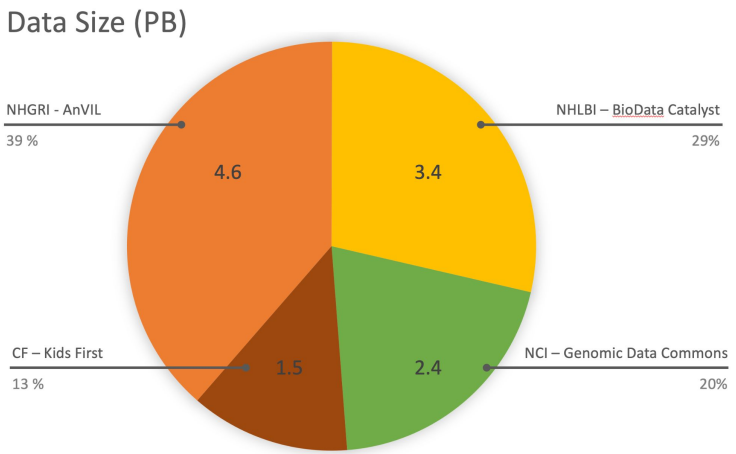
Aggregating outreach resources



Past workshop resources



NCPI Dataset Catalog



Researcher Auth Service

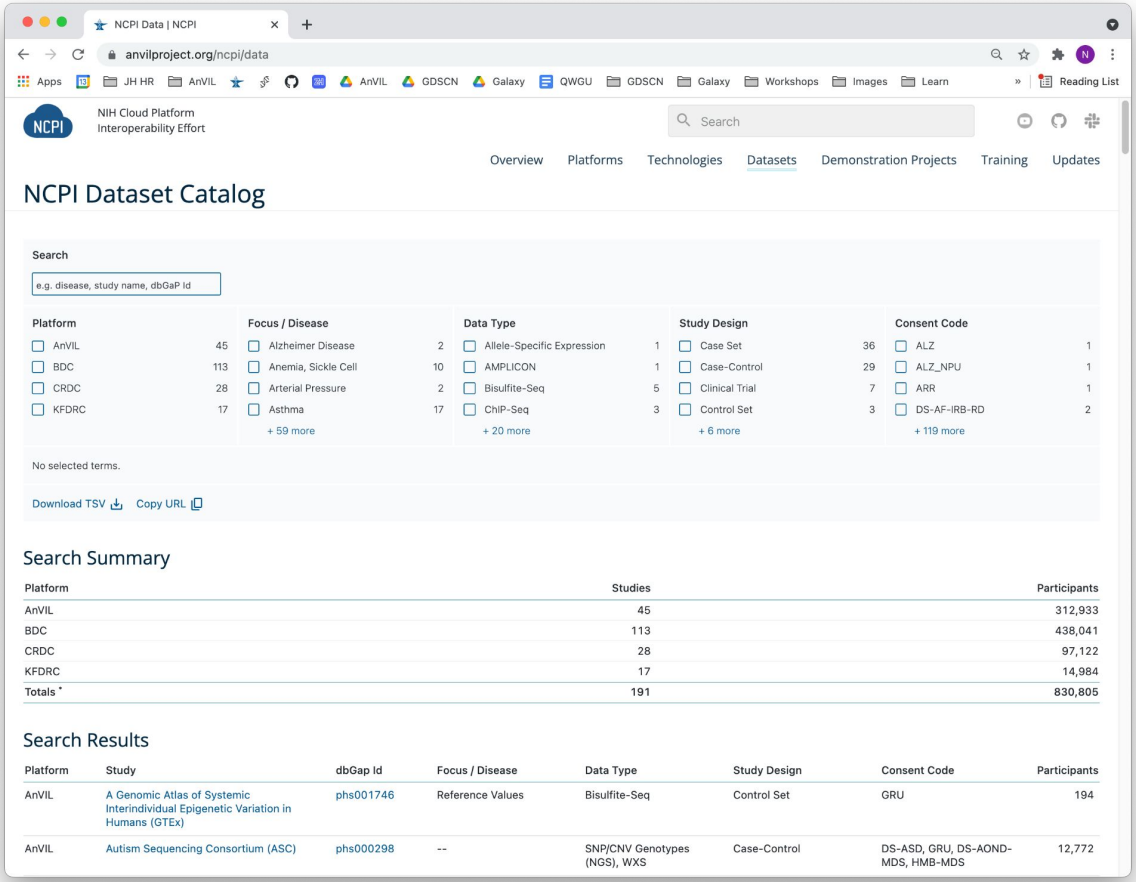
Data Repository Service

Fast Healthcare Interoperability Resources

12Pb / 830k participants and growing!

Cross-platform accessibility through several key technologies

Dataset Search (more details from Search WG)



Search by:

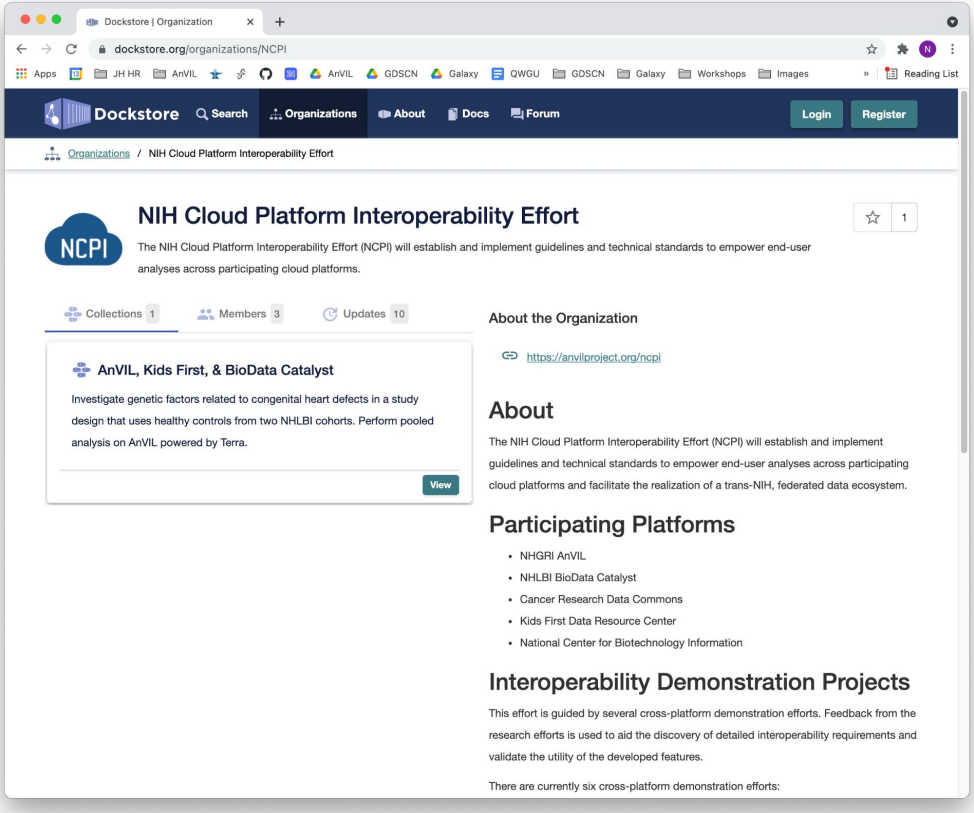
- Platform
- Focus or Disease
- Data type
- Study Design
- Consent Code

Budded off into new Search Working Group

Dockstore Organization for NCPI

Promoting FAIR practices in tool and workflow sharing

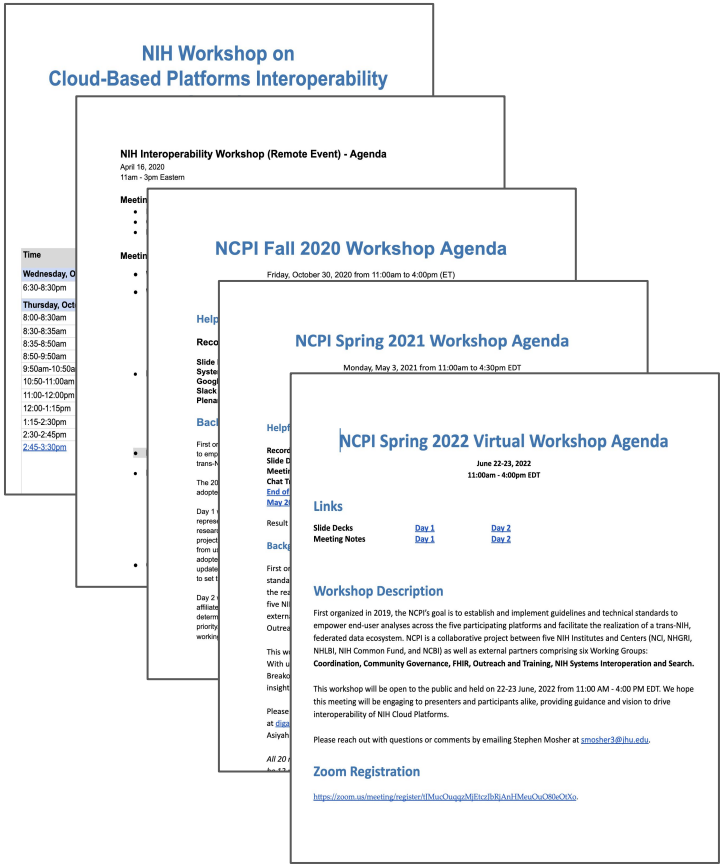
- Findable
- Accessible
- Interoperable
- Reusable



https://dockstore.org/organizations/NCPI

Supporting NCPI Workshops

Workshop	Date	Host
1st NCPI Workshop	03-04 October, 2019	BioData Catalyst
2nd NCPI Workshop	16 April, 2020	AnVIL
3rd NCPI Workshop	30 October, 2020	Kids First
4th NCPI Workshop	3-4 May, 2021	BioData Catalyst
5th NCPI Workshop	5-6 October, 2021	NCI CCDH
6th NCPI Workshop	22-23 June, 2022	AnVIL

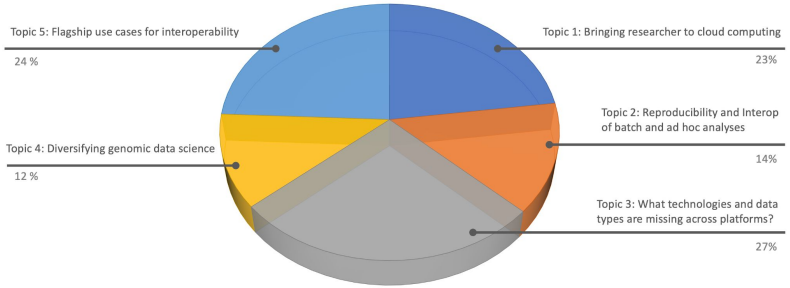


Today's Virtual Workshop

- Dedication from the Outreach WG, wider NCPI community and our partners to make today's event possible
- Planning across two days, four sessions of speakers, two breakout sessions, one panel discussion
  - 19 Speakers, 15 Breakout Moderators, 8 Note Takers, 3 Panelists, two MCs
  - 175 Registered Participants

	Session	Candidate 1	Candidate 2	Note taker
DAY1 2-4pm EDT 22JUN2022	Parallel Session 1	Allison Heath	Brian O'Connor	Beth Sheets
	Parallel Session 2	Valentina Di Francesco	Mike Feolo	Natalie Kucher
	Parallel Session 3	Chris Wellington	Stan Ahalt	David Higgins
	Parallel Session 4	Kathy Reinold	Adam Resnick	Marcia Fournier
	Parallel Session 5	Michael Schatz	Rachel Liao	Stephen Mosher
Day2 2:35-3:50pm EDT 23JUN2022	*Topic 1: Bringing researchers to cloud computing*	Tiffany Miller	NA	Helen Thompson
	Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses	Jack DiGiovanna	NA	Natalie Kucher
	Topic 3: What technologies and data types are missing across platforms?	Ken Wiley	NA	David Higgins
	Topic 4: Diversifying genomic data science	Asiyah Lin	NA	Marcia Fournier
	Topic 5: Flagship use cases for interoperability	Michael Schatz	NA	Cara Mason

Breakout 2



Future: Administrative Coordinating Center (ACC)

DEPARTMENT OF HEALTH & HUMAN SERVICES — Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

www.nih.gov

March 16, 2022

Research Opportunity Announcement

Research Opportunity Title: NIH Cloud Platform Interoperability Administrative Coordinating Center

OTA-22-004

Participating Organization(s): National Institutes of Health

Components: This Other Transactions Research Opportunity Announcement (OT ROA) is to support the *NIH Cloud Platform Interoperability* program (NCPI) and complements investments by NIH Institutes, Centers, and Offices (ICOs) in secure cloud-based platforms for data storage, sharing, and analysis. This research opportunity will be administered by the Office of Data Science Strategy (ODSS).

Funding Instrument: The funding instrument is the Other Transaction (OT) Award mechanism.

OT awards are not grants, cooperative agreements, or contracts, and use an Other Transactions Authority provided by law. Terms and conditions may vary between awards. Each award is therefore

Search WG

Dave Rogers (Clever Canary)
Kathy Reinold (Broad Institute)

Overview

- Mission, Vision, Strategy
- Search Use Cases

- ODSS Search RFI Response
- Search Landscape Survey of the NCPI search ecosystem
- Search Demonstration Projects

- Next Steps
- Discussion

Mission

The NCPI Search Working Group, formed in October 2021, aims to:

- Accelerate the **improvement of search interoperability** across the participating NCPI platforms in support of NCPI's shared vision of a trans-NIH, federated data ecosystem.

- Focus on supporting federated dataset discovery, cohort creation, and knowledge discovery.

See the NCPI Search Group Charter

Vision

- We envision an integrated, federated, FAIR data ecosystem, supporting
  - data interoperability,
  - transparency of data provenance and quality,
  - researcher and participant equity.

- The Search Working Group advances this vision by identifying, evaluating, promoting, and demonstrating the effective use of data interoperability standards and guidelines.

Target Search Use Cases / Modalities

Support search of studies and datasets across platforms by:

- experimental metadata such as assay, datatype, or study design,
- participant metadata such as medical history/treatment, behavioral metadata, environmental exposure, social determinants of health,
- observations made such as variants identified or the existence of other biomarkers,
- participant-consented allowable use.

Strategy

- Be driven by researcher scientific use-cases.

- Advocate for a federated search architecture.
- Advocate for common standards for data models and APIs.

- Foster knowledge sharing across the NCPI search community.
- Solicit and facilitate NCPI Search Demonstration Projects to provide concrete examples of standards and guidelines in action.
- Promote the best open access view of managed access datasets

ODSS Search RFI Response Overview

The NCPI Search Working Group's response to the NIH/ODSS Search RFI advocates:

- an open and federated data ecosystem,
- data standards adoption,
- exploring FHIR as an API solution for representing research data at the study metadata and individual level,
- investing in tools that enable the entire data collection, curation, submission and data sharing process to be infused with structured metadata/common data elements (CDEs).

See NOT-OD-21-187 Request for Information (RFI): Search Capabilities across the Biomedical Landscape for NIH-wide Data Discovery

RFI Response Overview

Specific recommendations included:

- Establishing a “Minimum Study Metadata” standard to drive consistent discovery of program data.
- Advocating for data catalog and data explorer code reusability and multi-tenancy to help accelerate implementation timelines and drive consistency across programs.
- Aligning on standard ways to “push” cohorts from data repositories to analysis environments, and “pull” selected clinical and genomic variables of interest from data repositories to analysis environments.
- Aligning on a mechanism to support pan-NIH dataset search.

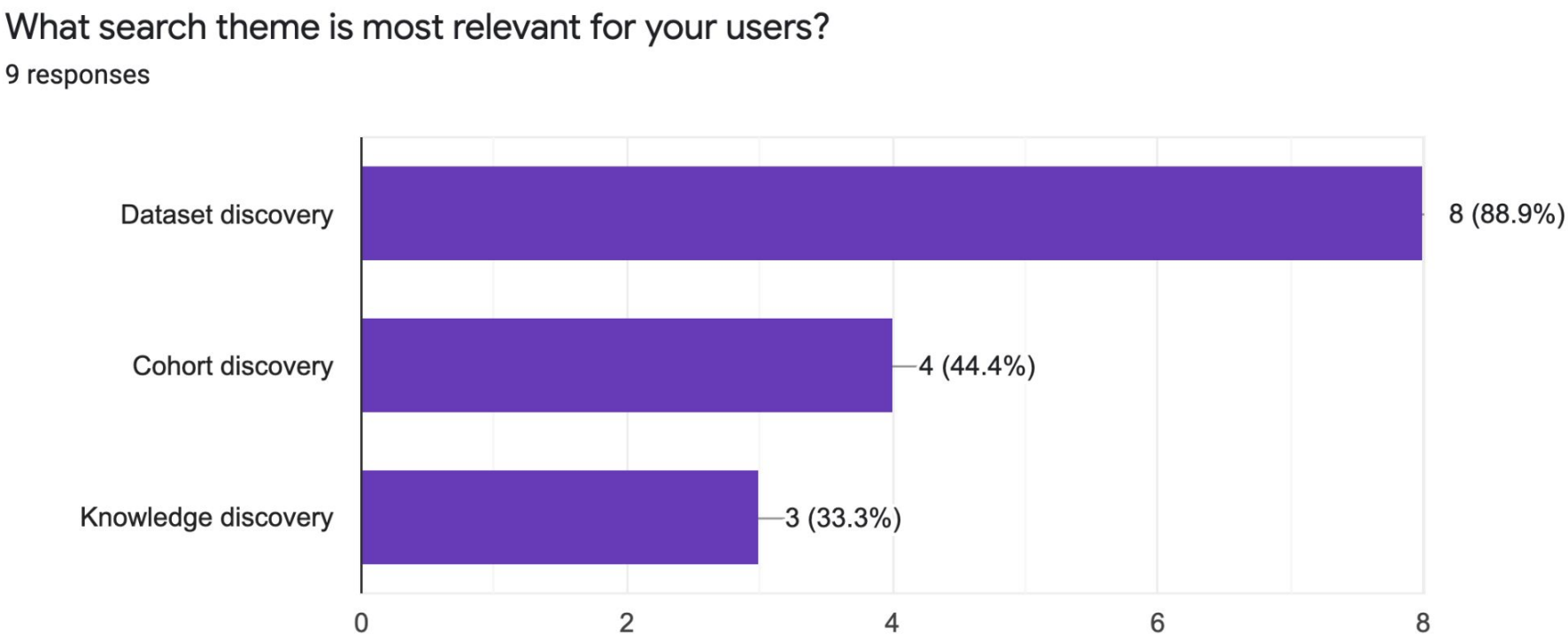
See the NCPI Search RFI Response.

Landscape Survey

- Purpose
  - Provide an overview of current search capabilities across NCPI platforms
  - Describe how we currently address search needs and understand the challenges

- Search capabilities represented in responses
  - AnVIL Gen3 Explorer, AnVIL Dataset Catalog
  - BioData Catalyst PIC-SURE, Dug
  - CRDC Cancer Data Aggregator (CDA) Search API
  - Kids First Data Portal, FHIR API
  - NCBI dbGaP Advanced Search, dbGaP FHIR API
  - NCPI Dataset Catalog

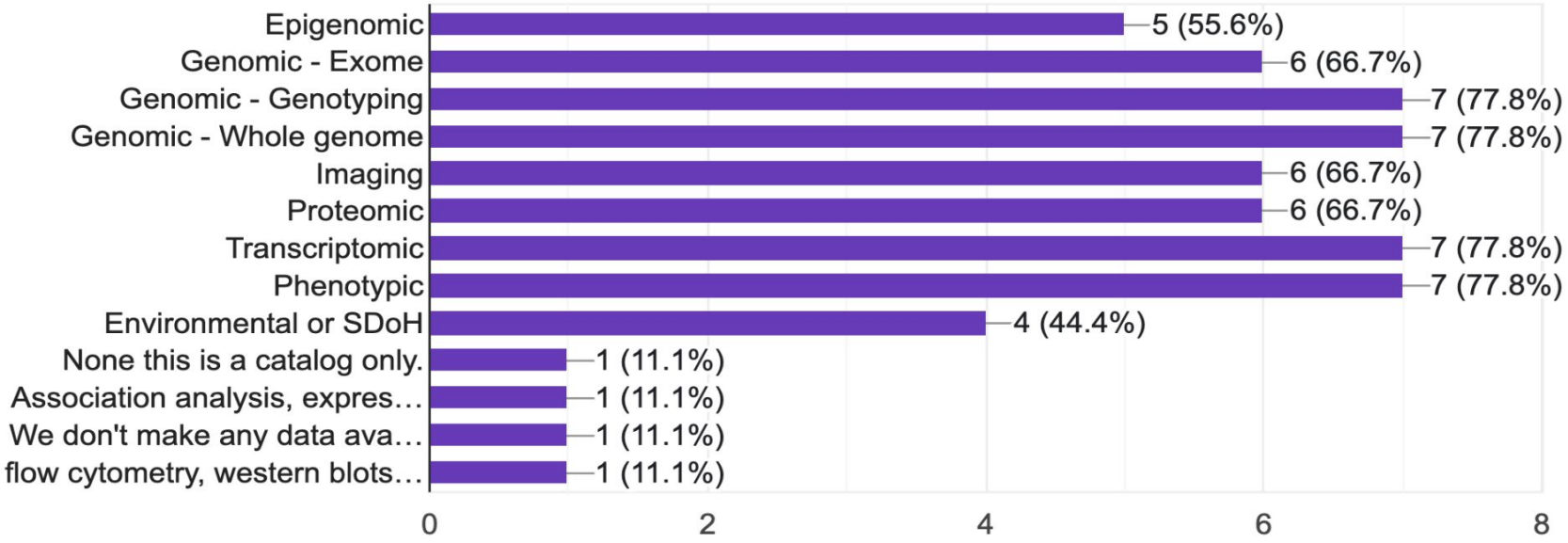
Landscape Survey - Theme



Landscape Survey - Data Modalities

What data modalities or types do you make available to users (assuming use has appropriate access rights)? Check all that apply.

9 responses



Landscape Survey - Phenotype Standards

- Summary
  - Most reference ontologies
  - Clearly some variation

Standard	**Responses**
HPO	3
MESH	1
PhenX	1
Follow dbGaP guide	1
Annotated w/ ontology ids	1
SNOMED	1
LOINC	1
NCIT	1
OMOP	1
CRDC Data Dictionaries/CRDC-H	1

Landscape Survey - Standards

Genotype Standards

Standard	Responses
Ensemble	1
Follow dbGaP guide	1
NCIT	1
MIAME	1
CRDC Data Dictionaries/CRDC-H	1
Whatever platform provides	1
n/a or no response	3

Other Data Standards

Standard	Responses
MIAME	1
Follow dbGaP guide	1
SRA	1
DUO	1
CRDC Data Dictionaries/CRDC-H	1
n/a or no response	5

Landscape Survey - Standards

- Non-phenotype data
  - Three responses reported this is not applicable
  - Of the other, generally one of the respondents reported the following
    - PubChem, EDAM, UBERON, OBI, dbGaP Submission Guide, SNOMED, LOINC, DICOM, OMOP, MONDO, ICD10, NCIT
  - Observation: Consider recommending specific ontologies for types of data
    - I.e. disease, lab tests, anatomy…
- Social Determinants of Health (SDoH)
  - One group reported storing this data in SQL database, another referenced dbGaP Submission criteria, others reported either not applicable or TBD
  - What standards cover this category well?

Landscape Survey - Key Points

- Key technology enablers of cross-platform search & cohort building
  - Internet, common terminology, open APIs, interoperable data models, elastic search, FHIR API, subject-level and file metadata
- Key metadata for search
  - Subject/Patient - demographic, phenotypic, whole organism tests, exposures
    - Does this include model organism or cell lines?
  - Samples/Biospecimen - diagnosis (disease, treatments), assays/analysis performed
  - Subject, sample counts and of course provenance - who, when, how…
  - Files - data modality/type of analysis/experimental strategy/data type, data format
- Consent
  - Four groups search open data only, others reference dbGaP consent groups, DUO consent codes, RAS
- Security
  - One reference to RAS, 5 responses cite FISMA-moderate and FedRAMP certifications.

Landscape Survey - Challenges

- Lack of metadata standards, lack of minimal standard
- Quality of metadata
- Lack of standardized APIs, APIs to pull data for indexing
- Different groups bringing their own data dictionaries
- Heterogeneity of data formats
- Lack of collaboration
- Better focus on the science
- Observation - changing nature of data, data formats – how to manage that?

Landscape Survey - Next Steps

- Continue to refine the survey with respect to data models and indexing methods.

- Publish the survey results on the NCPI Portal.

Demonstration Projects

Several demonstration projects for specific use cases are in the proposal phase including:

- Uniform search of public sample and sequence read information across NCBI and Kids First repositories. - Anne Deslattes Mays
- PIC-SURE NCPI Platform Integration - Paul Avillach
- Filter studies by DUO codes on the NCPI Dataset Catalog - Dave Rogers, Jonathan Lawson

See the NCPI Use case Tracker

Next Steps

- Recruit additional members.
- Solicit / recruit additional demonstration projects.
- Publish the landscape survey and additional analysis to the NCPI portal.
- Provide a survey of data model descriptions.
  - What are common tools used to describe data models?
  - Include those that allow for mapping/translation between data models or support schemas.
- Propose initial data model standards for discoverability.
  - Work closely with FHIR and Interop WGs
- Evolve strategy and refine near and longer term goals.

Questions/Discussion?

Break

1:05 PM - 1:35 PM EDT

Technical Aspects of Interoperability

1:35 PM - 2:35 PM EDT

The Texas Advanced Computing Center (TACC) as an Interoperable Cloud Resource for Biomedical Research

Dan Stanzione (TACC)

THE TEXAS ADVANCED COMPUTING CENTER (TACC) AS AN INTEROPERABLE CLOUD RESOURCE FOR BIOMEDICAL RESEARCH

Dan Stanzione
Executive Director, TACC
Associate Vice President for Research, UT-Austin

Cloud Platform Interoperability Workshop

June 2022

TACC - 2021

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

THE CHARGE FOR THIS TALK:

- How can TACC be leveraged for Biomedical Sciences?
- What resources are currently available?
- What technologies you are using to ensure interoperability with other systems?
- and some successful research examples for both basic and clinical research. . .

- (not necessarily in that order).

TACC AT A GLANCE - 2021

Personnel

185 Staff (~90 PhD)

Facilities

12 MW Data center capacity
Two office buildings, Three Datacenters, two visualization facilities, and a chilling plant.

Systems and Services

15 production platforms, the #1 and #3 US academic supercomputers

>Nine Billion compute hours per year
>5 Billion files, >100 Petabytes of Data,

Usage

>15,000 direct users in >4,000 projects, >50,000 web/portal users, User demand 4x available system time. Thousands of training/outreach participants annually

WHAT WE DO

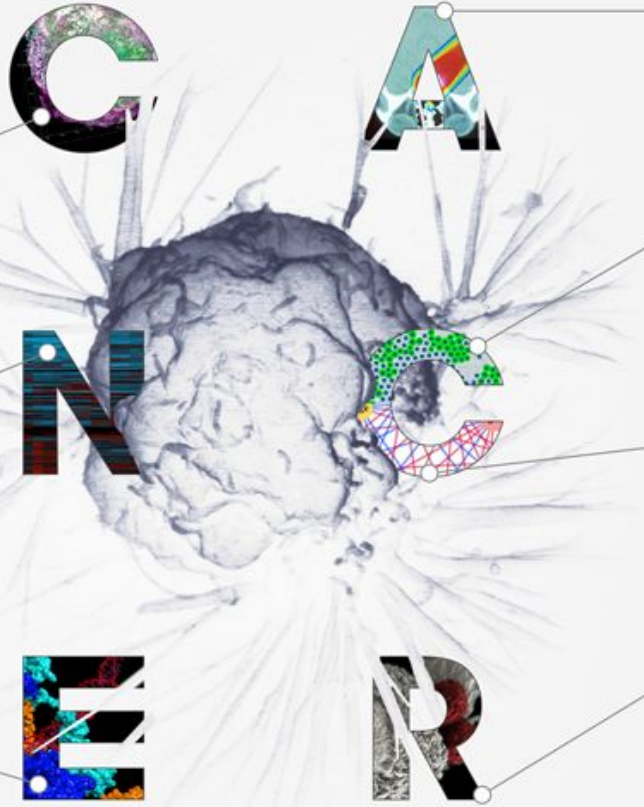
- Provide researchers with:
  - Computing, Data, AI , Software capabilities to support their research
  - The expert help to be able to use it!
  - In the ways they want to consume it
  - Help with grants/strategy
- Computation, AI, Data almost ubiquitous acro the sciences.

SYSTEMS UPDATES
A QUICK REMINDER ON OUR CURRENT MAJOR SYSTEMS

- Frontera, NSF Capability System, 2019-2025 (Currently #16)
- Stampede2, NSF Capacity System, 2017-2023 (Currently #47)
- Lonestar-6, Texas/Local System 2022-2027
- Longhorn – AI/DL GPU System, 2019-2025
- Jetstream2 - NSF "Cloud" System 2022-2027
- Chameleon – NSF CS Testbed 2015-2024 (multiple HW upgrades)
- Corral, Ranch, Stockyard – Storage Platforms
- *Aggregate: ~75PF, ~16,000 compute nodes, ~350PB*

The Texas Advanced Computing Center accelerates basic and applied cancer research to help save lives.

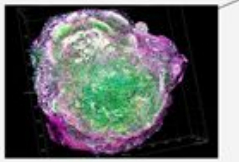
fighting



with supercomputers

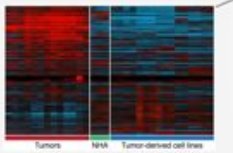
Computer Modeling

Researchers use advanced computing to model tissues, cells and drug interactions, and to design patient-specific treatments and identify new medicines.



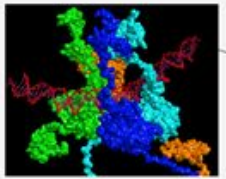
Big Data Analysis

Supercomputers allow researchers to find patterns in genomes and among patient outcomes to pinpoint risks and target treatments.



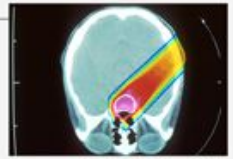
Molecular Dynamics Simulations

Simulating protein and drug interactions at the atomic level enables scientists to understand cancer and design more effective therapies.



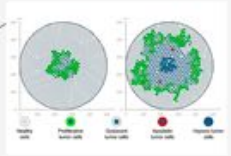
Quantum Calculations

Exploring how proton and x-ray beams interact with DNA on the quantum level helps explain why radiation treatments work and how they can be optimized.



Trial Design

Researchers use TACC's advanced computers to design clinical trials that can determine the combination of dosages that will be most effective.

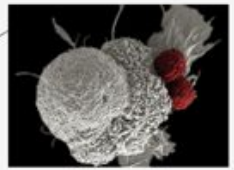


Clinical Planning

Supercomputers can test thousands of potential treatments in advance to help decide which one will work best.

Artificial Intelligence

AI on high-performance computers can uncover relationships among complex cellular networks and reverse-engineer interventions.

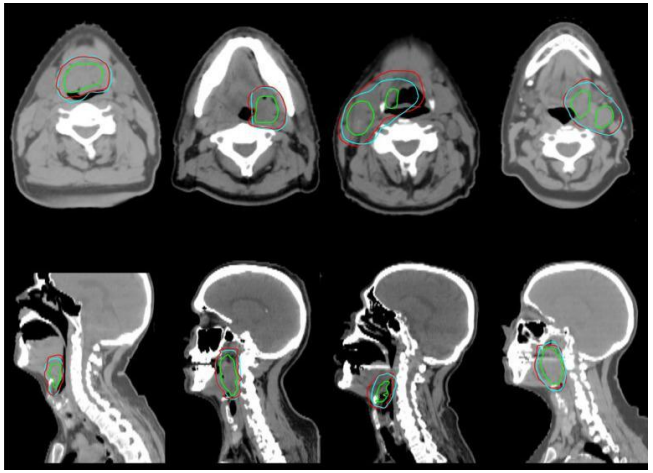


TACC
TEXAS ADVANCED COMPUTING CENTER

Artificial intelligence and deep neural networks increased speed and efficiency for identification of head and neck cancers

- **Problem:** Contouring is the process by which radiation oncologists carefully review medical images of the patient to identify the gross tumor volume, then design patient-specific clinical target volumes that include surrounding tissues, since these regions can hide cancerous cells and provide pathways for metastasis. The process is quite subjective, and there is wide variability in how trained physicians contour the same patient's computed tomography (CT) scan.

- **Importance:** In the case of head and neck cancer, countouring is a particularly sensitive task due to the presence of vulnerable tissues in the vicinity. Better contouring can lead to determining best practices, so standards of care can emerge.



Comparison between computer-predicted ground-truth clinical target volume (CTV1) (blue) and physician manual contours (red)

- **Approach:** Carlos Cardenas (MD Anderson) used Maverick to analyze data from 52 oropharyngeal cancer patients who had been treated at MD Anderson between January 2006 to August 2010, and had previously had their gross tumor volumes and clinical tumor volumes contoured for their radiation therapy treatment. He developed deep learning algorithm using auto-encoders — a form of neural networks that can learn how to represent datasets — to identify and recreate physician contouring patterns.

- **Result:** Cardenas and his collaborators tested the method on a subset of cases that had been left out of the training data. They found that their results were comparable to the work of trained oncologists. The predicted contours agreed closely with the ground-truth and could be implemented clinically, with only minor or no changes.

TACC

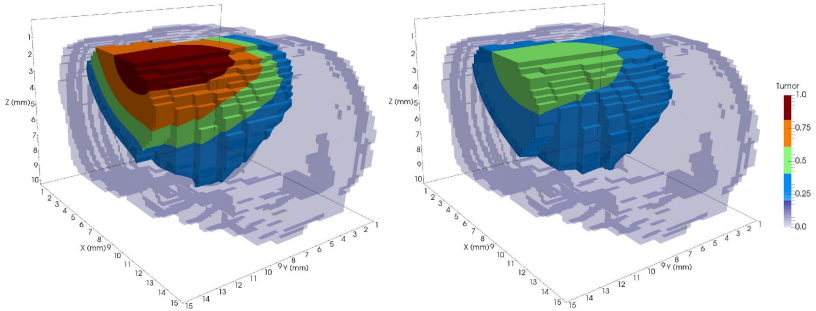
https://www.tacc.utexas.edu/-/an-ai-oncologist-to-help-cancer-patients-worldwide

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

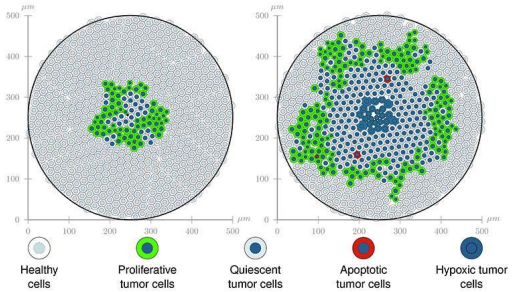
Complex **computer models** and **analytic tools** to predict how cancer will progress in a specific individual

- **Problem:** The current state of cancer research is data-rich, but lacking governing laws and models. The solution may not be to mine large quantities of patient data, but to *mathemetize* cancer: to uncover the fundamental formulas that represent how cancer behaves.

- **Importance:** Accurate models could be used to predict the growth and decline of cancer and reactions to various therapies.



Model of tumor growth in a rat brain before radiation treatment (left) and after one session of radiotherapy (right)



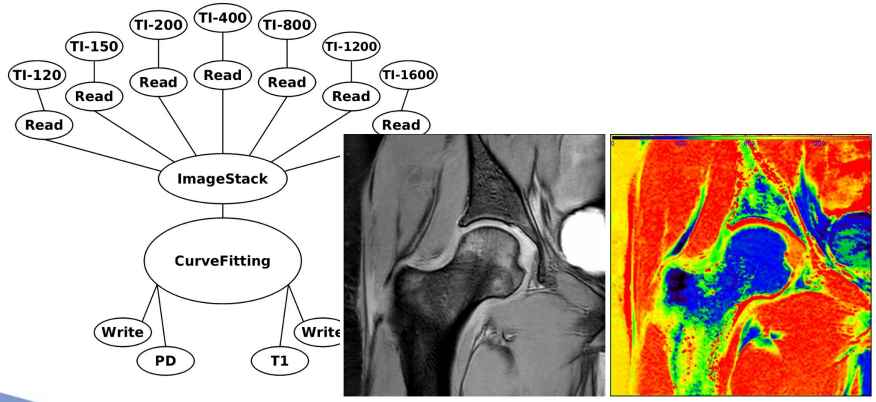
Snapshots of a tumor model with tumor cells growing in a healthy tissue at two time points and under different nutrient conditions

- **Approach:** Researchers from Dell Medical School used Stampede2 to analyze patient-specific data from magnetic resonance imaging, positron emission tomography, x-ray computed tomography, biopsies and other factors, in order to develop their computational model.

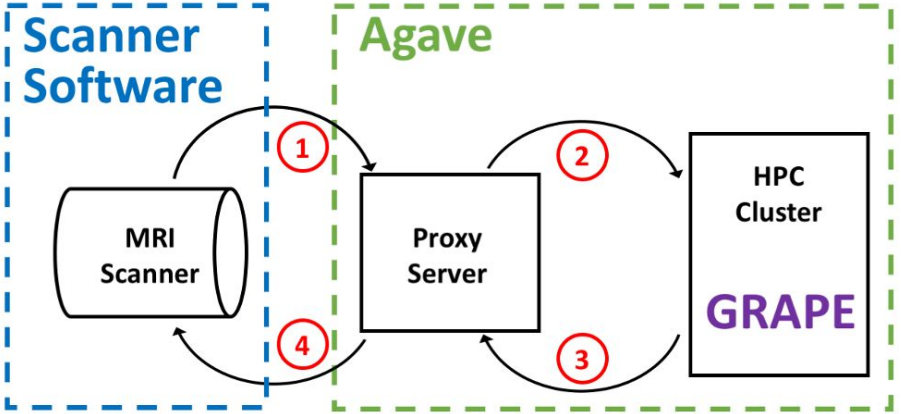
- **Result:** The group was able to predict with 87 percent accuracy whether a breast cancer patient would respond positively to treatment after just one cycle of therapy.

TAPIS and Jetstream enabled automated, real-time, quantitative magnetic resonance imaging

- **Problem:** Quantitative analysis of MR images is typically performed after the patient has left the scanner. Corrupted or poor quality images can result in patient call backs, delaying disease intervention.
- **Importance:** Real-time analytics of MRI scans can enable same-session quality control, reducing patient call backs, and it can enable precision medicine.



Quantitative calculations performed during scan session



Platform to automate analysis tied to HPC resources

- **Approach:** Dr. Refaat Gabr (UTHealth) and Dr. Joe Allen (TACC) used the CyVerse SDK and Agave to help develop an automated platform for real-time MRI,
- **Result:** Scan data can now be automatically processed on high performance computing resources in real-time with no human intervention.

https://www.tacc.utexas.edu/-/real-time-mri-analysis-powered-by-supercomputers

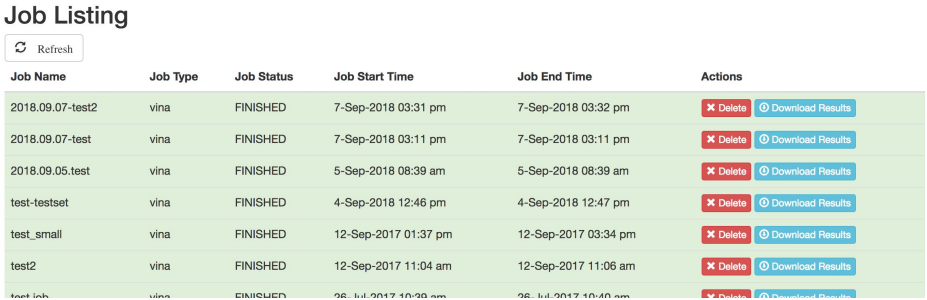
TACC

UTHealth

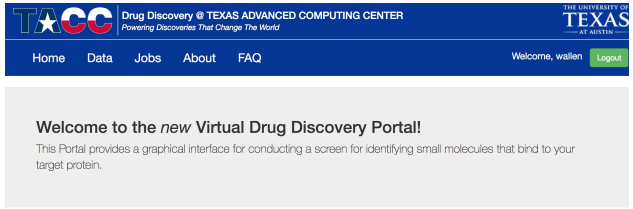
The Drug Discovery Portal empowers researchers worldwide to perform virtual screens on TACC HPC resources

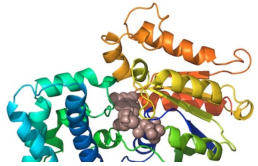
- **Problem:** While *virtual screening* has compelling advantages over experimental methods alone, it requires high-performance computational resources, software licenses, and technical expertise, which may be unattainable for small academic labs.

- **Importance:** Successful structure-based virtual screening methods save time and resources in the drug discovery pipeline.



Job outputs are available for download in a web interface





The DrugDiscovery@TACC web portal

- **Approach:** Dr. Stan Watowich (UTMB Galveston) partnered with researchers at TACC to provide an accessible and free virtual screening service called DrugDiscovery@TACC to investigators across the state of Texas and around the world.

- **Result:** Users upload proteins of interest into a friendly web interface, choose a ZINC library to screen, and results are returned typically within 24 hours. The efforts have led to dozens of documented drug candidate hits.

https://drugdiscovery.tacc.utexas.edu/

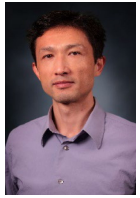
TACC

utmb Health

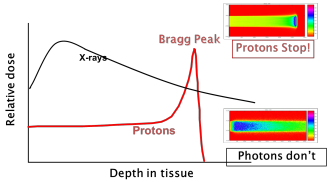
Particle/Proton Therapy Translational Research Platform

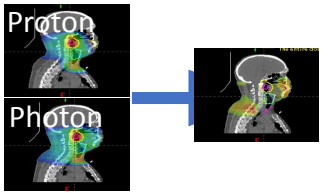
Xiaodong Zhang (MDACC)

Hang Liu (TACC)



- Radiation Therapy: shooting high-energy particles to kill tumors while sparing healthy tissues





Photon vs Proton

25 GY unnecessary photon radiation

- 25000 x of the general public annual radiation limit
- 5000000 x of the intraoral X-ray

- Intensity Modulated Proton Therapy (IMPT) is the most advanced radiation therapy
- IMPT plan is to search all available solutions for how each proton beam modulated to deliver prescribed radiation
- Ideal IMPT plan is impossible to be achieved in the current clinically available computing environment
- The huge advantages of IMPT have NOT been fully utilized for majority of cancer patients

TEXAS The University of Texas at Austin

The University of Texas at Austin Oden Institute for Computational Engineering and Sciences

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

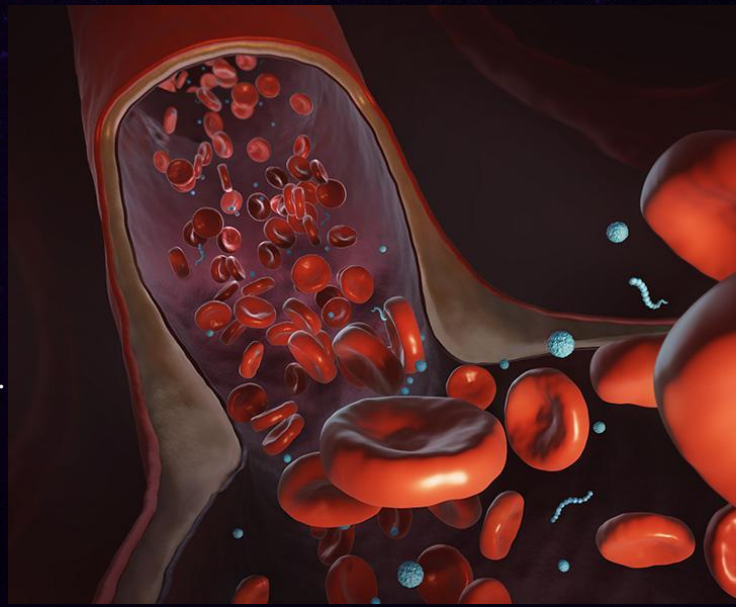
TACC TEXAS ADVANCED COMPUTING CENTER

- Multi-Center
- Protected Health Data Storage
- Protected Computing
- Virtual Biospecimen Data Repository
- Web browser accessible portal

TARGETING TUMORS WITH NANOWORMS

YING LI, UCONN

- "My research is centered on how to build high-fidelity, high-performance computing platforms to understand the complicated behaviors of these materials and the biological systems down to the nanoscale,"
- Nanoworms are long, thin, engineered encapsulations of drug contents.
- Modeled how these structures move in blood vessels of different geometries mimicking the constricted microvasculature.
  - Nanoworms can travel more efficiently through the bloodstream, passing through blockages where spherical or flat shapes get stuck.
  - Can use magnetic fields to influence flow.
- Can increase percentage of (highly toxic) drugs delivered directly to tumor.
- Published in *Soft Matter,* 2021.



TECHNOLOGIES THAT HELP MOVE THINGS AROUND

- Containerization:
  - We support Singularity, Charliecloud, Apptainer, a few others – the containerized workflows you build elsewhere will work at TACC
  - Push your Docker images into Biocontainers or other repositories, we can run them in Singularity.
  - At this point, that's just good software engineering
- Standard Orchestration tools:
  - We support Slurm (for batch), Kubernetes (Services, Interactive sessions), JupyterLab (notebooks)
- Our data storage and formats are, umm, not exotic.
  - POSIX Files in repository
  - Standard connectors for relational databases.
  - We do have object stores if you really like them (S3 interface, like AWS)… codes like them more than people.

TECHNOLOGIES THAT HELP MOVE THINGS AROUND

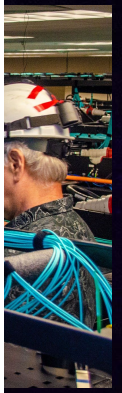
- Standard tools for interfacing, getting stuff in and out.
  - ssh/scp/gridftp for remote access
  - Google authenticator or others for multi-factor auth, where needed.
  - Open source TAPIS API for RESTFul web service access:
    - We've run this in AWS and Azure, as well as at TACC, and you could use it for free.
  - *There are no "TACC specific" access/workflow/API tools.*
  - *Maybe the cloud should run more like us. . .*
- We have computers, networks, storage systems, and a really good Linux image; you can run layers of your choice on that. . . What we recommend though:

TECHNOLOGIES THAT HELP MOVE THINGS AROUND

- Don't build on vendor-specific services. . . Almost all have open equivalents.
- Use containers that run anywhere, methods to fetch from central repositories.
- But even when portable, data migration has a cost – in money and time. And this adds up fast, so think about where your data is or should be.

- Plenty of our staff move back and forth ☺.

TH




FRONTERA
TACC
NSF
TEXAS

FHIR for Genomics: The Path Forward

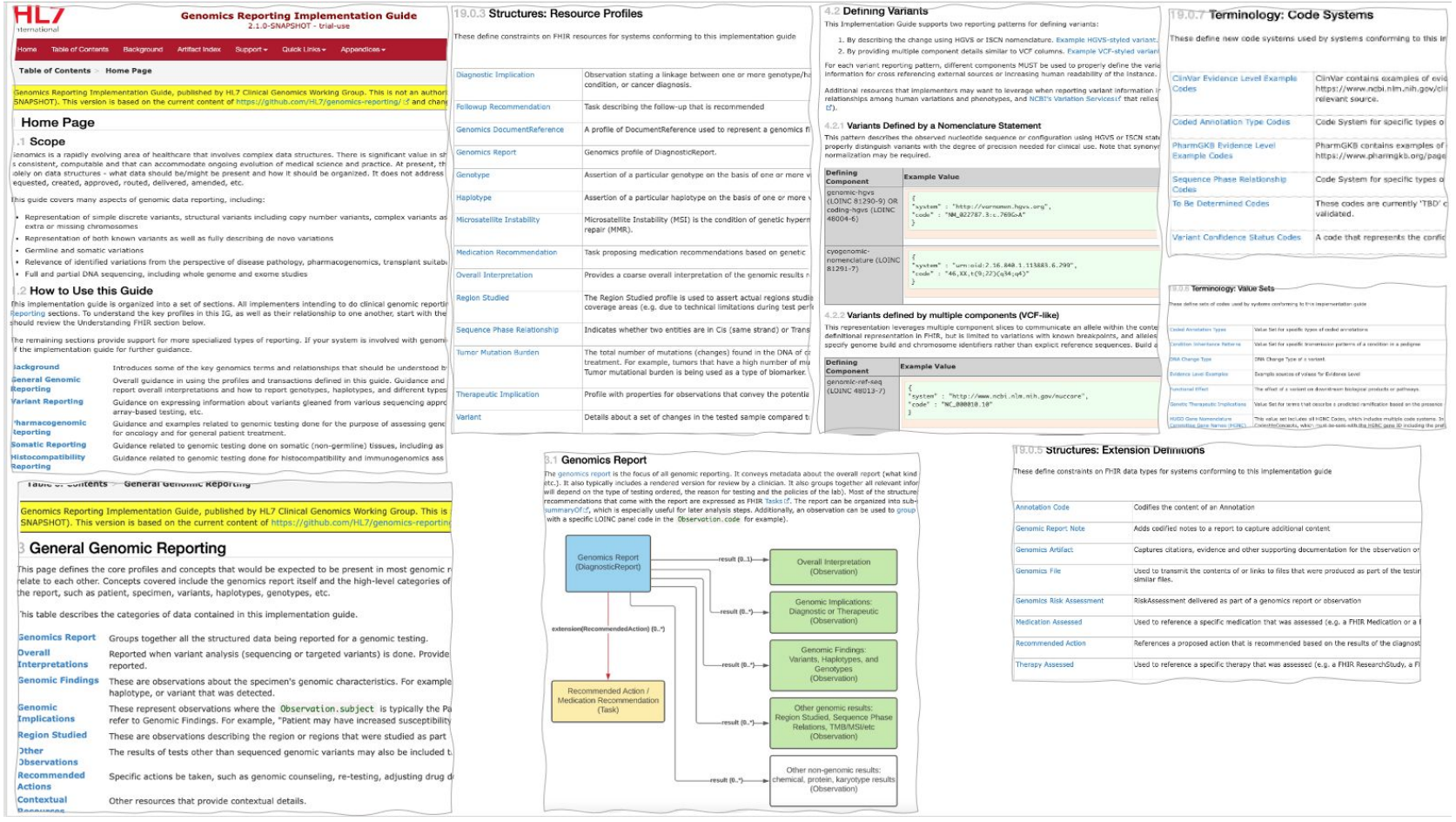
Mullai Murugan (Baylor College of Medicine)

Overview - HL7 FHIR for Genomics

FHIR & CG Overview

- HL7
  - **Healthcare Standards** for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services
- FHIR (Core Specification)
  - *FHIR® – Fast Healthcare Interoperability Resources – is a next generation standards framework created by HL7. FHIR combines the best features of HL7's v2 , HL7 v3 and CDA product lines while leveraging the latest web standards and applying a tight focus on implementability.*
  - RESTful API
  - Development heavily driven by implementations (see Argonaut)
  - Insufficient genomics representation in R4 (latest release)
- Clinical Genomics FHIR Implementation Guide (Specification)
  - Profiles of existing FHIR resources to support exchange of genomic data
  - Supports variant level data, variant level interpretations (inherited disease, somatic, PGx), report level interpretations, recommended follow-ups, report

Clinical Genomics Genomics Reporting IG



New Implementers

- Getting Started with Clinical Genomics for FHIR
- Clinical Genomics Working Group Participation
- Chat/Discussion boards
- Tracking and ticketing system
- Genomics Reporting STU2 Implementation Guide
- Genomics Reporting Working Draft Implementation Guide

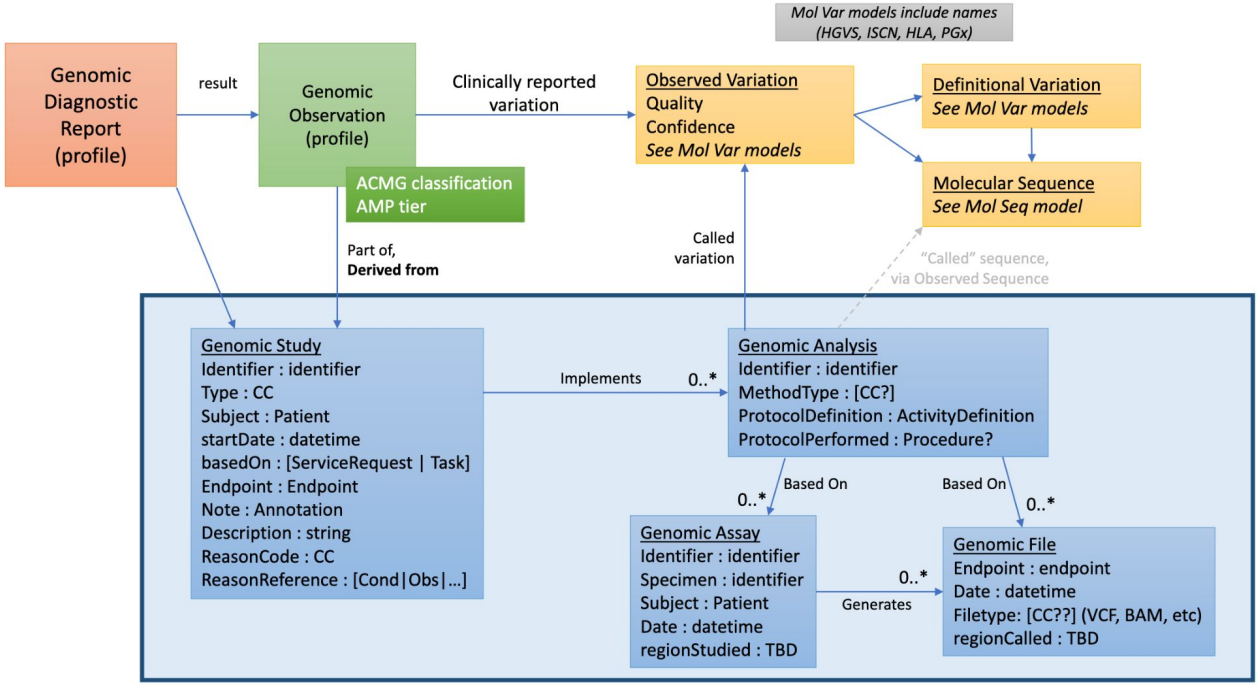
FHIR Genomics - New Initiatives & Ongoing Effort

Genomics FHIR Initiatives

- Genomics Reporting Implementation Guide - STU2 Publication
  - General Clinical Genomic Reporting
  - Information for expressing information about variants
  - Pharmacogenomic Reporting
  - Histocompatibility Reporting
- New - Genomic Study
- Other efforts
  - GenomeX, housed under the CodeX FHIR Accelerator
  - FHIR to OMOP

Genomic Study

Led by:
Robert Freimuth, Mayo Clinic
HL7 FHIR Clin Gen WG IM Lead



Use Cases:
- Reports with multiple components
-Multiple studies for same patient
- Consortia programs
-Trio, T/N testing etc.

Not shown: implication, interpretation, etc

http://build.fhir.org/ig/HL7/genomics-reporting/branches/operations/OperationDefinition-find-study-metadata.html

Challenges, and the path forward

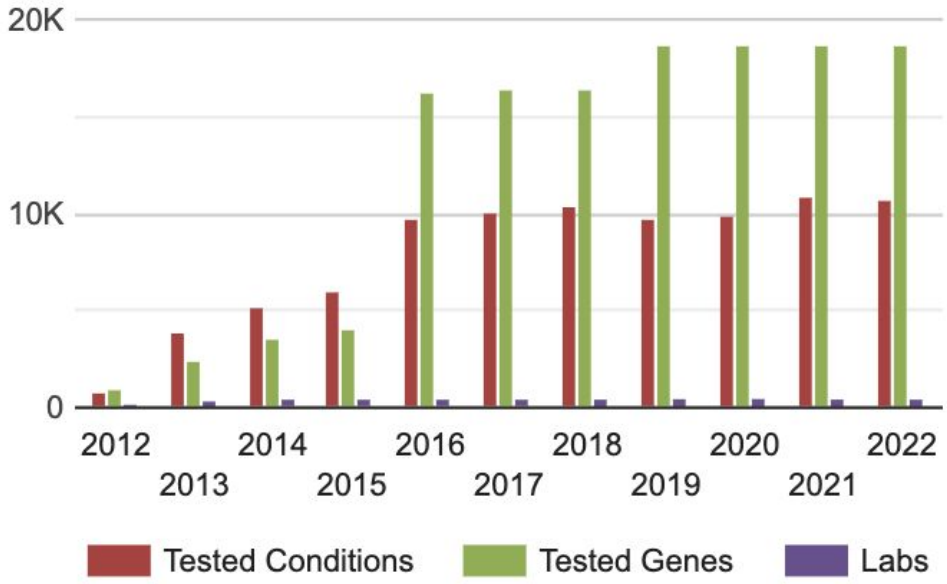
Challenges, and the path forward

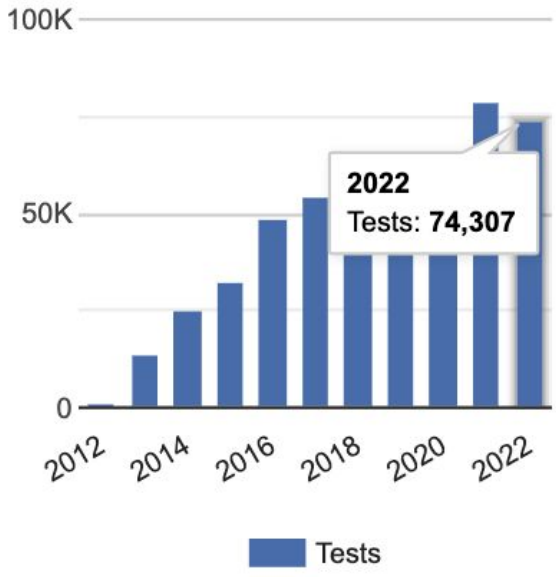
1. Learning curve
2. Ease of
3. Mul
4. Dive
5. Adc

The publication of FHIR DSTU2 included the creation of the FHIR Maturity Model (FMM). When new Resources are created, they are not

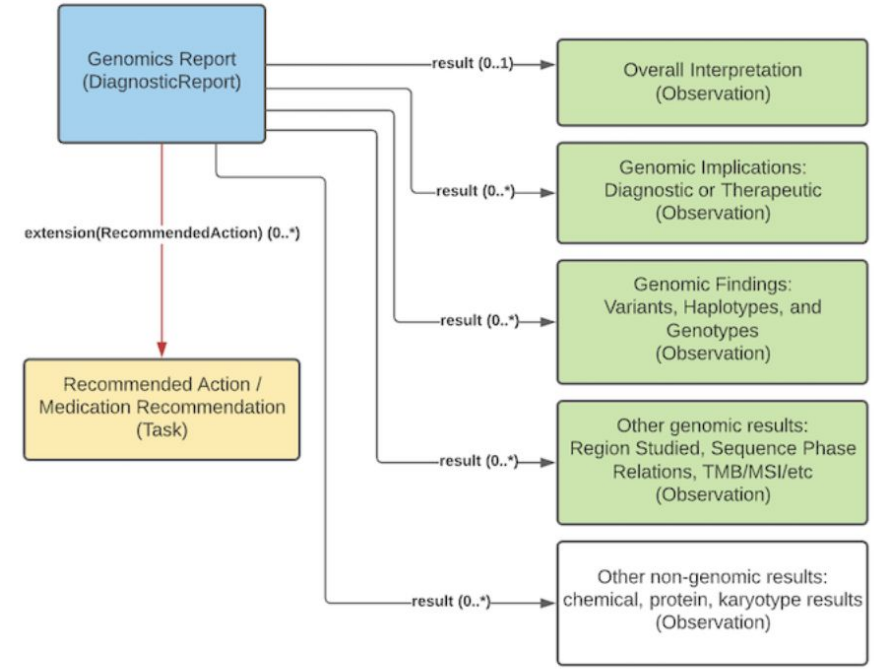
- **FMM0 (Draft)** – The resource is still in early development but has been accepted into the FHIR standard.

GTR Data

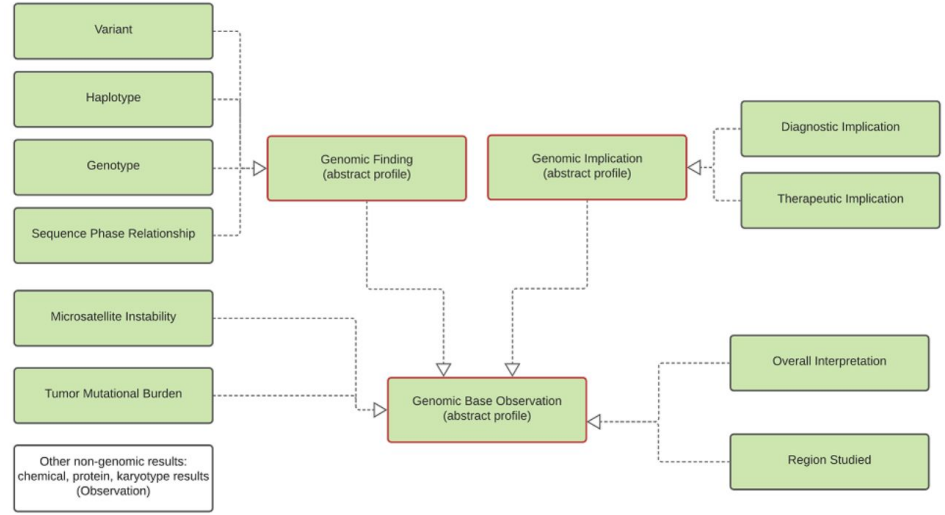




1. Clinical Genomics IG Learning Curve



Genomic Report Overview



Genomic Observations

4.2 Defining Variants

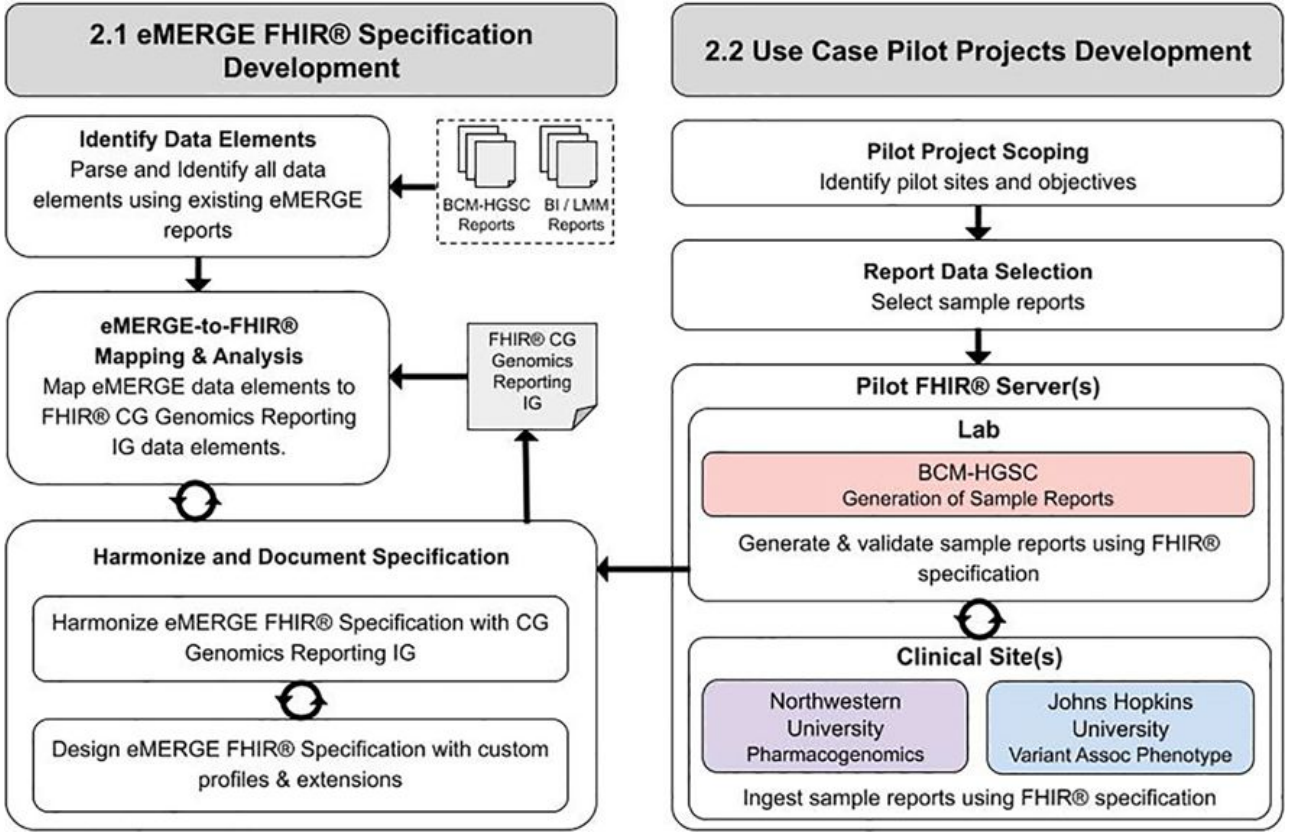
This Implementation Guide supports two reporting patterns for defining variants:

1. By describing the change using HGVS or ISCN nomenclature. Example HGVS-styled variant.
2. By providing multiple component details similar to VCF columns. Example VCF-styled variant.

For each variant reporting pattern, different components MUST be used to properly define the variant where possible. Other components MAY be used to provide additional information for cross referencing external sources or increasing human readability of the instance.

[Clinical Genomics GR IG]

2. Ease of Implementation



From publication "Genomic considerations for FHIR; eMERGE implementation lessons"

eMERGE III FHIR Pilot:
Larry Babb, Broad Institute
Luke Rasmussen, NU
Casey Overby Taylor, JHU
Mullai Murugan, BCM

Getting Started? Go here

3.Multiple Pilot Efforts

1. Creation of a FHIR specification and a pilot implementation for eMERGE Phase III;
2. Creation of a HLA Reporting IG based on the Genomics Reporting IG (STU1) led by Bob Milius at the NMDP;
3. A pilot project that utilizes the Genomics Reporting IG (STU1) at Cerner, in collaboration with a Diagnostic Laboratory;
4. Repr
5. An c
gGM

- *Gen*
- FHI

1.Completed Major

1. Composite Report - Section Grouping
2. Lab Defined Tests - Methodology, References, etc... (PlanDefinition)
3. Report Level Comments - Observation
4. Recommendations (Proposed) - (RecommendedAction - Task)
5. Nested & Indirect Result Referencing - hasMembers & derivedFrom?
6. Addition of chromosome to Variant

2.Completed Minor

1. New Identifier Type Code(s)
2. InhDisPath phenotype cardinality change
3. InhDisPath value (CC) made extensible
4. DR category cardinality changed to 0..*

2.Completed Minor (cont'd)

5. RelatedArtifact extension in Observation Components - Assessed Meds Citations (CG)
6. Distinction between Report Sign-Out/Off Date and Report Sent Date - (Sign Out = Issue) (OO)

3.Pending

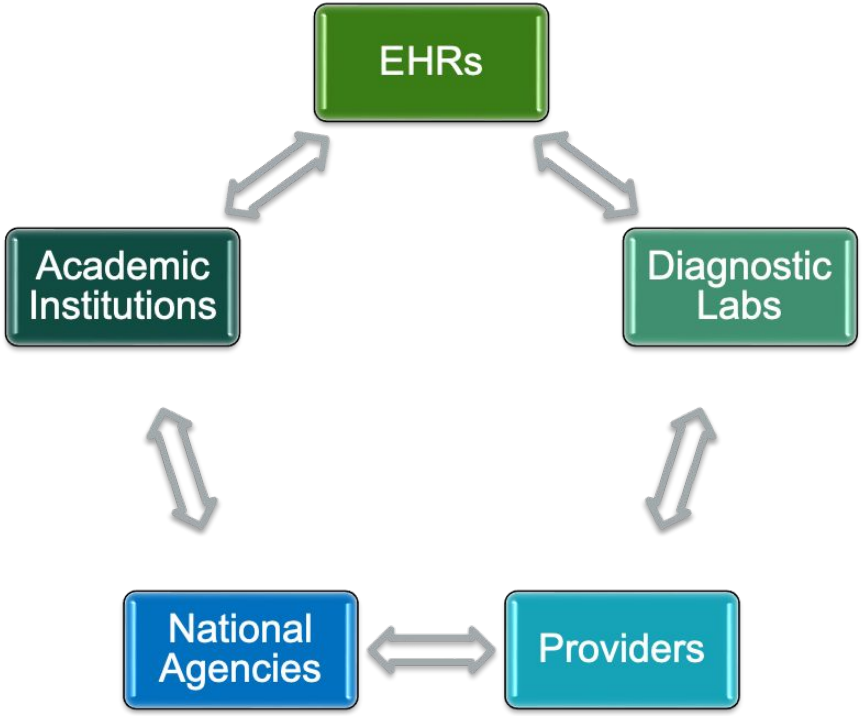
1. RecommendedAction Task reasonRef cardinality to 0..* (OO)
2. Add Age to US-Core Patient Profile (PatAdm)
3. Clinical vs Research Flag (Core)
4. Why is DR.code fixed to LOINC 81247-9? (CG)
5. RecommendedAction profile "code" should be extensible (CG)

4. Diversity of the tech landscape

- Open Source
  - HAPI
  - Microsoft FHIR Server
  - Etc.
- Industry Sponsored
  - SMILE CDR
  - Microsoft Azure Based
  - AWS
  - Google
- EHR Vendors' FHIR servers
- SMART Apps

5. Adoption and direction

- EHR Systems/DLs Engagement
- Path setting research effort
- Standards integration
- Tech growth
- Mandates



Acknowledgements

eMERGE Phase III
EHRI subgroup
FHIR Pilot subgroup
Larry Babb, Broad Institute
Ken Wiley, NHGRI
Luke Rasmussen, NU
Casey Overby Taylor, JHU

HL7 FHIR Clinical Genomics (CG)
CG working group chairs
CG working group members
Robert Freimuth, Mayo Clinic, IM
Ali Khalifa, Mayo Clinic, IM
Arthur Hermann, GenomeX, KP
May Terry, Mitre Corporation
FHIR Core working group

ONC Sync for Genes Phase 3
Allison Dennis, ONC
Kevin Chaney, ONC
Robert Freimuth, Mayo Clinic
Robert Milius, NMDP
Audacious Inquiry

Baylor College of Medicine
Richard Gibbs
Eric Venner
Fei Yan
Victoria Yi

Supporting Genomic Data Sharing through the Global Alliance for Genomics and Health

Heidi Rehm (Broad Institute/MGH)

The Global Alliance for Genomics and Health Mission...

The GA4GH aims to accelerate progress in genomic science and human health by **developing standards and framing policies for responsible genomic and health-related data sharing.**

GA4GH achieves this by...

- **Convening** stakeholders
- **Creating** standards and harmonized approaches through community consensus
- **Catalyzing** sharing of data
- But **does not** generate data, nor build primary infrastructure or perform research/clinical care that our standards support

GA4GH Organization Structure

Global Alliance for Genomics & Health

STRATEGIC ADVISORY/FUNDING

- SAB
- Funders Forum

OICR · Wellcome Sanger Institute · BROAD INSTITUTE · EMBL-EBI · NIH · wellcome · NIHR National Institute for Health Research · MRC Medical Research Council · GenomeCanada · CIHR IRSC

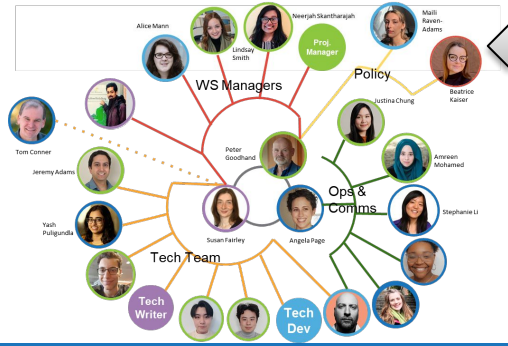
ORGANIZATIONAL GOVERNANCE

- Board of Directors
  - Finance, Audit, & Risk Committee
  - Executive Committee
  - Nominating Committee
- CEO
- Secretariat
  - Ops Team
  - Technical Standards Team
  - Work Stream Managers
  - Comms Team

TECHNICAL GOVERNANCE

- Standards Steering Committee

21 full-time staff



Executive Leadership


Peter Goodhand
Chief Executive Officer


Ewan Birney
Chair


Kathryn North
Vice Chair


Heidi Rehm
Vice Chair


Angela Page
Director of Strategy and Engagement

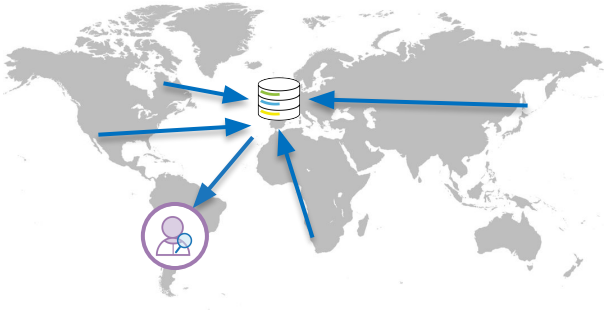

Susan Fairley
Chief Standards Officer

Assigned Experts & Active Contributors

Different Approaches to Data Sharing

Central Database

Genomic Knowledgebase

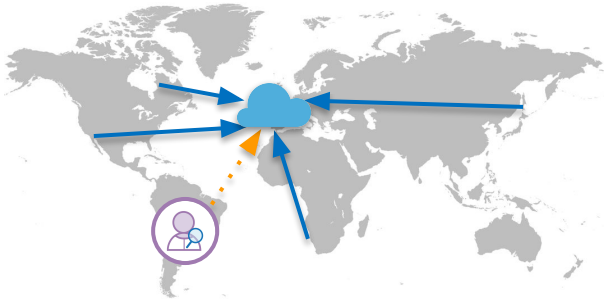


Aggregate data globally

Download, analyze locally

Secure Cloud

Large scale research datasets

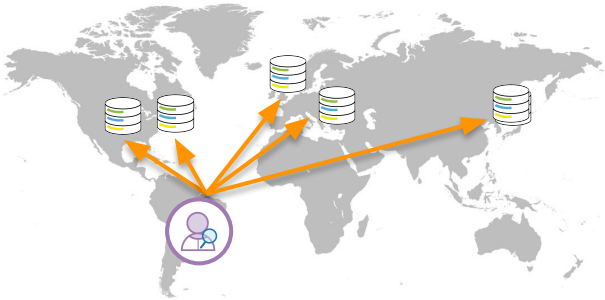


Aggregate data globally

Analyze centrally in secure cloud

Federation

Connecting national genomics initiatives



Host data locally

Visit data remotely and collate results

User

Data transmission

Secure access

How GA4GH Works

Technical Work Streams

- Discovery
- Large-Scale Genomics
- Data Use & Researcher IDs
- Cloud
- Genomic Knowledge Standards
- Clinical & Phenotypic Data Capture

Foundational Work Streams

- Regulatory & Ethics
- Data Security

Input from Driver Projects as exemplars of community

Output of standards that are taken up by the community

Partner Engagement

RARE DISEASES

Matchmaker Exchange

COMPLEX TRAITS

AUTISM SPEAKS

TOPMed

GEM Japan

Australian Genomics Health Alliance

All of Us Research Program

Genomics England

ClinGen

BRCA Exchange

CANCER

ARGO

NIH National Cancer Institute

EU Can Can.

VICC Variant Interpretation for Cancer Consortium

elixir

ENA

monarch initiative

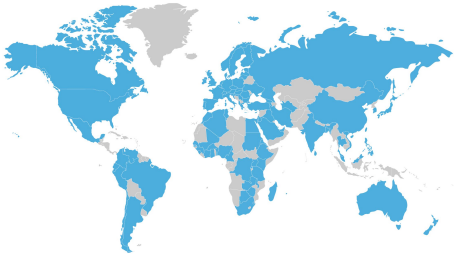
epi share

Human Cell Atlas

BASIC BIOLOGY

24 Driver Projects

Real-World Driver Projects



Federated Analysis Systems Project (FASP)

Starter Kit

Technical Alignment Subcommittee (TASC)

Equity, Diversity and Inclusion Advisory Board

GA4GH 2020-2022 Strategic Roadmap

Global Alliance for Genomics & Health

Clin/Pheno Data Capture	Phenopackets	Pedigree Representation			
Cloud	Workflow Execution Service API	Tool Registry Service	Data Repository Service API	Task Execution Service	Cloud Testbed Interoperability
Discovery	Beacon API	Service Info	Service Registry	Data Connect API	
Data Security	Authentication & Authorization Infrastructure	Data Security and Infrastructure Policy	Risk Assessment	Breach Response Protocol	Cloud Security & Privacy
Data Use & Researcher Identity	Data Use Ontology	GA4GH Passports	Machine Readable Consent Guidance		
Genomic Knowledge Standards	Variation Representation	Variation Annotation			
Large Scale Genomics	htsget API	refget API	Read file formats	Variation File Formats	Crypt4GH	rnaget API	BED File Format
Regulatory & Ethics	Framework	Policy Initiatives	Responsible Data Sharing Resource Documents	GDPR Forum	

Ethics Review Recognition, Accountability, Consent, Privacy & Security
Technical Standards & IP, Return of Results

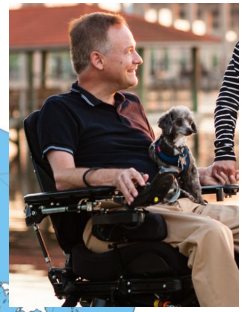
Challenges in rare disease gene discovery

- 75% of rare disease cases remain unsolved
- 4,631 genes implicated in at least one disease but evidence for >10,000 more genes yet to be discovered for Mendelian disease (Bamshad, et al. AJHG 105, 448–455, 2019)
- The remaining genetic diseases are very, very rare – difficult for any one investigator to amass enough cases to implicate a new disease gene

Principles of Gene Matching

Individual with rare disease

Clinical geneticist



Individual with rare disease

Rare disease researcher

Phenotypic Data
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5

Genotypic Data
Gene D

→

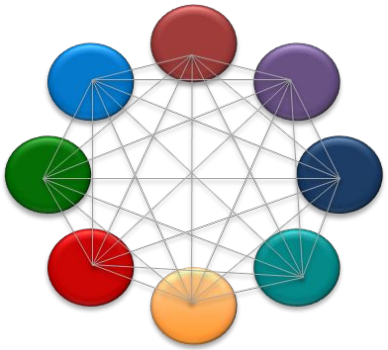
Genomic Matchmaker

←

Genotypic Data
Gene D

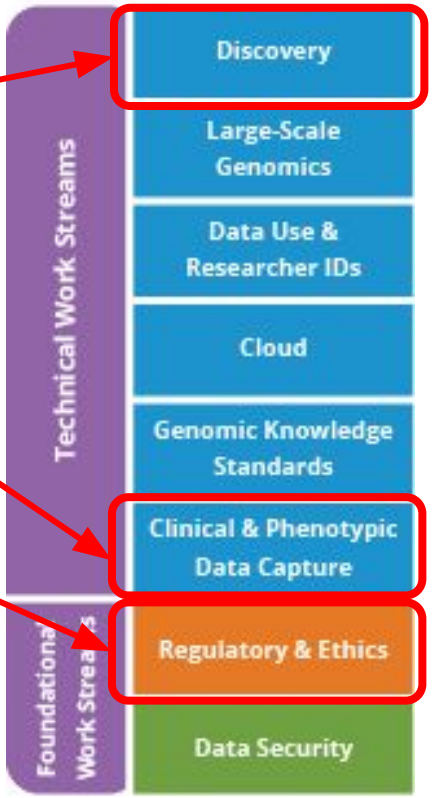
Phenotypic Data
Feature 1
Feature 3
Feature 4
Feature 5
Feature 6

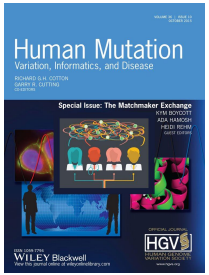
Developing the MME Federated Network using GA4GH Standards



Use of GA4GH standards:

- API for data exchange
  - ID (Mandatory) +/- Label
  - Submitter (Mandatory)
  - Phenotypic Features and/or Gene Names (Mandatory)
  - Disorders (Optional) - OMIM or OrphaNet
  - Sex, Age of Onset, Inheritance (Optional)
- Clinical and phenotypic data capture standards
- Consent framework for data sharing



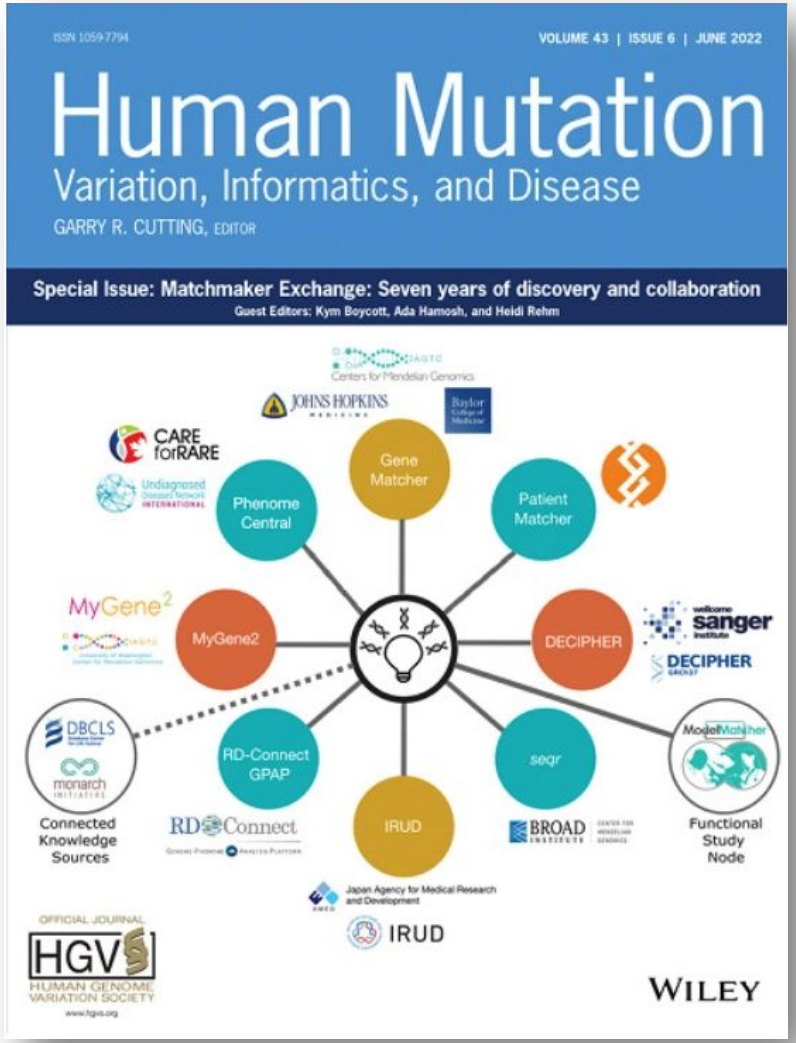


Philippakis et al. **The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery.** Hum Mutat. 2015;36(10):915-21.

Buske et al. **The Matchmaker Exchange API: automating patient matching through the exchange of structured phenotypic and genotypic profiles.** Hum Mutat. 2015;36(10):922-7

16 papers in a special issue of Human Mutation (Vol 36, Issue 10, Oct 2015)

www.matchmakerexchange.org



EDITORIAL INTRODUCTION

Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking

Kym M. Boycott, Danielle R. Azzariti, Ada Hamosh, Heidi L. Rehm

Human Mutation. 2022;43:659–667. https://doi.org/10.1002/humu.24373

- The impact of **GeneMatcher** on international data sharing and collaboration
- **PhenomeCentral:** 7 years of rare diseas
- **DECIPHER:** Supporting the interpretatio variant data to advance diagnosis and res
- **seqr**: A web-based analysis and collabo
- **PatientMatcher**: A customizable Pythor rare disease patients via the Matchmaker
- The **RD-Connect Genome-Phenome An** gene discovery for rare diseases
- Advances in the development of **PubCa** interface and matching algorithm
- **ModelMatcher**: A scientist-centric online platform to facilitate collaborations between stakeholders of rare and undiagnosed disease research
- Discovery of over 200 new and expanded genetic conditions using GeneMatcher
- A clinical laboratory's experience using GeneMatcher—Building stronger gene–disease relationships
- Diagnostic testing laboratories are valuable partners for disease gene discovery: 5-year experience with GeneMatcher
- **Variant-level matching** for diagnosis and discovery: Challenges and opportunities
- **Beacon v2** and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond
- **Genomics4RD**: An integrated platform to share Canadian deep-phenotype and multiomic data for international rare disease gene discovery

Over 10,000 candidate genes
from ~200,000 patients
from >12,000 contributors
from 98 countries
Over 1000 genes discovered through matchmaking

GeneDx
Illumina
Ambry

Three clinical labs had 1040/3819 (27%) gene discoveries validated through MME

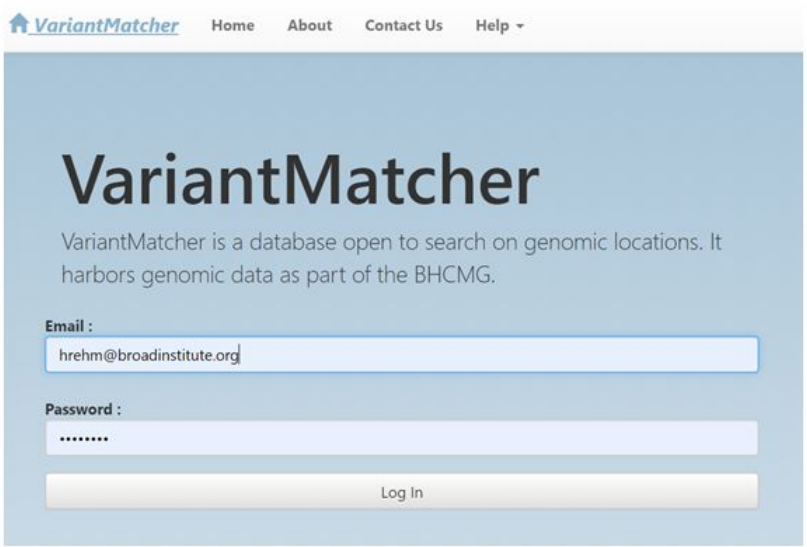
(a) Two-sided matchmaking

Matchmaker Exchange

(b) One-sided matchmaking

(c) Zero-sided matchmaking

Boycott KM, Azzariti DR, Hamosh A, Rehm HL. Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking. Human Mutation. June 2022;43:659–667.



VariantMatcher (VM) created by:

- Nara Sobreira
- François Schiettecatte
- Ada Hamosh
- BHCMG Center for Mendelian Genomics

Your search included the following features:
Hypotonia, Microcephaly, Global Developmental delay, Esotropia

---

A submission match notification, for your search: '6:34004293:T>C', was sent to the following:

BHXXXX - Patient - Affected - 6:34004293:T>C
Salmo Raskin - genetika@genetika.com.br - PUC Brazil
Bilateral Cleft

BHXXXX - Patient - Affected - 6:34004293:T>C
Hamza Aziz - haziz2@jhmi.edu - JHU
Bicuspid Aortic valve, Aneurysm, ascending aortic

BHXXXX - Patient - Affected - 6:34004293:T>C
Samantha Penney - penney@bcm.edu - Baylor College of Medicine
Encephalopathy, Ataxia, Hypotonia

BHXXXX - Patient - Affected - 6:34004293:T>C
Samantha Penney - penney@bcm.edu - Baylor College of Medicine
Ataxia, Spasticity, adult onset spinocerebellar ataxia

BHXXXX - Mother - Unaffected - 6:34004293:T>C
Filippo Vairo - fvairo@hcpa.edu.br - Hospital de Clinicas de Porto Alegre

BHXXXX - Father - 6:34004293:T>C
Daryl Scott - dscott@bcm.edu - Baylor College of Medicine

BHXXXX - Mother - 6:34004293:T>C
Samantha Penney - penney@bcm.edu - Baylor College of Medicine

BHXXXX - Father - 6:34004293:T>C
Samantha Penney - penney@bcm.edu - Baylor College of Medicine

Please do not reply to this email, it was sent from an unattended email address; however, you can email us at variantmatcher@jhmi.edu or use the contact form.

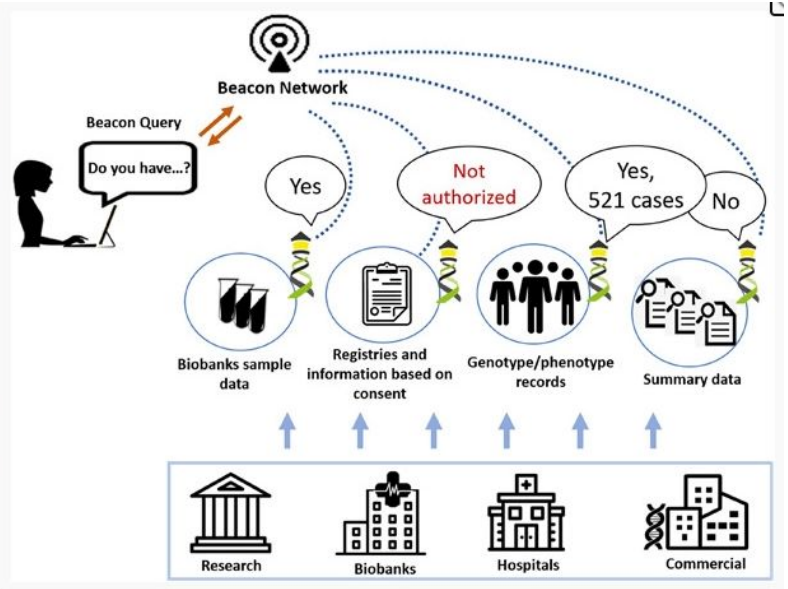
Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

INFORMATICS | Open Access

Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond

Jordi Rambla, Michael Baudis, Roberto Ariosa, Tim Beck, Lauren A. Fromont, Arcadi Navarro, Rahel Paloots, Manuel Rueda, Gary Saunders, Babita Singh, John D. Spalding ... See all authors

First published: 17 March 2022 | https://doi.org/10.1002/humu.24369 | Citations: 1



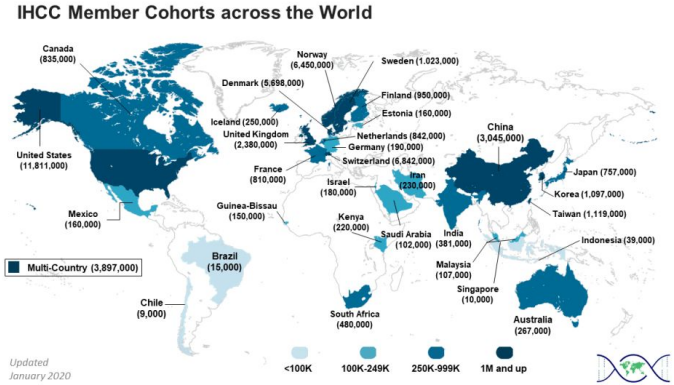
Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

DATABASES | Open Access

Variant-level matching for diagnosis and discovery: Challenges and opportunities

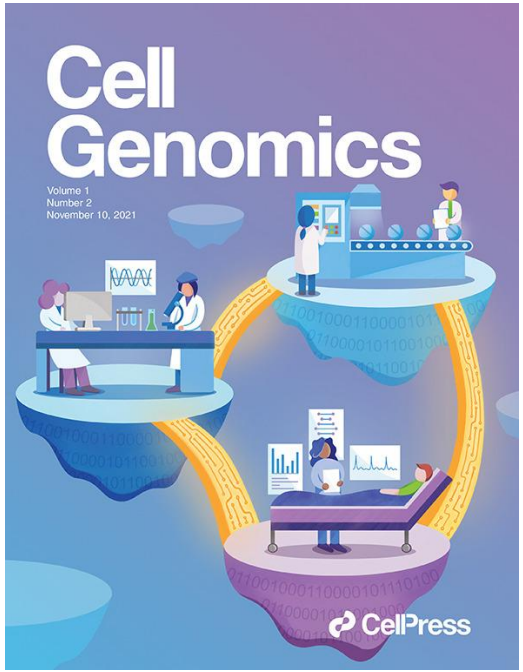
Eliete da S. Rodrigues, Sean Griffith, Renan Martin, Corina Antonescu, Jennifer E. Posey, Zeynep Coban-Akdemir, Shalini N. Jhangiani, Kimberly F. Doheny, James R. Lupski, David Valle ... See all authors

First published: 22 February 2022 | https://doi.org/10.1002/humu.24359 | Citations: 1



MyGene2, Geno2MP, VariantMatcher, Franklin

GA4GH Marker Paper and other GA4GH Work Product Publications in November 2021 Issue of Cell Genomics



Cell Genomics

CellPress
OPEN ACCESS

Perspective

GA4GH: International policies and standards for data sharing across genomic research and healthcare

Heidi L. Rehm,[1,2,47] Angela J.H. Page,[1,3,*] Lindsay Smith,[3,4] Jeremy B. Adams,[3,4] Gil Alterovitz,[5,47] Lawrence J. Babb,[1] Maxmillian P. Barkley,[6] Michael Baudis,[7,8] Michael J.S. Beauvais,[3,9] Tim Beck,[10] Jacques S. Beckmann,[11] Sergi Beltran,[12,13,14] David Bernick,[1] Alexander Bernier,[9] James K. Bonfield,[15] Tiffany F. Boughtwood,[16,17] Guillaume Bourque,[9,18] Sarion R. Bowers,[15] Anthony J. Brookes,[10] Michael Brudno,[18,19,20,21,38] Matthew H. Brush,[22] David Bujold,[9,18,38] Tony Burdett,[23] Orion J. Buske,[24] Moran N. Cabili,[1] Daniel L. Cameron,[25,26] Robert J. Carroll,[27] Esmeralda Casas-Silva,[123] Debyani Chakravarty,[29] Bimal P. Chaudhari,[30,31] Shu Hui Chen,[32] J. Michael Cherry,[33] Justina Chung,[3,4] Melissa Cline,[34] Hayley L. Clissold,[15] Robert M. Cook-Deegan,[35] Mélanie Courtot,[23] Fiona Cunningham,[23] Miro Cupak,[6] Robert M. Davies,[18] Danielle Denisko,[19] Megan J. Doerr,[36] Lena I. Dolman,[19] Edward S. Dove,[38] L. Jonathan Dursi,[20,39] Stephanie O.M. Dyke,[9] James A. Eddy,[37] Karen Eilbeck,[40] Kyle P. Ellrott,[22] Susan Fairley,[3,23] Khalid A. Fakhro,[41,42] Helen V. Firth,[15,43] Michael S. Fitzsimons,[44] Marc Fiume,[6] Paul Flicek,[23] Ian M. Fore,[29] Mallory A. Freeberg,[23] Robert R. Freimuth,[45] Lauren A. Fromont,[52] Jonathan Fuerth,[6] Clara L. Gaff,[16,17] Weiniu Gan,[33] Elena M. Ghanaim,[46] David Glazer,[47] Robert C. Green,[1,48,49] Malachi Griffith,[50] Obi L. Griffith,[50] Robert L. Grossman,[44] Tudor Groza,[51] Jaime M. Guidry Auvil,[29] Roderic Guigó,[13,52] Dipayan Gupta,[23] Melissa A. Haendel,[53] Ada Hamosh,[54] David P. Hansen,[16,83] Reece K. Hart,[1,100,124] Dean Mitchell Hartley,[55] David Haussler,[35] Rachele M. Hendricks-Sturrup,[56] Calvin W.L. Ho,[57] Ashley E. Hobb,[6] Michael M. Hoffman,[19,20,21] Oliver M. Hofmann,[26] Petr Holub,[58,59] Jacob Shujui Hsu,[60] Jean-Pierre Hubaux,[61] Sarah E. Hunt,[23] Ammar Husami,[62] Julius O. Jacobsen,[63] Saumya S. Jamuar,[64,65] Elizabeth L. Janes,[3,66] Francis Jeanson,[129] Aina Jené,[52] Amber L. Johns,[67,68] Yann Joly,[9] Steven J.M. Jones,[69] Alexander Kanitz,[8,70] Kazuto Kato,[71] Thomas M. Keane,[23,72] Kristina Kekesi-Lafrance,[3,9] Jerome Kelleher,[73] Giselle Kerry,[23] Seik-Soon Khor,[74,75] Bartha M. Knoppers,[9] Melissa A. Konopko,[76] Kenjiro Kosaki,[77] Martin Kuba,[59] Jonathan Lawson,[1] Rasko Leinonen,[23] Stephanie Li,[1,3] Michael F. Lin,[78] Mikael Linden,[79,80] Xianglin Liu,[86] Isuru Udara Liyanage,[23] Javier Lopez,[101] Anneke M. Lucassen,[81] Michael Lukowski,[44] Alice L. Mann,[3,15] John Marshall,[68] Michele Mattioni,[82] Alejandro Metke-Jimenez,[83] Anna Middleton,[84,85] Richard J. Milne,[84,85] Fruzsina Molnar-Gabor,[86] Nicola Mulder,[87] Monica C. Munoz-Torres,[53] Rishi Nag,[23] Hidewaki Nakagawa,[88,89] Jamal Nasir,[90] Arcadi Navarro,[52,91,92,93] Tristan H. Nelson,[94] Ania Niewielska,[23] Amy Nisselle,[17,26,95] Jeffrey Niu,[20] Tommi H. Nyrönen,[79,80] Brian D. O'Connor,[1] Sabine Oesterle,[6] Soichi Ogishima,[96] Laura A.D. Paglione,[97,98] Emilio Palumbo,[13,52] Helen E. Parkinson,[23] Anthony A. Philippakis,[1] Angel D. Pizarro,[99] Andreas Prlic,[100] Jordi Rambla,[13,52] Augusto Rendon,[101] Renee A. Rider,[46] Peter N. Robinson,[102,103] Kurt W. Rodarmer,[104] Laura Lyman Rodriguez,[105] Alan F. Rubin,[25,26] Manuel Rueda,[52] Gregory A. Rushton,[1] Rosalyn S. Ryan,[106] Gary I. Saunders,[76] Helen Schuilenburg,[23] Torsten Schwede,[8,70] Serena Scollen,[76] Alexander Senf,[107] Nathan C. Sheffield,[108] Neerjah Skantharajah,[3,4] Albert V. Smith,[109] Heidi J. Sofia,[45] Dylan Spalding,[79,80] Amanda B. Spurdle,[110] Zornitza Stark,[16,17,26] Lincoln D. Stein,[4,19] Makoto Suematsu,[77] Patrick Tan,[64,111,112] Jonathan A. Tedds,[76] Alastair A. Thomson,[33] Adrian Thorogood,[9,113] Timothy L. Tickle,[1] Katsushi Tokunaga,[75,114] Juha Törnroos,[79,80] David Torrents,[92,116] Sean Upchurch,[115] Alfonso Valencia,[92,116] Roman Valls Guimera,[26] Jessica Vamathevan,[23] Susheel Varma,[23,117] Danya F. Vears,[17,26,95,118] Coby Viner,[19,20] Craig Voisin,[119] Alex H. Wagner,[31,32] Susan E. Wallace,[10] Brian P. Walsh,[22] Vivian Ota Wang,[29] Marc S. Williams,[94] Eva C. Winkler,[120] Barbara J. Wold,[115] Grant M. Wood,[1] J. Patrick Woolley,[73] Chisato Yamasaki,[71] Andrew D. Yates,[23] Christina K. Yung,[4,121] Lyndon J. Zass,[87] Ksenia Zaytseva,[9,122] Junjun Zhang,[4] Peter Goodhand,[4,3] Kathryn North,[17,26] and Ewan Birney[23,123]

Get Involved! Visit GA4GH.ORG

Join a Work Stream!

Contact secretariat@ga4gh.org

Become an Organizational Member

ga4gh.org/members

Subscribe to GA4GH Updates

ga4gh.org/subscribe

Interoperability Opportunities & Challenges with the Cloud and STRIDES

Nick Weber (NIH STRIDES)

Interoperability Opportunities & Challenges with STRIDES & Cloud

NCPI Spring Workshop

Nick Weber

Program Lead, NIH STRIDES Initiative | Program Manager, Cloud Services
Center for Information Technology

NIH National Institutes of Health

June 23, 2022

NIH STRIDES Initiative

The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability

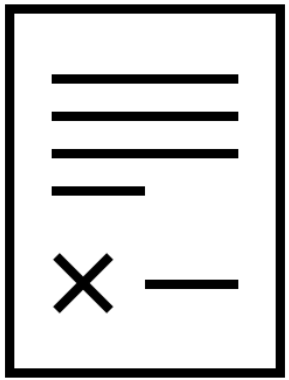
- State-of-the-art data storage and computational capabilities
- Training and education for researchers
- Innovative technologies such as artificial intelligence and machine learning
- Professional engineering and technical support

Partnerships with aws Google Cloud Microsoft Azure

Two Core Components of STRIDES

1) Other Transaction Agreement

Enables NIH-funded institutions to leverage STRIDES benefits



2) NIH Enterprise Cloud Platforms & Services

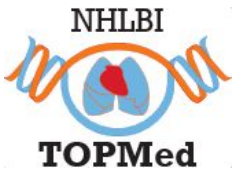
Supports efficient and secure NIH-wide use of the cloud for IRP needs and/or ICs' institutional management requirements

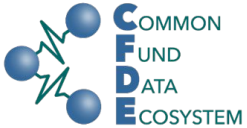
Example: U-Pitt enrolled in STRIDES. NIH-funded PIs supported by NIGMS (U24), NIDDK (U01), & NIDCD (R44) benefit from STRIDES discounts using the cloud to support their award/research activity

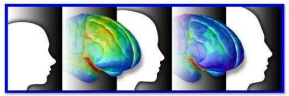
Example: NIA's Laboratory of Neurogenetics analyzes WGS data on the cloud for Parkinson's, Alzheimer's, and other dementias, and manages general lab infrastructure for data storage and deposition into the AMP PD data repository & knowledge platform

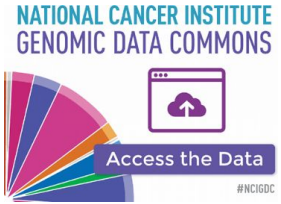
Cross-Cutting: Discounts, Training, Professional Services, & Vendor Support

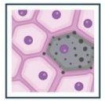
Sample of STRIDES-Supported Research Programs



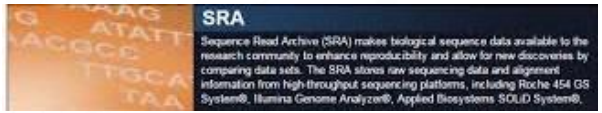








SenNet



NEW: NIH Cloud Lab Offering

A cloud testbed allowing researchers to “try before they buy”

Primary Cloud Lab Use Cases

Exploring the Cloud Consoles
Researchers can gain an understanding of the look and feel of cloud environments before they jump into a full STRIDES account for research

Supplementing Cloud Training
Researchers can use the sandbox to strengthen their understanding of cloud training or follow along with training content in a separate environment.

Experimenting with Simple Cloud Solutions
Researchers interested in solutions for specific scientific tasks can use the sandbox to build proof of concept or other simple solutions to understand LOE and other details for production.

Benchmarking Costs
Testing out different tools and configurations (instance types, sizes, etc.) to optimize research analyses

NIH Cloud Lab (continued)

NIH Cloud Lab is a no-cost (to you), 90-day pilot program that enables NIH-funded researchers to try commercial cloud services in an NIH-approved environment. The Cloud Lab provides training and guardrails to protect against financial and security risks.

Full Access to the Cloud Console

- Deploy a full range of resources
- CPU or GPU VMs
- Managed Jupyter notebooks
- Advanced AI/ML capabilities
- Bioinformatic workflow managers
- Access to compute clusters

Bioinformatic Tutorials to Speed Uptake

- Variant Calling
- GWAS
- Medical Imaging
- RNA seq
- Single Cell RNA seq
- Proteomics
- Using HPC environments in the cloud

Broad Access Across the NIH Community

- Intramural
  - AWS – Beta Testing
  - GCP – Beta Testing
- Extramural
  - AWS – Limited Beta Testing
  - GCP – *Conditional* Limited Beta Testing

Let us know you're interested at: cloud.NIH.gov/resources/cloudlab

Interoperability Challenges & Considerations

- New Data Management & Sharing Policy
- Modularity / portability / reusability
- Cross-cloud billing integration
- Cost enforcement
- Cost estimation
- Institution-level data mesh "nodes"?
- Pilot programs for standardization around products like Kubernetes, Docker, etc.?
- RAS as an underpinning for billing auth?
- NIH Cloud Lab examples / source code?
- NIH Cloud Lab & community contributions?

Interoperability is a challenge not only for data resources and analysis platforms built on the cloud, but for core cloud infrastructure *itself*

Build Research Capacity *in Partnership with* Central IT's Cloud Ops Team

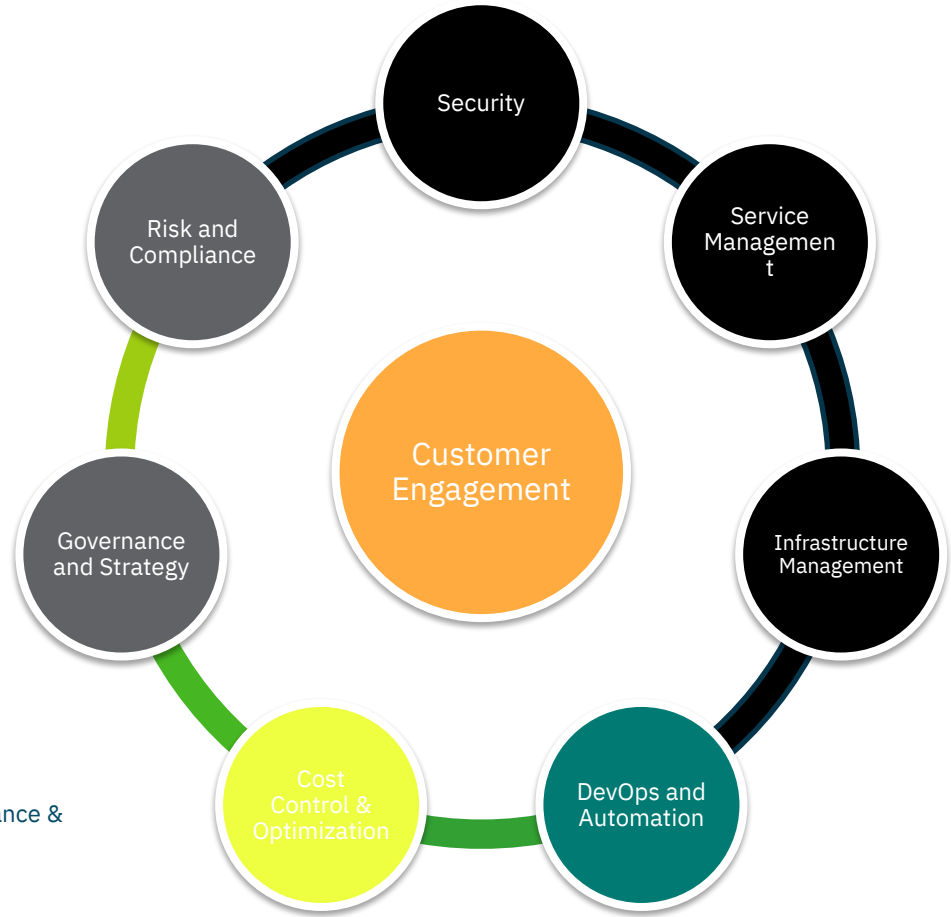
Interoperability in general requires mastery of the fundamentals (see: RAS); cloud infrastructure interoperability is no different

Customer Engagement
- Assessment & planning
- Onboarding
- Architecture consultation
- Shared responsibility
- Cloud migration

Risk & Compliance
- FISMA, FedRAMP, & CSF
- NIST 800-37, -53, & -171
- Continuous monitoring

Governance and Strategy
- Cloud demand prioritization
- Service roll-out
- Standards, guardrails, & reference architectures
- Cloud operating model & transformation office
- Policy roll-out
- Disaster recovery & COOP strategy

Cost Control & Optimization
- Consolidated billing
- Cost allocation & optimization
- Budget alerting & control
- Workload optimization for performance & cost



Security
- Identity & access management
- Vulnerability management
- Data protection & privacy
- Security monitoring
- Infrastructure security hardening
- Incident response
- Cloud access security broker

Service Management
- Automated monitoring, ticketing & alerting
- 24/7 service desk operations
- Change & configuration management
- Incident & problem management
- Monitoring & event management
- Self service & service catalog

Infrastructure Management
- Platform & technologies setup
- Infrastructure provisioning
- Network provisioning and management
- Core infrastructure maintenance and modernization
- Disaster recovery & COOP

DevOps and Automation
- Release management
- Continuous integration
- Continuous deployment
- Cloud automation pipeline

Concurrent Breakout Session

Topic 1: Bringing researchers to cloud computing	Tiffany Miller
Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses	Jack DiGiovanna
Topic 3: What technologies and data types are missing across platforms?	Ken Wiley
Topic 4: Diversifying genomic data science	Asiyah Lin | Kim Albero
Topic 5: Flagship use cases for interoperability	Michael Schatz

2:35 PM - 3:50 PM EDT

Topic 1: Bringing researchers to cloud computing

Barriers to bringing researchers to cloud computing	Strategies for getting around barrier
"Expensive"- Academics can often view "on prem" as free, but everything that is not free is expensive. Furthermore, there is a notion of direct and indirect costs that must be budgeted. (Mike S)	
"Cost education/Fear of overspend" - Not understanding how much stuff costs in this new way of working	1. Cloud Lab from Strides (maybe? If the user could make use of this on an analysis platform)
"Learning curve for doing science"- There is a learning curve and time must be spent preparing to use the cloud, translating pipelines to it, etc.	1. Incentivizing learning w/ training awards?
"Value proposition"-Is the value of the cloud worth the time to learn?	1. If we can educate folks on the 'jump off point' when working on the cloud can improve their ROI of time and money, a lot of the other barriers might become easier to address (Ravinder)
"Policy"- Aligning data policy w/ technology	- Educate Policy people and program officers and include in development - Ex. Pick IC w/ knowledge of cloud and transfer knowledge over to NIBIB (just for example). Perhaps policy people transfer knowledge to other policies across ICs
"Which analysis platform is for you?" Do I use native compute, Terra, SBG? Etc.	1. Map that shows where things are… and why you'd choose this or that to learn

For notes and the table see here: https://docs.google.com/document/d/1NnYE84dRLSRtCBtVc2j8aOskQfDAEIXcT-nPsDan3XQ/edit#

Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses

Provenance is a higher priority than perfect reproducibility

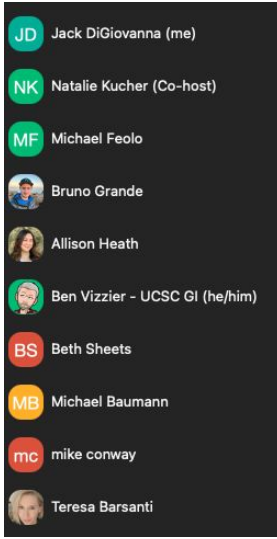
First step would be more information about data used

- Metadata exchange (dataset level, aggregate, subject level)
- Accessioning space (am I speaking AnVIL or KidsFirst, DOIs?)

Two types of data releases important for different goals

Provenance would help for multiple situations (retractions, submissions, bug-fixes, tool improvements)

We have many of the components for analysis reproducibility but are not yet at the point of checkpoint and restart

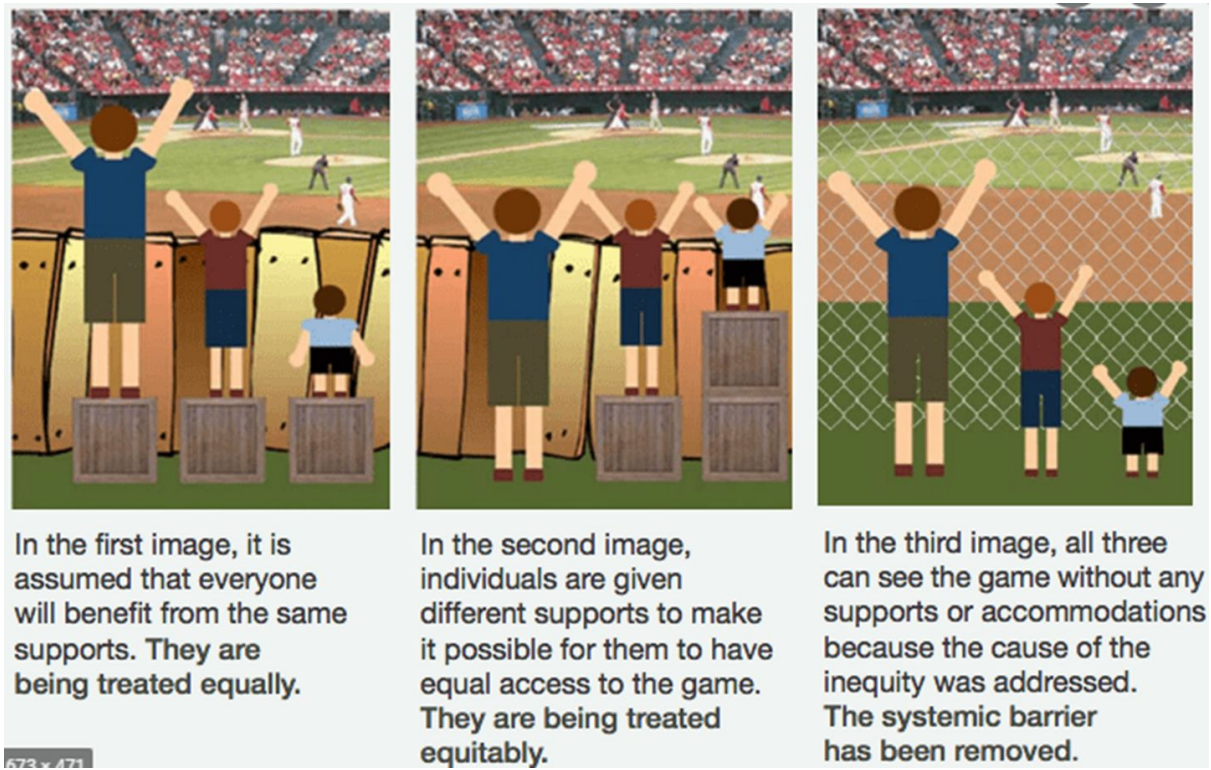


Topic 3: What technologies and data types are missing across platforms?

- Linking by phenotypes
  - Highly valuable for combining datasets together, but a lot of difficulties.
    - Phenotypes need to be standardized.
    - Need provenance - how were these collected?
    - Negative phenotypes - was a phenotype observed to be absent? Or not measured?
  - Tools that translate codes across ontologies would be helpful here.
- Clinical data notes
  - Can information be extracted out of these? Medical NLP tools?
    - One person's experience: still needs a bit to go.
    - Confused participants and their family members.
    - Can't translate and assign HPO terms.
  - Notes are not for the purpose of telling researchers info, they are for the patient care team.
    - Generally, physicians put notes all over the place. Professional note takers would help.
    - Billing codes could be useful, but again, not clinical focused.

Topic 4: Diversifying genomic data science

Discussant: Asiyah Lin (NIH), Kim Albero (MITRE), Jay Ronquillo (NIH), Rabia Begum(Genome Medicine), Matthew Meersman (MITRE), Marcia Fournier (NIH), Michelle Salter(Deloitte)



Link to Dr. Albero's slides

Key points

- Data diversity in NCPI cloud platforms?
- Pull data together for small under-represented populations – larger cohort building
- Utilize All of Us data
- Ethical issues – pulling data – re-identify – data privacy and security
- Provide a safe and secure environment for the under-represented or minority groups to involve in the science
- Missing the emphasize on diversity in our activities!
- Funding:
  - Congressional funding support for diversity related research
  - Adding diversity into the Funding Opportunity Announcement for NCPI

Next step

- Starting point: A small **data diversity investigation** to all NCPI platform datasets.
  - report back to the next workshop.
- Call for participation: asiyah.lin@nih.gov
- Still a lot needs to be done in diversity, equity, and inclusive area

Topic 5: Flagship use cases for interoperability

- We've heard quite a bit about Small Fish
  - Enabling small scale projects to effectively use what's already been built.
- Big Fish
  - Enable organizations and large scale projects
- Big Fish and small fish - NCPI's success will be in achieving both
- New NIH data management sharing policy will enable broader sharing of processed data outcomes
  - Important to make interoperable
  - challenging to harmonize given that they have already been analyzed
- Generalist repositories : May be most effective for partially processed, open access data. The repositories do account for the long tail of data sharing.
  - How can researchers find data across the 7 or 8 generalized repositories?
  - How can we consistently share metrics across the repositories?

Summary and Future Directions

Michael Schatz (Johns Hopkins University)