

# 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

---

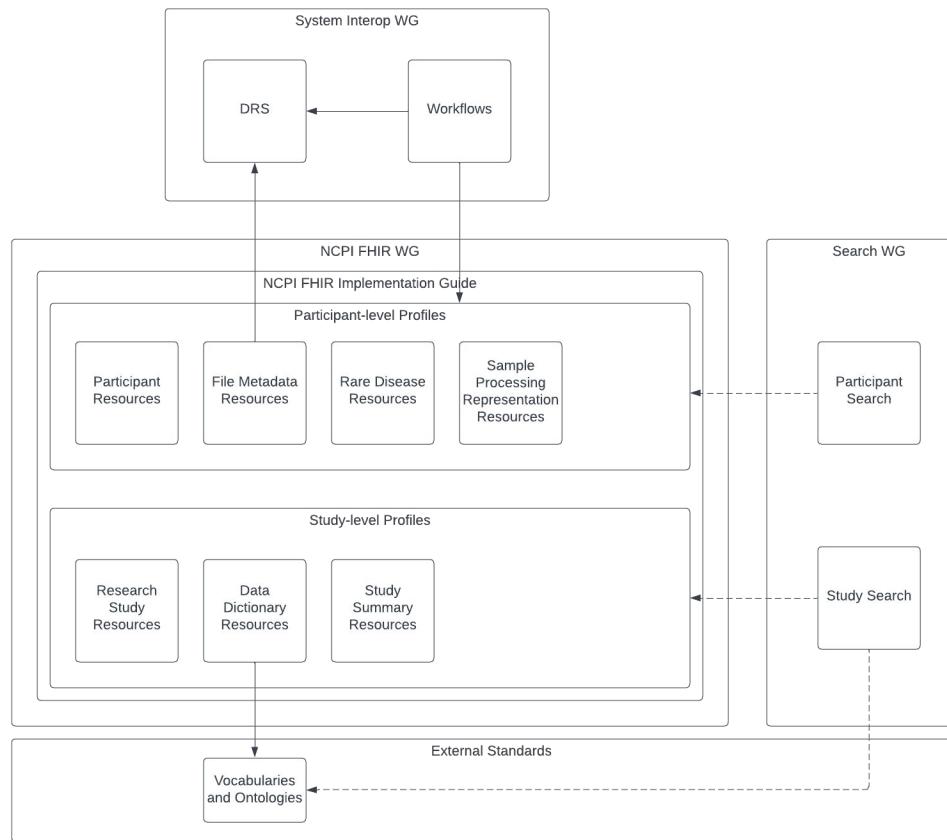


1. To provide an API to allow access to study and participant level data.
2. To provide standardized structures for study and participant data.
3. To enable structured semantics for data where available.

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



- Highlighted community groups



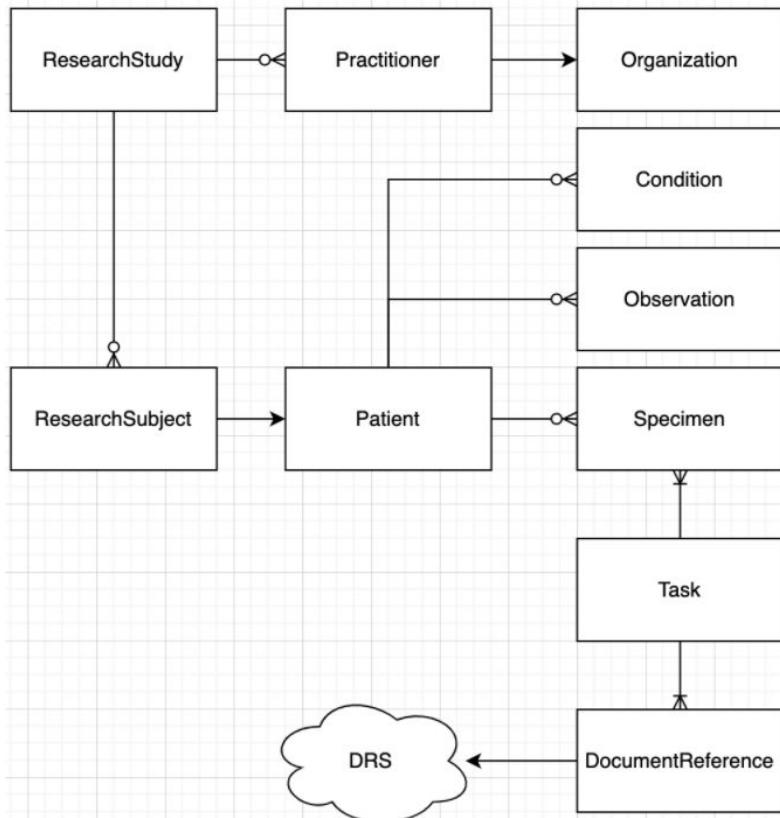
- Kids First: Production FHIR Services deployed
  - <https://kf-api-fhir-service.kidsfirstdrc.org/>
  - Open access data, requires login to KF Portal
- dbGaP: Public data services deployed
  - <https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/>
  - Study level data only
  - Work in progress on controlled access data, pilot implementations complete
- AnVIL: Non-production service pilots
  - Test deployment indexing AnVIL data across Terra
  - Pilot study specific ETL

- 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!

Study Information

Diseases and Syndromes

Laboratory Measures

Participant Demographics

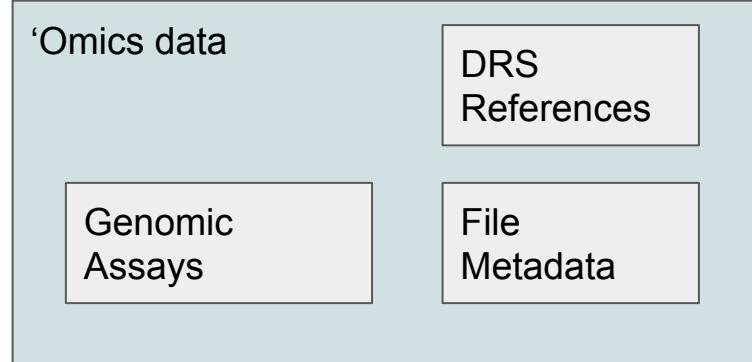
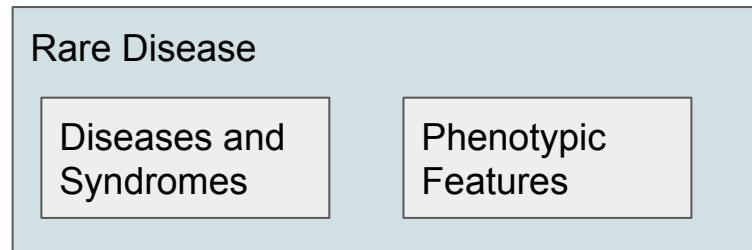
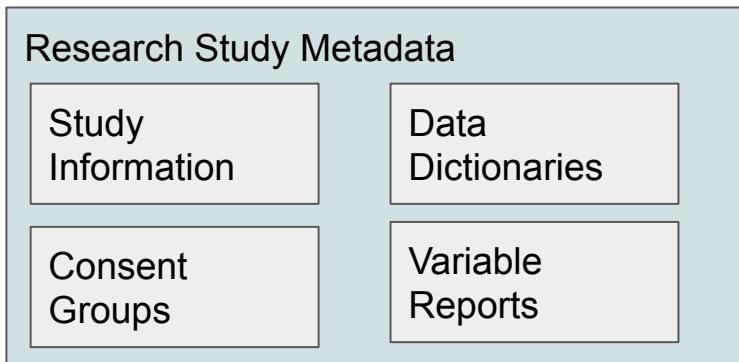
File Metadata

Genomic Assay Metadata

Phenotypic Features

# 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

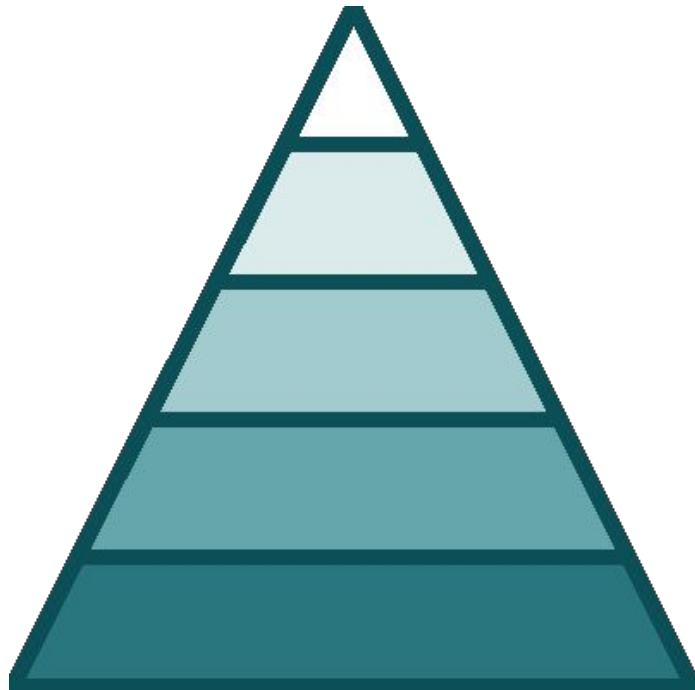
---

*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>

The screenshot shows the NCPI Portal homepage. At the top, there's a navigation bar with tabs for Overview, Platforms, Technologies, Datasets, Demonstration Projects, Training, and Updates. Below the navigation bar, there's a sidebar with links for About Us, Guiding Principles, and Working Groups. The main content area features a heading "NIH Cloud Platform Interoperability Effort" and a sub-section "Helping to create a federated genomic data ecosystem". It includes a paragraph about the effort's purpose and a diagram illustrating the federated ecosystem. The diagram shows five interconnected cloud icons: AnVIL, BioData CATALYST, National Cancer Institute Data Commons, Kids First Data Resource Center, and National Library of Medicine National Center for Biotechnology Information.

## Participating platforms

The screenshot shows the "Participating platforms" page. At the top, there's a navigation bar with tabs for Overview, Platforms, Technologies, Datasets, Demonstration Projects, Training, and Updates. Below the navigation bar, there's a sidebar with links for On This Page, Overview, AnVIL, BioData Catalyst, CRDC, Kids First, and NCBI. The main content area features a heading "Overview of Participating Platforms" and a section for "NHGRI AnVIL". It includes a paragraph about AnVIL's purpose and a link to its website. To the right, there's a sidebar with links for On This Page, NHGRI AnVIL, NCI Cancer Research Data Commons (CRDC), NH Common Fund - Kids First Data Resource Center, and National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM).

# NCPI Portal

## Technologies enabling science

The screenshot shows the 'Technologies' tab selected in the top navigation bar. The main content area is titled 'Interoperation Technologies'. It features a summary box stating: 'NCPI members are exploring / developing the following technologies in support of cloud platform interoperability.' Below this are two sections: 'Researcher Auth Service (RAS)' and 'Data Repository Service (GA4GH DRS)'. Each section contains a brief description and links to further documentation.

**Interoperation Technologies**

NCPI members are exploring / developing the following technologies in support of cloud platform interoperability.

**Researcher Auth Service (RAS)**

Researcher Auth Service (RAS) is an effort by the NIH's Center for Information Technology (CIT) to provide a common mechanism by which researchers can establish their identity and access data they are authorized to use across the systems outlined above. The RAS Application Programming Interface (API) allows seamless access to researchers for integrated data repositories.

Using RAS a researcher accessing NIH data resources can log in with their RRA Commons credentials and they would then be able to access any integrated repository without having to log in again. Existing rules for authorization will be enforced so a user can only access data that he or she has been authorized to view.

RAS uses open standards and protocols and provides integrating systems with many standards-based options for integration. RAS is part of the NIH CIT IAM General Support System (GSS) which is a Federal Information Security Management Act (FISMA) High system. As such, RAS adheres to NIST (National Institute of Standards and Technology) 800-53 and 800-57 guidelines pertaining to configuration management, least privilege, and cryptographic key establishment & management.

For detailed documentation of the RAS API see [Researcher Auth Service \(RAS\) Project Service Offerings](#).

**Data Repository Service (GA4GH DRS)**

GA4GH DRS. The Global Alliance for Genomics and Health (GA4GH) is an international coalition formed to enable the sharing of genomic and clinical data. The GA4GH Data Repository Service (DRS) provides a generic interface to data repositories so data consumers, including workflow systems, can access data objects in a single, standard way regardless of where they are stored and how they are managed.

The primary functionality of DRS is to map a logical ID to a means for physically retrieving the data represented by the ID. There are two styles of DRS URIs, Hostname-based and Compact Identifier-based, both using the drs://URI scheme. The API defines the characteristics of those IDs, the types of data supported, how they can be pointed to using URIs, and how clients can use these URIs to ultimately make successful DRS API requests.

For more information on the most recent version of this API (1.1) see the [Data Repository Service 1.1 Documentation](#).

## The science driving the tech

The screenshot shows the 'Demonstration Projects' tab selected in the top navigation bar. The main content area is titled 'NCPI Interoperability Demonstration Projects'. It features a summary box stating: 'The NCPI interoperability effort is guided by cross-platform demonstration projects which exercise specific scientific and technical use cases related to cloud-platform interoperability. Feedback from the projects is used to aid the discovery of detailed interoperability requirements and validate the utility of the developed features.' Below this are three project descriptions: 'Genetic Bases of Congenital Heart Defects (Goldmuntz)', 'LINE-1 Retrotransposon Expression (McKerrow)', and 'Sex as a Biological Variable (Wilson)'. Each project has a brief description and links to further details.

**NCPI Interop Demonstration Projects**

The NCPI interoperability effort is guided by cross-platform demonstration projects which exercise specific scientific and technical use cases related to cloud-platform interoperability. Feedback from the projects is used to aid the discovery of detailed interoperability requirements and validate the utility of the developed features.

The following demonstration projects are under development and will be updated with details on methods and results as they become available:

**Genetic Bases of Congenital Heart Defects (Goldmuntz)**

**Platforms** - NHLBI BioData Catalyst + Kids First DRC

In this research, we intend to study the genetic bases of congenital heart defects using variant and gene set analysis approaches, machine learning methods, amongst other statistical and genetic analysis models to help fill in the gaps that exist in the understanding of the etiology of CHDs. This will help the scientific community to better understand cardiogenesis and to better assess the risk of disease. Access to this whole-genome sequence data will facilitate our work. [Read more...](#)

**LINE-1 Retrotransposon Expression (McKerrow)**

**Platforms** - AnVIL + CRDC

This interoperability project aims to find a path to connect the GTEx data on the AnVIL platform to further processing and also combination with a prior analysis on the CRDC. This "normals" use case is a frequent request from our users, so finding a solution would be extremely valuable for a large number of cancer researchers. [Read more...](#)

**Genetic Bases of Congenital Heart Defects (Manning)**

**Platforms** - NHGRI AnVIL + Kids First DRC + NHLBI BioData Catalyst

In this project we investigate genetic factors related to congenital heart defects in a study design that uses healthy controls from two NHLBI cohorts and perform pooled analysis on AnVIL, powered by Terra. [Read more ...](#)

**Sex as a Biological Variable (Wilson)**

# NCPI Portal

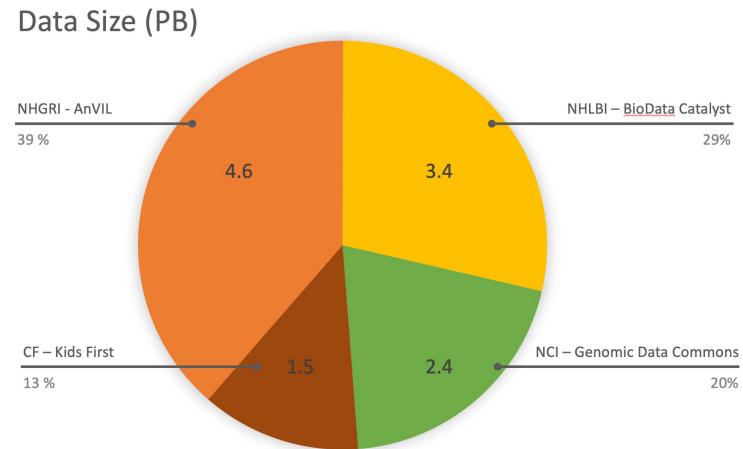
## Aggregating outreach resources

The screenshot shows the 'NCPI Training and Outreach' page. At the top, there's a sidebar with links to 'Overview', 'Train Your Colleague', and 'Cloud Cost Estimation'. The main content area has a heading 'NCPI Training and Outreach' with a sub-section 'A collection of training materials for NCPI resources'. Below this, a text block explains the purpose of the document. The page is divided into several sections: 'AnVIL' (with a detailed description of training resources for major components like Terra, Gen3, Galaxy, and Dockstore), 'Training Resources' (with links to documentation and video tutorials), 'User Support' (with a help desk link), 'Social links' (with links to Twitter, Slack, GitHub, and YouTube), and 'NHLBI BioData Catalyst' (with a brief description of its features). A sidebar on the right lists 'On This Page' items: AnVIL, NHLBI BioData Catalyst, CRDC, and Kids First.

## Past workshop resources

The screenshot shows the 'Past workshop resources' page. It features a sidebar with links to 'Overview', 'Platforms', 'Technologies', 'Datasets', 'Demonstration Projects', 'Training', and 'Updates'. The main content area has a heading 'Progress Updates' with a sub-section 'Progress Updates' explaining the purpose of workshops. Below this is a section for the '5th NCPI Workshop - October 5, 2021' with an 'Overview' section and a list of agenda items. Another section for the '4th NCPI Workshop - May 3rd, 2021' is also shown with its own overview and agenda. On the far right, a sidebar titled 'On This Page' lists past workshops: '5th NCPI Workshop - October 5, 2021', '4th NCPI Workshop - May 3rd, 2021', '3rd NCPI Workshop - October 30th, 2020', '2nd NCPI Workshop - April 16, 2020', and '1st NCPI Workshop - October 03, 2019'.

# 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)

The screenshot shows the NCPI Dataset Catalog search interface. At the top, there's a navigation bar with links for Overview, Platforms, Technologies, Datasets, Demonstration Projects, Training, and Updates. Below the navigation bar is a search bar with placeholder text "e.g. disease, study name, dbGaP Id". The main content area is divided into several sections:

- Platform:** A list of platforms: AnVIL, BDC, CRDC, KFDRC.
- Focus / Disease:** A list of diseases: Alzheimer Disease (45), Anemia, Sickle Cell (113), Arterial Pressure (28), Asthma (17), plus 59 more.
- Data Type:** A list of data types: Allele-Specific Expression (2), AMPLICON (10), Bisulfite-Seq (2), ChIP-Seq (17), plus 20 more.
- Study Design:** A list of study designs: Case Set (1), Case-Control (1), Clinical Trial (5), Control Set (3), plus 6 more.
- Consent Code:** A list of consent codes: ALZ (36), ALZ\_NPU (29), ARR (7), DS-AF-IRB-RD (3), plus 119 more.

Below these sections, there's a "Search Summary" table showing the number of studies and participants per platform:

Platform	Studies	Participants
AnVIL	45	312,933
BDC	113	438,041
CRDC	28	97,122
KFDRC	17	14,984
<b>Totals *</b>	<b>191</b>	<b>830,805</b>

At the bottom, there's a "Search Results" table showing two specific datasets:

Platform	Study	dbGap Id	Focus / Disease	Data Type	Study Design	Consent Code	Participants
AnVIL	A Genomic Atlas of Systemic Interindividual Epigenetic Variation in Humans (GTEx)	phs001746	Reference Values	Bisulfite-Seq	Control Set	GRU	194
AnVIL	Autism Sequencing Consortium (ASC)	phs000298	--	SNP/CNV Genotypes (NGS), WXS	Case-Control	DS-ASD, GRU, DS-AOND-MDS, HMB-MDS	12,772

Search by:

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

Buddied off into new  
Search Working Group

# Dockstore Organization for NCPI

Promoting FAIR practices in tool and workflow sharing

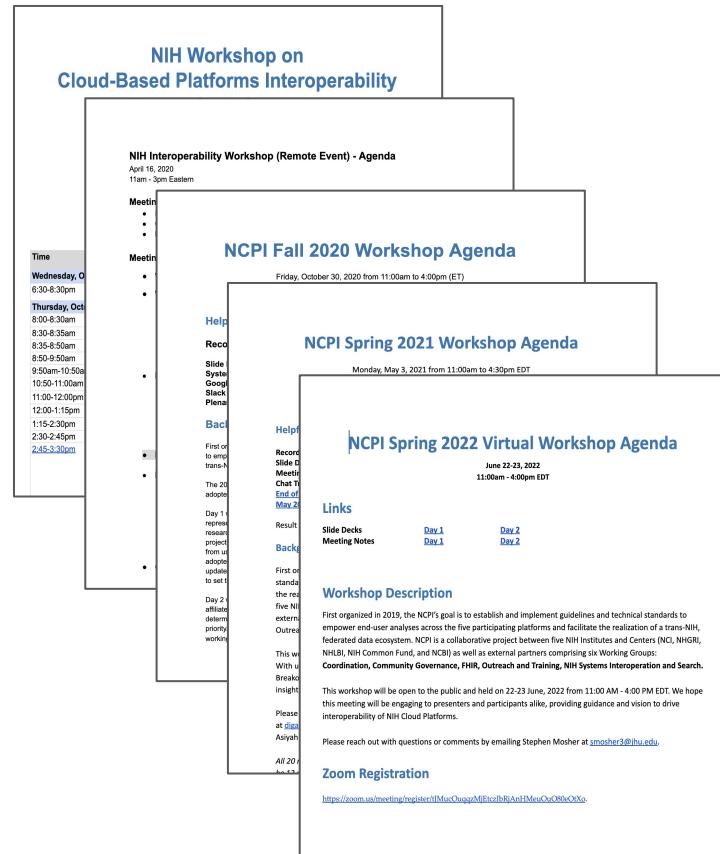
- Findable
- Accessible
- Interoperable
- Reusable

The screenshot shows a web browser window displaying the Dockstore Organization for NCPI. The URL in the address bar is <https://dockstore.org/organizations/NCPI>. The page features a dark blue header with the Dockstore logo, a search bar, and navigation links for 'Organizations', 'About', 'Docs', and 'Forum'. Below the header, a breadcrumb trail shows 'Organizations / NIH Cloud Platform Interoperability Effort'. The main content area has a title 'NIH Cloud Platform Interoperability Effort' with a blue cloud icon containing 'NCPI'. A sub-section titled 'About the Organization' includes a link to <https://anvilproject.org/ncpi>. Another section, 'About', describes the NIH Cloud Platform Interoperability Effort's goal of establishing guidelines and standards for cross-platform analysis. A 'Participating Platforms' section lists partner organizations: NHGRI AnVIL, NHLBI BioData Catalyst, Cancer Research Data Commons, Kids First Data Resource Center, and National Center for Biotechnology Information. The 'Interoperability Demonstration Projects' section notes six ongoing projects guided by cross-platform efforts. At the bottom, a note states there are currently six demonstration efforts.

<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

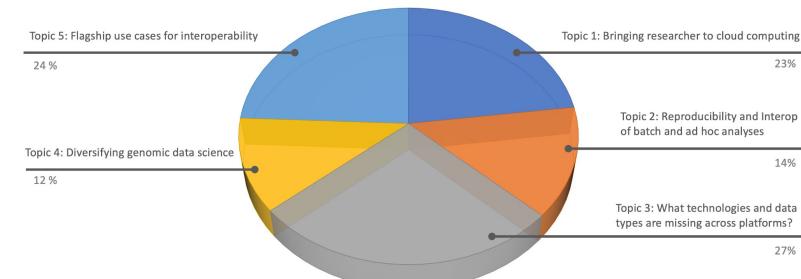
---

- 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](http://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

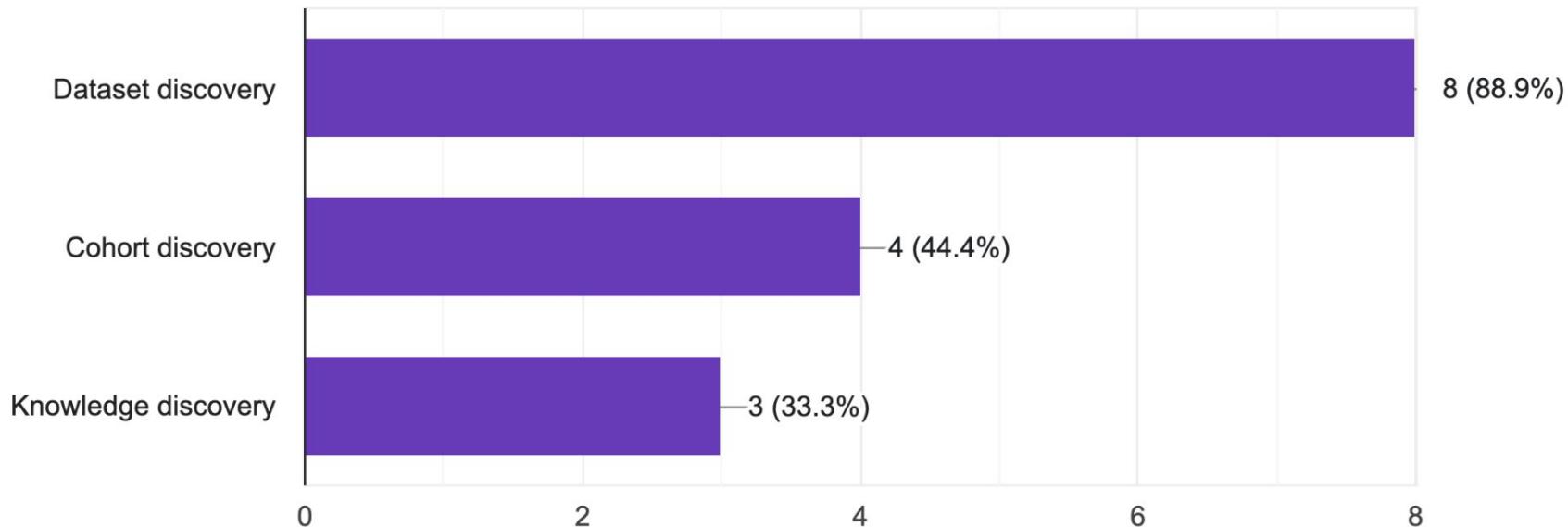
# Landscape Survey - Theme

---



What search theme is most relevant for your users?

9 responses

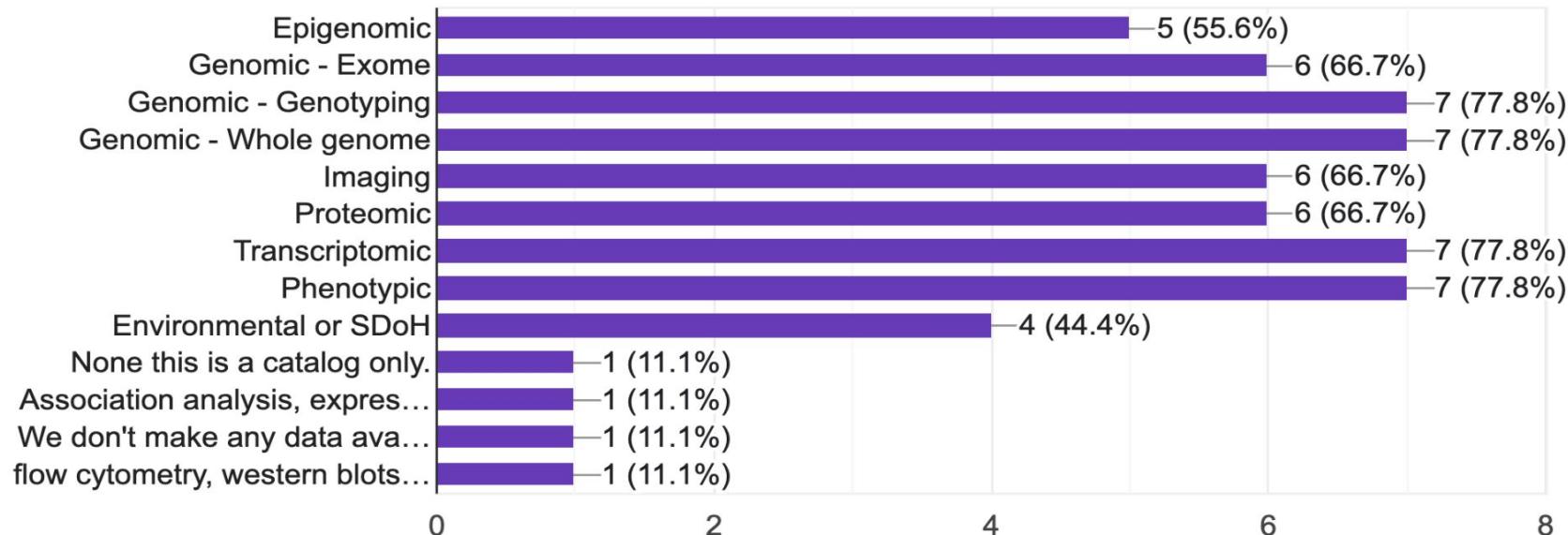


# Landscape Survey - Data Modalities

---

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
<i>HPO</i>	3
<i>MESH</i>	1
<i>PhenX</i>	1
<i>Follow dbGaP guide</i>	1
<i>Annotated w/ ontology ids</i>	1
<i>SNOMED</i>	1
<i>LOINC</i>	1
<i>NCIT</i>	1
<i>OMOP</i>	1
<i>CRDC Data Dictionaries/CRDC-H</i>	1

# Landscape Survey - Standards

---

## Genotype Standards

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

## Other Data Standards

<b>Standard</b>	<b>Responses</b>
<i>MIAME</i>	1
<i>Follow dbGaP guide</i>	1
<i>SRA</i>	1
<i>DUO</i>	1
<i>CRDC Data Dictionaries/CRDC-H</i>	1
<i>n/a or no response</i>	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 across 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

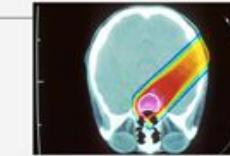
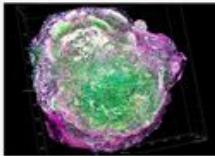


**The Texas Advanced Computing Center accelerates basic and applied cancer research to help save lives.**

*fighting*

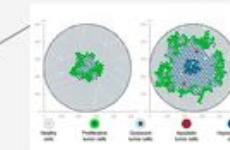
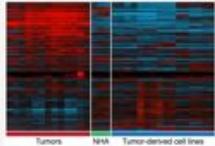
#### **Computer Modeling**

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



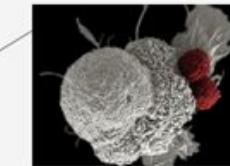
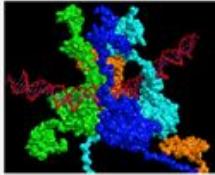
#### **Big Data Analysis**

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.

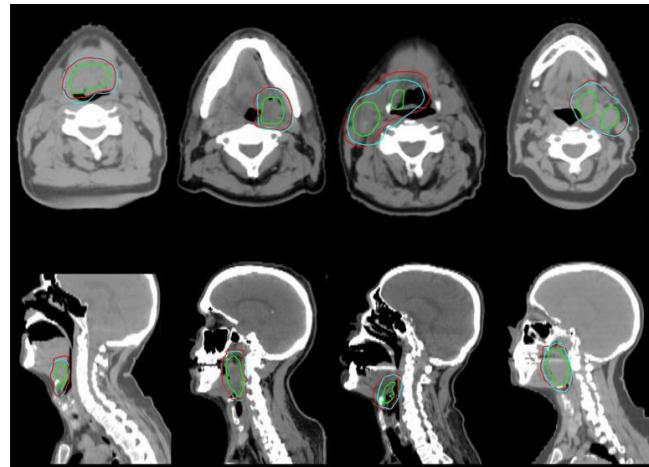
*— with supercomputers —*

# 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, contouring 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.

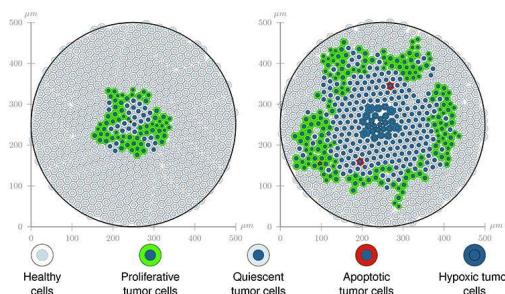
- **Approach:** Carlos Cardenas (MD Anderson) used Maverick to analyze data from 52 oropharyngeal cancer patients who had been treated at MD Anderson between January 2006 to August 2010, and had previously had their gross tumor volumes and clinical tumor volumes contoured for their radiation therapy 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.



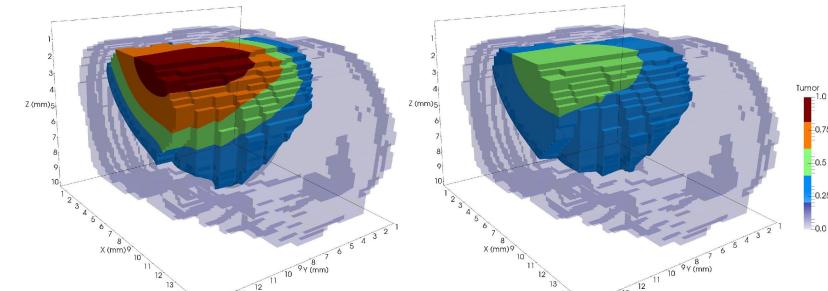
*Comparison between computer-predicted ground-truth clinical target volume (CTV1) (blue) and physician manual contours (red)*

# 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 *mathematize* 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.



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

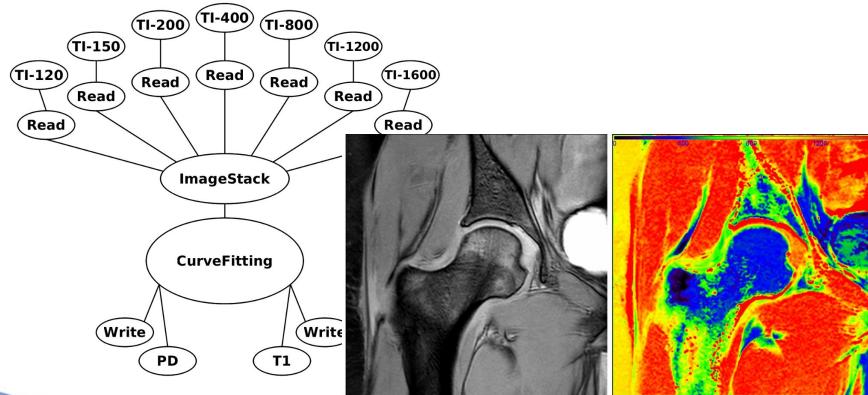


Model of tumor growth in a rat brain before radiation treatment (left) and after one session of radiotherapy (right)

- 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.

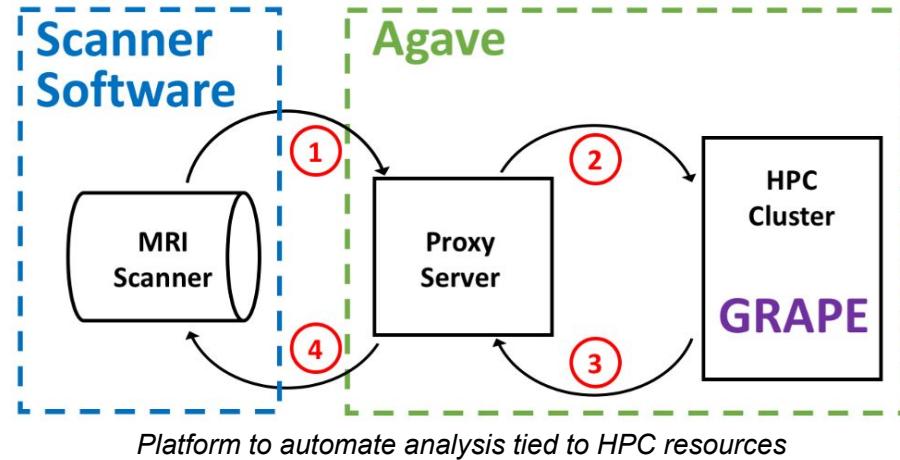
# TAPIS and Jetstream enabled automated, real-time, quantitative magnetic resonance imaging

- 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

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



- 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.

# 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 Listing

Refresh

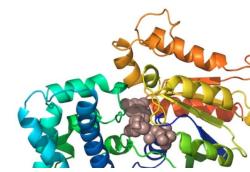
Job Name	Job Type	Job Status	Job Start Time	Job End Time	Actions
2018.09.07-test2	vina	FINISHED	7-Sep-2018 03:31 pm	7-Sep-2018 03:32 pm	
2018.09.07-test	vina	FINISHED	7-Sep-2018 03:11 pm	7-Sep-2018 03:11 pm	
2018.09.05.test	vina	FINISHED	5-Sep-2018 08:39 am	5-Sep-2018 08:39 am	
test-testset	vina	FINISHED	4-Sep-2018 12:46 pm	4-Sep-2018 12:47 pm	
test_small	vina	FINISHED	12-Sep-2017 01:37 pm	12-Sep-2017 03:34 pm	
test2	vina	FINISHED	12-Sep-2017 11:04 am	12-Sep-2017 11:06 am	
test3	vina	FINISHED	28-Jul-2017 10:20 pm	28-Jul-2017 10:20 pm	

Job outputs are available for download in a web interface



## Welcome to the new Virtual Drug Discovery Portal!

This Portal provides a graphical interface for conducting a screen for identifying small molecules that bind to your target protein.



The DrugDiscovery@TACC web portal

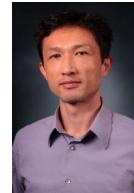
- Approach:** Dr. Stan Watowich (UTMB Galveston) partnered with researchers at TACC to provide an accessible and free virtual screening service called DrugDiscovery@TACC to investigators across the state of Texas and around the world.
- Result:** Users upload proteins of interest into a friendly web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

# Particle/Proton Therapy Translational Research Platform

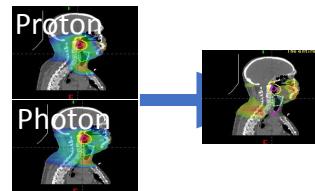
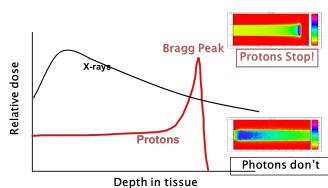
Xiaodong Zhang  
(MDACC)



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

# Acute to Chronic Pain Signatures

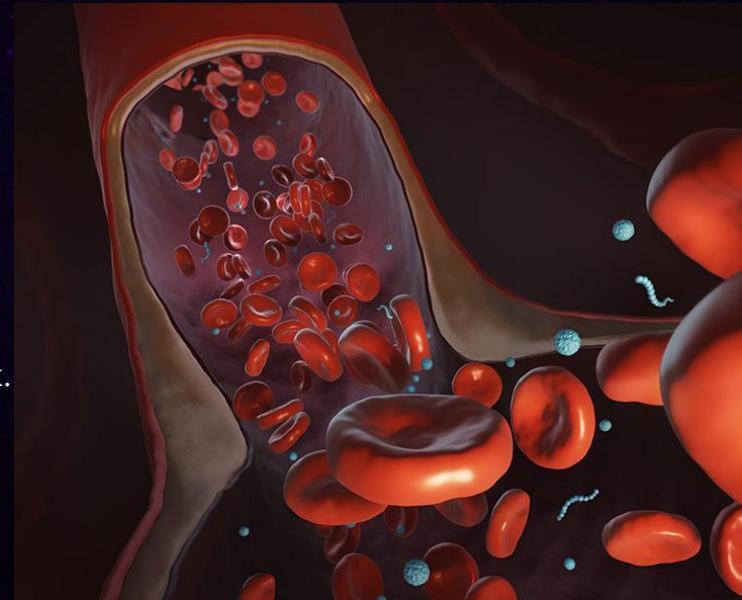
A bold research initiative to identify biomarkers and advance pain science

- 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



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

**Genomics Reporting Implementation Guide**  
2.1-SNAPSHOT - trial-use

[Home](#) [Table of Contents](#) [Background](#) [Artifact Index](#) [Support](#) [Quick Links](#) [Appendices](#)

**Table of Contents** [Home Page](#)

Genomics Reporting Implementation Guide, published by HL7 Clinical Genomics Working Group. This is not an author SNAPSHOT. This version is based on the current content of <https://github.com/HL7/genomics-reporting> and change

## Home Page

### 1 Scope

1.1 Scope

Genomics is a rapidly evolving area of healthcare that involves complex data structures. There is significant value in it's consistent, computable and that can accommodate ongoing evolution of medical science and practice. At present, they rely on data structures - what data should be present and how it should be organized. It does not address equested, created, approved, routed, delivered, amended, etc.

This guide covers many aspects of genomic data reporting, including:

- Representation of simple discrete variants, structural variants including copy number variants, complex variants as extra or missing chromosomes
- Representation of both known variants as well as fully describing de novo variations
- Germine and somatic variations
- Relevance of identified variations from the perspective of disease pathology, pharmacogenomics, transplant suitability
- Full and partial DNA sequencing, including whole genome and exome studies

### 1.2 How to Use this Guide

This implementation guide is organized into a set of sections. All implementers intending to do clinical genomic reporting sections. To understand the key profiles in this IG, as well as their relationship to one another, start with the should review the Understanding FHIR section below.

The remaining sections provide support for more specialized types of reporting. If your system is involved with genomics if the implementation guide for further guidance.

**Background** Introduces some of the key genomics terms and relationships that should be understood by overall guidance in using the profiles and translations defined in this guide. Guidance and report overall interpretations and how to report genotypes, haplotypes, and different types of variants. Guidance on expressing information about variants gleaned from various sequencing approaches, testing, etc.

**General Genomic Reporting** Guidance and examples related to genomic testing done for the purpose of assessing genetic risk for oncology and for general patient treatment.

**Variant Reporting** Guidance related to genomic testing done on somatic (non-germline) tissues, including as guidance related to genomic testing done for histocompatibility and immunogenomics as

**Somatic Genomic Reporting**

**Histocompatibility Reporting**

**Histocompatibility Reporting**

**Table of Contents** [General Genomic Reporting](#)

Genomics Reporting Implementation Guide, published by HL7 Clinical Genomics Working Group. This is SNAPSHOT. This version is based on the current content of <https://github.com/HL7/genomics-reporting>

## 3 General Genomic Reporting

This page defines the core profiles and concepts that would be expected to be present in most genomic reports relate to each other. Concepts covered include the genomics report itself and the high-level categories of the report, such as patient, specimen, variants, haplotypes, genotypes, etc.

This table describes the categories of data contained in this implementation guide.

<b>Genomics Report</b>	Groups together all the structured data being reported for a genomic testing.
<b>Overall Interpretations</b>	Reported when variant analysis (sequencing or targeted variants) is done. Provide reported.
<b>Genomic Findings</b>	These are observations about the specimen's genomic characteristics. For example haplotype, or variant that was detected.
<b>Genomic Implications</b>	These represent observations where the <code>Observation.subject</code> is typically the Patient refer to Genomic Findings. For example, "Patient may have increased susceptibility"
<b>Region Studied</b>	These are observations describing the region or regions that were studied as part of the tests other than sequenced genomic variants may also be included.
<b>Other Observations</b>	Specific actions be taken, such as genomic counseling, re-testing, adjusting drug dosages.
<b>Recommended Actions</b>	Other resources that provide contextual details.
<b>Contextual Resources</b>	

**19.0.3 Structures: Resource Profiles**

These define constraints on FHIR resources for systems conforming to this implementation guide

<b>Diagnostic Implication</b>	Observation stating a linkage between one or more genotype/phenotype condition, or cancer diagnosis.
<b>Followup Recommendation</b>	Task describing the follow-up that is recommended.
<b>Genomics DocumentReference</b>	A profile of DocumentReference used to represent a genomics document.
<b>Genomics Report</b>	Genomics profile of DiagnosticReport.
<b>Genotype</b>	Assertion of a particular genotype on the basis of one or more variants.
<b>Haplotype</b>	Assertion of a particular haplotype on the basis of one or more variants.
<b>Microsatellite Instability</b>	Microsatellite Instability (MSI) is the condition of genetic hyper-repair (MMR).
<b>Medication Recommendation</b>	Task proposing medication recommendations based on genetic variants.
<b>Overall Interpretation</b>	Provides a coarse overall interpretation of the genomic results.
<b>Region Studied</b>	The Region Studied profile is used to assert actual regions/study coverage areas (e.g. due to technical limitations during test performance).
<b>Sequence Phase Relationship</b>	Indicates whether two entities are in <i>Cis</i> (same strand) or <i>Trans</i> (opposite strand).
<b>Tumor Mutation Burden</b>	The total number of mutations (changes) found in the DNA of a tumor sample. For example, tumors that have a high number of mutations. Tumor mutational burden is being used as a type of biomarker.
<b>Therapeutic Implication</b>	Profile with properties for observations that convey the potential therapeutic benefit of a variant.
<b>Variant</b>	Details about a set of changes in the tested sample compared to a reference genome.

**4.2 Defining Variants**

This Implementation Guide supports two reporting patterns for defining variants:

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

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

Additional resources that implementers may want to leverage when reporting variant information are relationships among human variations and phenotypes, and NCBI's Variation Services.<sup>1</sup> that relates to this.

**4.2.1 Variants Defined by a Nomenclature Statement**

This pattern describes the observed nucleotide sequence or configuration using HGVS or TSCN name properly distinguish variants with the degree of precision needed for clinical use. Note that synonymous nomenclature may be required.

<b>Defining Component</b>	<b>Example Value</b>
genomic-coordinate-report (LOINC 81290-9) OR coding-hgvs (LOINC 48004-6)	{ "system" : "http://varnomen.hgv.org/", "code" : "NM_022787.3..c..769cA*" }
cycloameric-nomenclature (LOINC 81291-7)	{ "system" : "http://www.ncbi.nlm.nih.gov/muccore/", "code" : "46..XX..C9..22344*4*" }

**4.2.2 Variants defined by multiple components (VCF-like)**

This representation leverages multiple component details to communicate an allele within the context of its definition representation in FHIR, but is limited to variations with known breakpoints, and aliases specific genome build and chromosome identifiers rather than explicit reference sequences. Build a

<b>Defining Component</b>	<b>Example Value</b>
genomic-ref-seq (LOINC 48013-7)	{ "system" : "http://www.ncbi.nlm.nih.gov/muccore/", "code" : "NC_000019.10" }

**19.0.5 Structures: Extension Definitions**

These define constraints on FHIR data types for systems conforming to this implementation guide

<b>Annotation Code</b>	Specifies the content of an Annotation.
<b>Genomic Report Note</b>	Adds codified notes to a report to capture additional content.
<b>Genomics Artifact</b>	Captures citations, evidence and other supporting documentation for the observation or report.
<b>Genomics File</b>	Used to transmit the contents of or links to files that were produced as part of the test or similar files.
<b>Genomics Risk Assessment</b>	RiskAssessment delivered as part of a genomics report or observation.
<b>Medication Assessed</b>	Used to reference a specific medication that was assessed (e.g. a FHIR Medication or a FHIR MedicationStatement).
<b>Recommended Action</b>	References a proposed action that is recommended based on the results of the diagnostic test.
<b>Therapy Assessed</b>	Used to reference a specific therapy that was assessed (e.g. a FHIR ResearchStudy, a FHIR Medication or a FHIR MedicationStatement).

**19.0.7 Terminology: Code Systems**

These define new code systems used by systems conforming to this implementation guide

<b>CinVar Evidence Level Example Codes</b>	CinVar contains examples of evidence level codes. <a href="https://www.ncbi.nlm.nih.gov/cinvar/">https://www.ncbi.nlm.nih.gov/cinvar/</a>
<b>Coded Annotation Type Codes</b>	Code System for specific types of annotations.
<b>PharmGKB Evidence Level Example Codes</b>	PharmGKB contains examples of evidence level codes. <a href="https://www.pharmgkb.org/page/evidence-level-codes">https://www.pharmgkb.org/page/evidence-level-codes</a>
<b>Sequence Phase Relationship Codes</b>	Code System for specific types of sequence phase relationship codes.
<b>To Be Determined Codes</b>	These codes are currently 'TBD' or待定.
<b>Variant Confidence Status Codes</b>	A code that represents the confidence status of a variant.

**19.0.8 Terminology: Value Sets**

These define sets of codes used by systems conforming to this implementation guide

<b>Coded Annotation Types</b>	Value Set for specific types of coded annotations.
<b>Condition Inheritance Patterns</b>	Value Set for specific inheritance patterns of a condition in a pedigree.
<b>DNA Change Type</b>	DNA Change Type of a variant.
<b>Evidence Level Examples</b>	Example sources of values for Evidence Level.
<b>Genetic Therapeutic Implications</b>	The effect of a variant on downstream biological products or pathways.
<b>Genomic Gene Nomenclature</b>	Value Set for terms that describe a predicted nomenclature based on the presence of a gene.
<b>Genomic Gene Names (HGNC)</b>	This value set includes all HGNC entries, which includes multiple code systems. See <a href="https://www.ncbi.nlm.nih.gov/gene/">https://www.ncbi.nlm.nih.gov/gene/</a> for more details.

**Diagram of Genomics Report Structure**

```

graph TD
    GR[Genomics Report (DiagnosticReport)] -- "extension(extension/recommendedAction)(0..*)" --> RA[Recommended Action / Medication Recommendation (Task)]
    RA -- "result(0..*)" --> OI[Overall Interpretation (Observation)]
    RA -- "result(0..*)" --> GI[Genomic Implications: Diagnostic or Therapeutic (Observation)]
    RA -- "result(0..*)" --> GF[Genomic Findings: Variants, Haplotypes, and Genotypes (Observation)]
    RA -- "result(0..*)" --> OR[Other genomic results: Region Studied, Sequence Phase Relations, TMB/MSI/etc (Observation)]
    RA -- "result(0..*)" --> ON[Other non-genomic results: chemical, protein, karyotype results (Observation)]
  
```

# 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



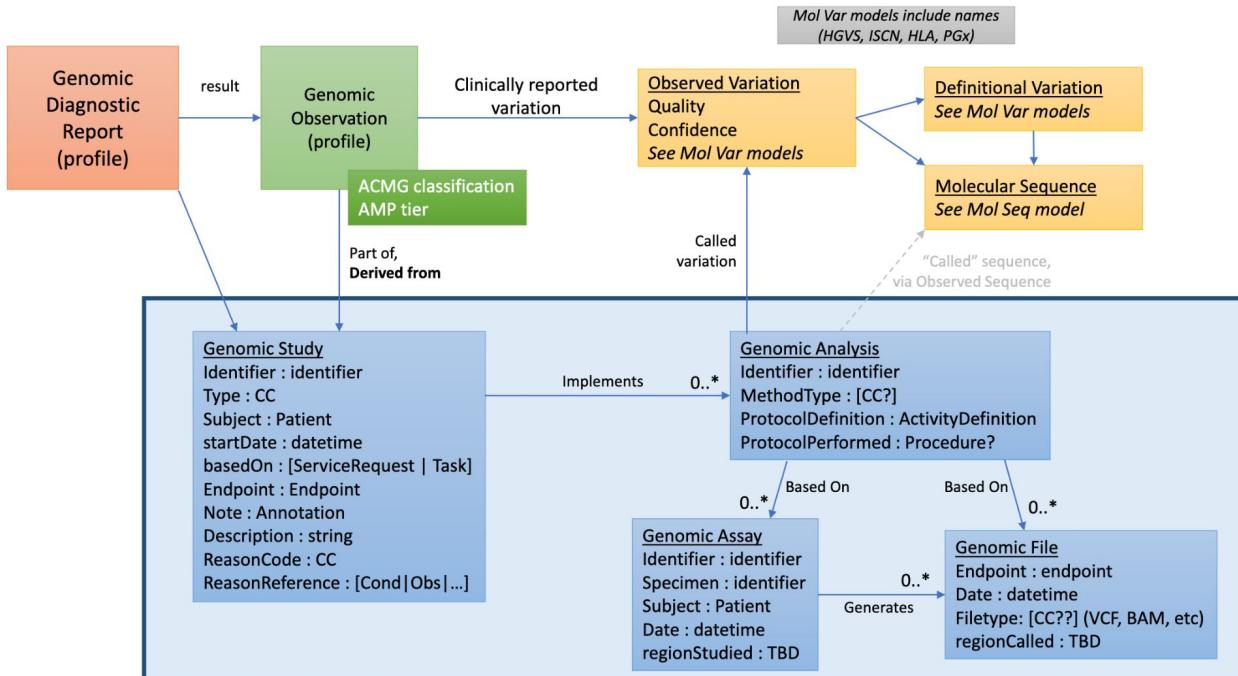
# Genomics FHIR Initiatives

---

- 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.

# Challenges, and the path forward



# Challenges, and the path forward

---

## 1. Learning Curve

The publication of FHIR DSTU2 included the creation of the FHIR Maturity Model (FMM).

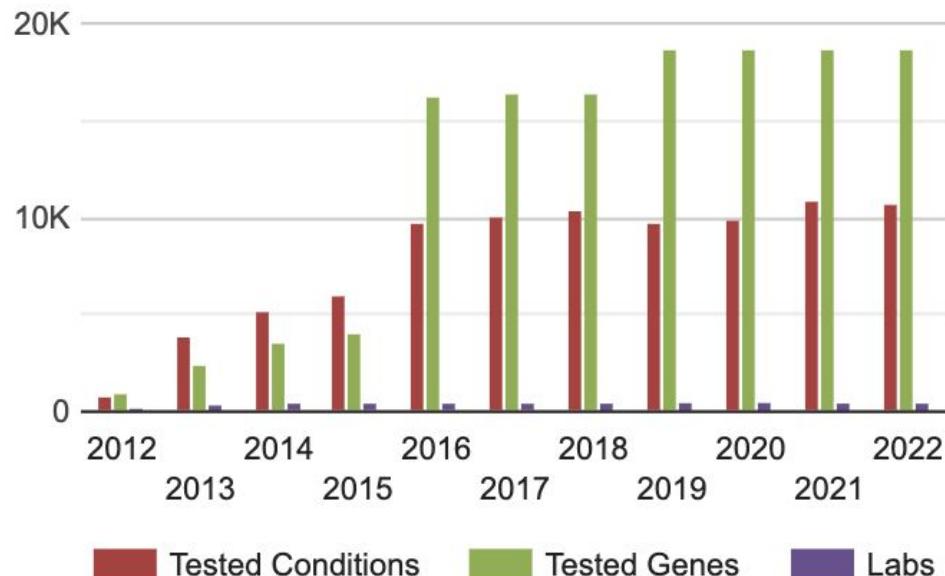
## 2. Ease of GTR Data

When new Resources are created, they are not

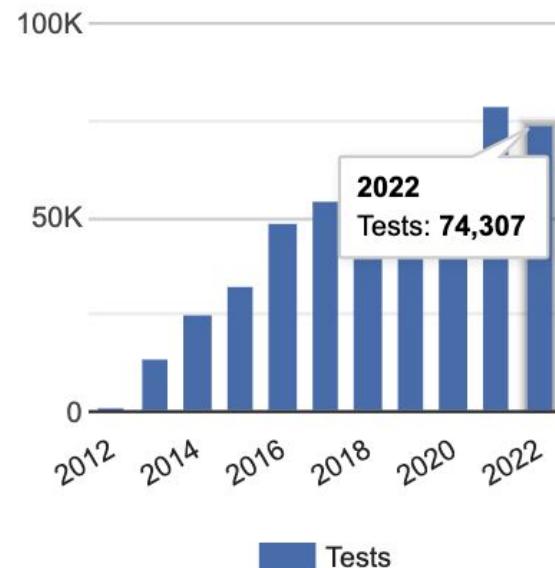
## 3. Mul

## 4. Diver

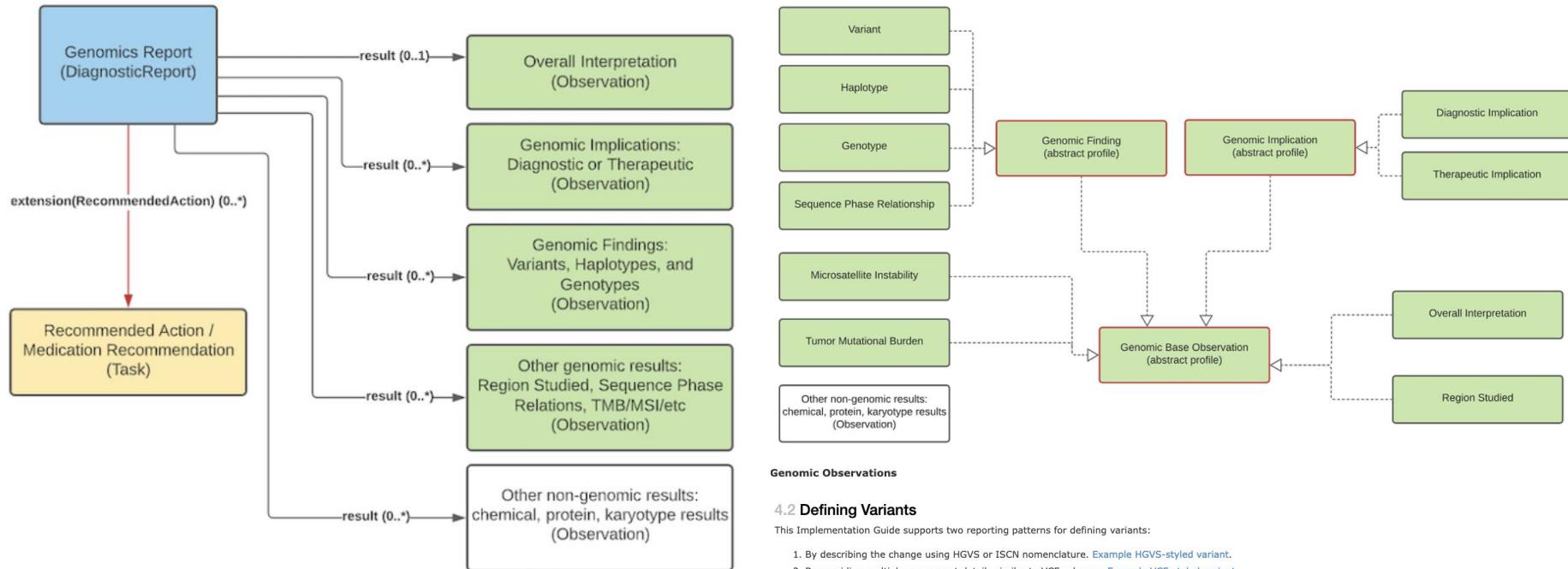
## 5. Adc



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



# 1. Clinical Genomics IG Learning Curve



## Genomic Report Overview

### 4.2 Defining Variants

This Implementation Guide supports two reporting patterns for defining variants:

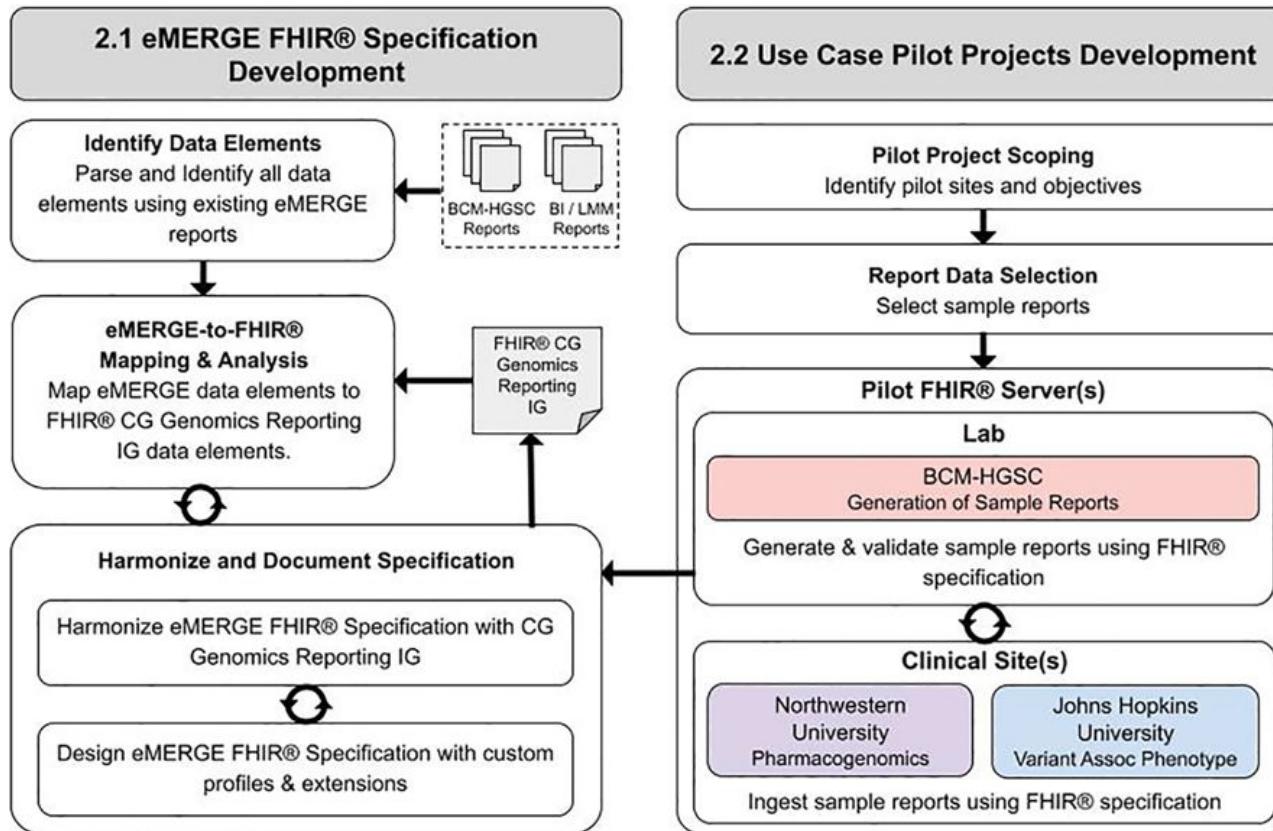
1. By describing the change using HGVS or ISCN nomenclature. [Example HGVS-styled variant](#).
2. By providing multiple component details similar to VCF columns. [Example VCF-styled variant](#).

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.



## 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.

- |  |   |
|--|---|
| <p>4. Repi<br/>5. An c<br/>gGM</p> <ul style="list-style-type: none"><li>● <a href="#">Gen</a></li><li>● <a href="#">FHI</a></li></ul> | <ol style="list-style-type: none"><li>1. Completed Major<ol style="list-style-type: none"><li>1. Composite Report - Section Grouping</li><li>2. Lab Defined Tests - Methodology, References, etc...<br/><u>(PlanDefinition)</u></li><li>3. Report Level Comments - Observation</li><li>4. Recommendations (Proposed) -<br/><u>(RecommendedAction - Task)</u></li><li>5. <u>Nested &amp; Indirect Result Referencing - hasMembers &amp; derivedFrom?</u></li><li>6. <u>Addition of chromosome to Variant</u></li></ol></li><li>2. Completed Minor<ol style="list-style-type: none"><li>1. <u>New Identifier Type Code(s)</u></li><li>2. <u>InhDisPath phenotype cardinality change</u></li><li>3. <u>InhDisPath value (CC) made extensible</u></li><li>4. <u>DR category cardinality changed to 0..*</u></li></ol></li></ol> |
|--|---|
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

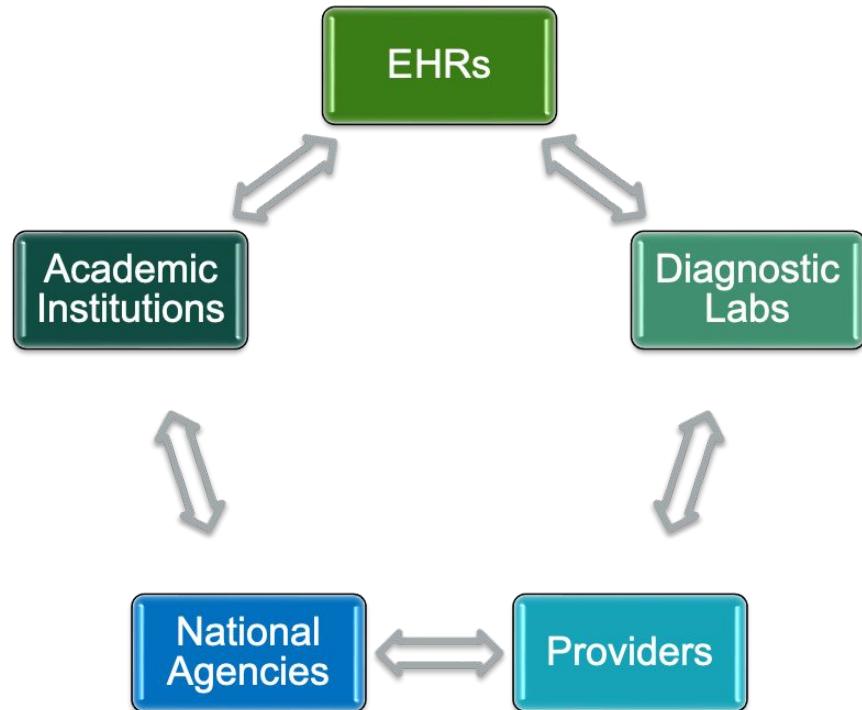
---

- 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)

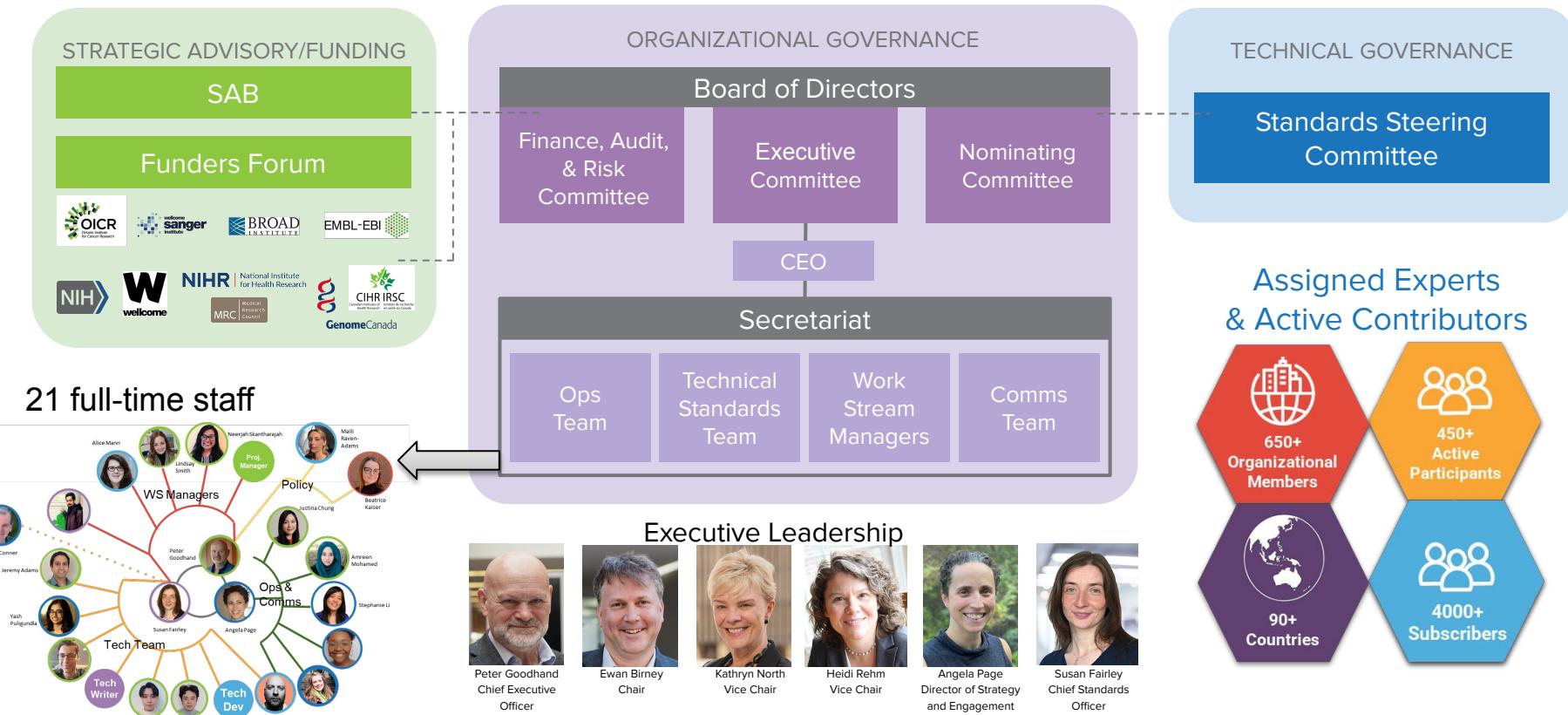
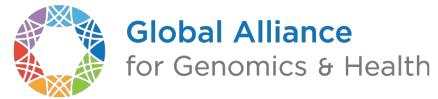
# The Global Alliance for Genomics and Health Mission...

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

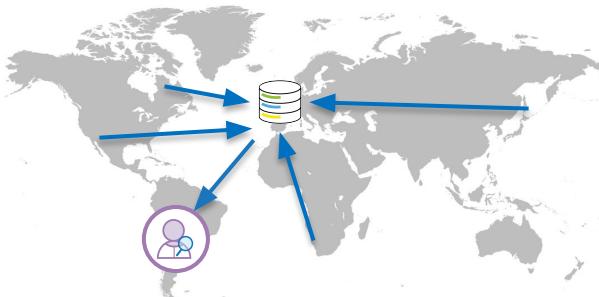
# GA4GH Organization Structure



# Different Approaches to Data Sharing

## Central Database

Genomic Knowledgebase

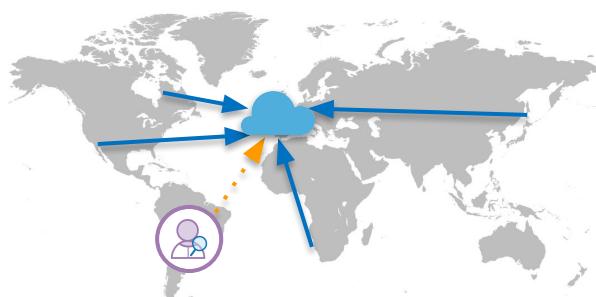


Aggregate data globally

Download, analyze locally

## Secure Cloud

Large scale research datasets

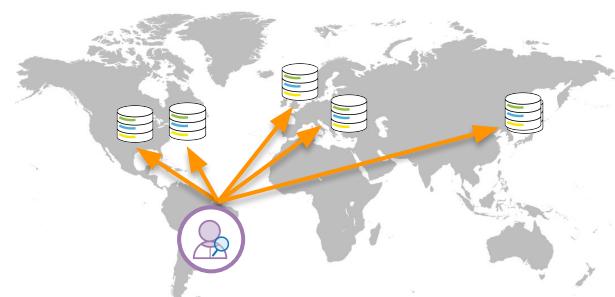


Aggregate data globally

Analyze centrally in secure cloud

## Federation

Connecting national genomics initiatives



Host data locally

Visit data remotely and collate results

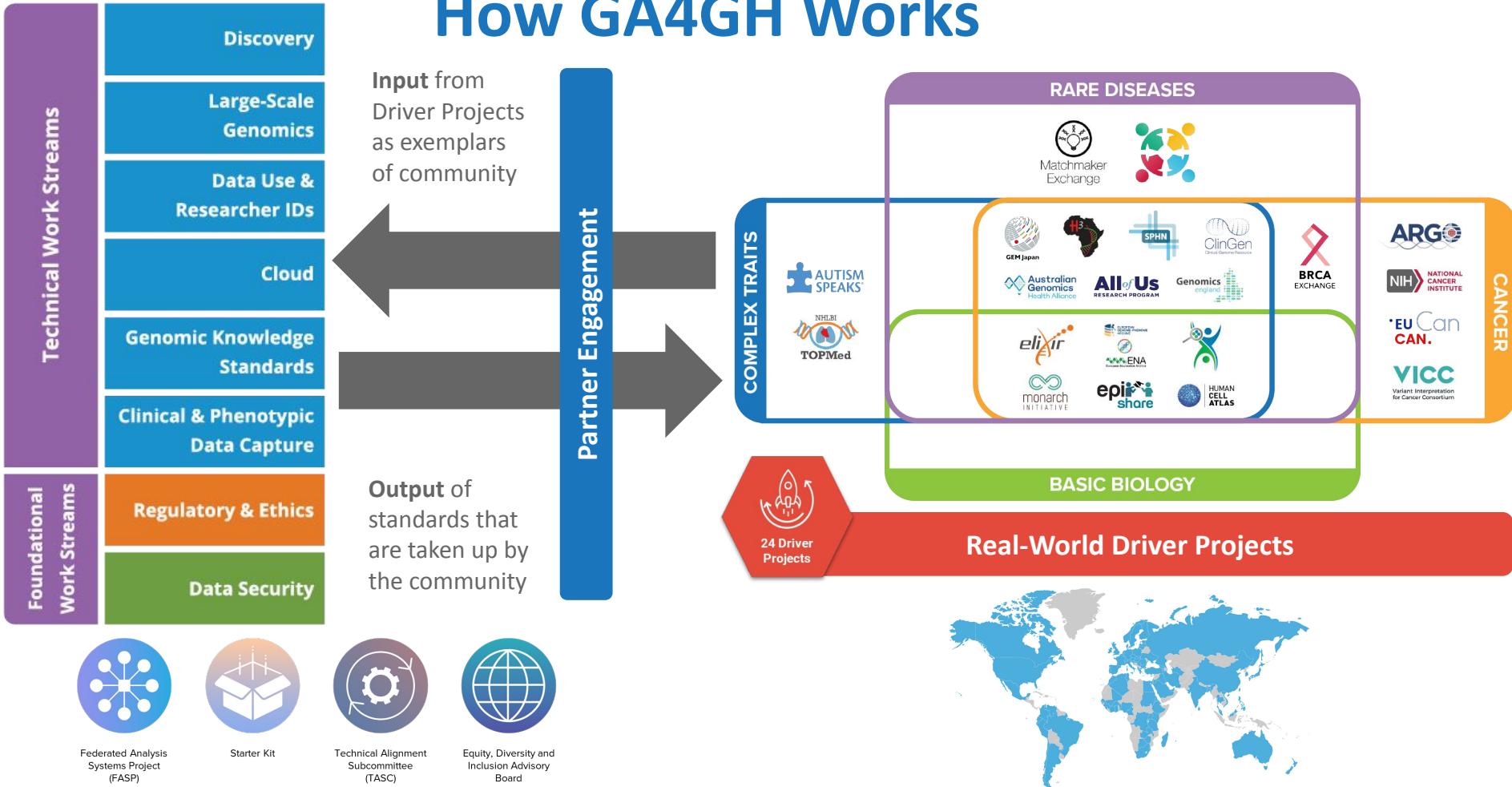


User

→ Data transmission

→ Secure access

# How GA4GH Works



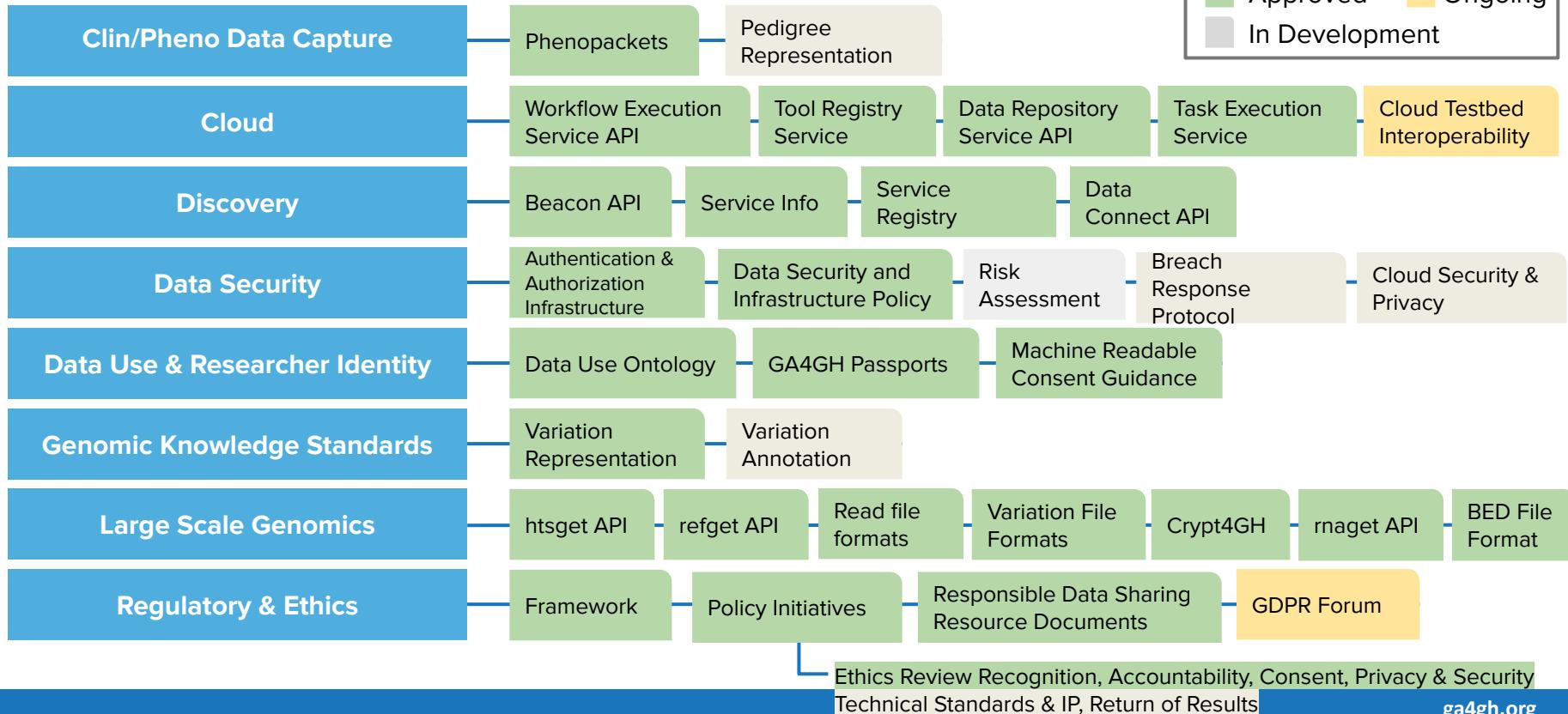
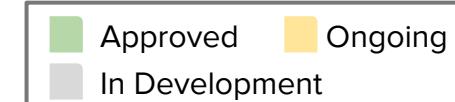
Federated Analysis  
Systems Project  
(FASP)

Starter Kit

Technical Alignment  
Subcommittee  
(TASC)

Equity, Diversity and  
Inclusion Advisory  
Board

# GA4GH 2020-2022 Strategic Roadmap



# 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



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

Genotypic Data  
Gene D

Genomic Matchmaker

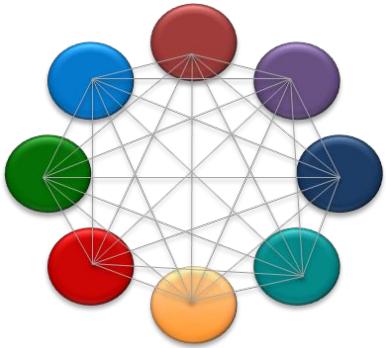
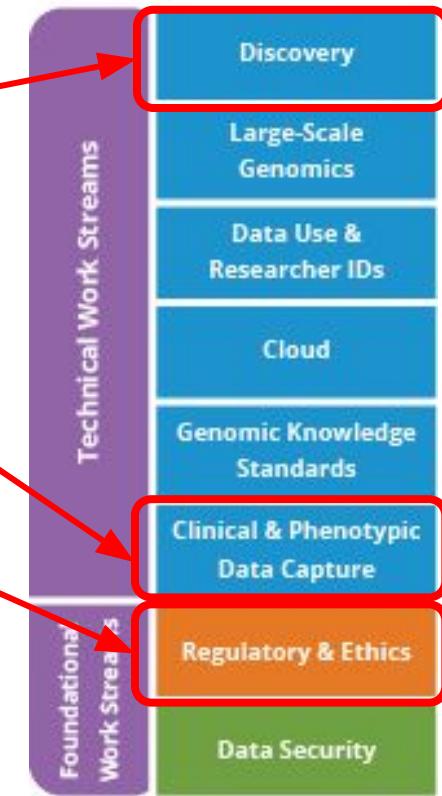
Genotypic Data  
Gene D

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

# Developing the MME Federated Network using GA4GH Standards

## Use of GA4GH standards:

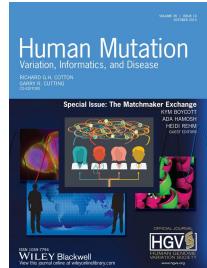
- API for data exchange
  - ID (Mandatory) +/- Label
  - Submitter (Mandatory)
  - Phenotypic Features and/or Gene Names (Mandatory)
  - Disorders (Optional) - OMIM or OrphaNet
  - Sex, Age of Onset, Inheritance (Optional)
- Clinical and phenotypic data capture standards
- Consent framework for data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

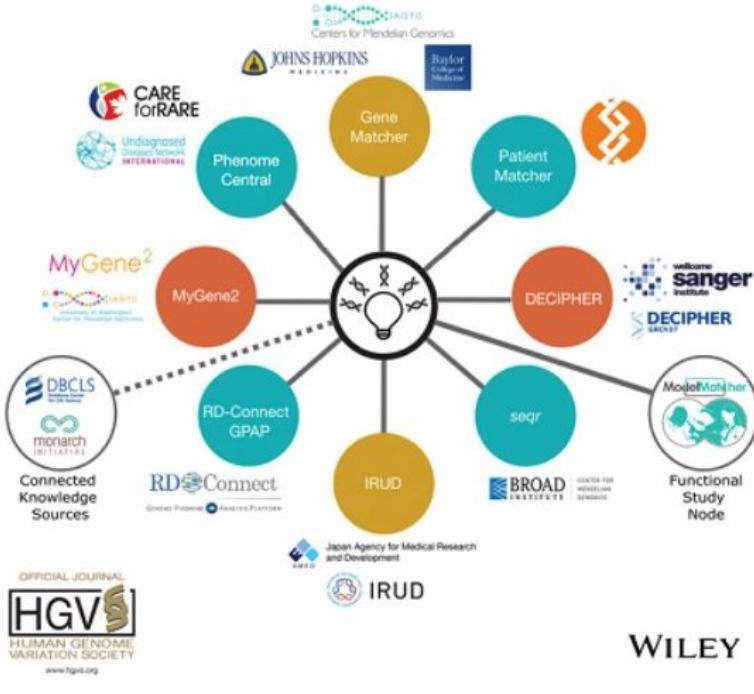
# Human Mutation

## Variation, Informatics, and Disease

GARRY R. CUTTING, EDITOR

## Special Issue: Matchmaker Exchange: Seven years of discovery and collaboration

Guest Editors: Kym Boycott, Ada Hamosh, and Heidi Rehm



## 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 disease

- DECIPHER:** Supporting the interpretation of variant data to advance diagnosis and research

- seqr:** A web-based analysis and collaboration tool

- PatientMatcher:** A customizable Python library for rare disease patients via the Matchmaker Exchange

- The **RD-Connect Genome-Phenome Analysis Platform** for gene discovery for rare diseases

- Advances in the development of **PubCrawl**, an 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

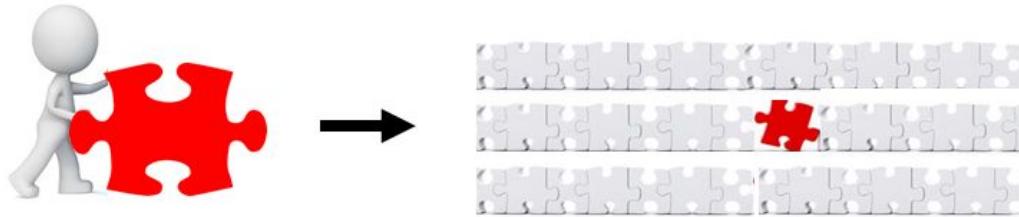
Three clinical labs had 1040/3819 (27%) gene discoveries validated through MME

### (a) Two-sided matchmaking



# Matchmaker Exchange

### (b) One-sided matchmaking



### (c) Zero-sided matchmaking



# VariantMatcher

VariantMatcher is a database open to search on genomic locations. It harbors genomic data as part of the BHCMG.

Email :

Password :

## 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](mailto:genetika@genetika.com.br) - PUC Brazil

**Bilateral Cleft**

BHXXXX - Patient - Affected - 6:34004293:T>C

Hamza Aziz - [haziz2@jhmi.edu](mailto:haziz2@jhmi.edu) - JHU

**Bicuspid Aortic valve, Aneurysm, ascending aortic**

BHXXXX - Patient - Affected - 6:34004293:T>C

Samantha Penney - [penney@bcm.edu](mailto:penney@bcm.edu) - Baylor College of Medicine

**Encephalopathy, Ataxia, Hypotonia**

BHXXXX - Patient - Affected - 6:34004293:T>C

Samantha Penney - [penney@bcm.edu](mailto: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](mailto:fvairo@hcpa.edu.br) - Hospital de Clinicas de Porto Alegre

BHXXXX - Father - 6:34004293:T>C

Daryl Scott - [dscott@bcm.edu](mailto:dscott@bcm.edu) - Baylor College of Medicine

BHXXXX - Mother - 6:34004293:T>C

Samantha Penney - [penney@bcm.edu](mailto:penney@bcm.edu) - Baylor College of Medicine

BHXXXX - Father - 6:34004293:T>C

Samantha Penney - [penney@bcm.edu](mailto: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](mailto:variantmatcher@jhmi.edu) or use the [contact form](#).

# Human Mutation

Variation, Informatics, and Disease

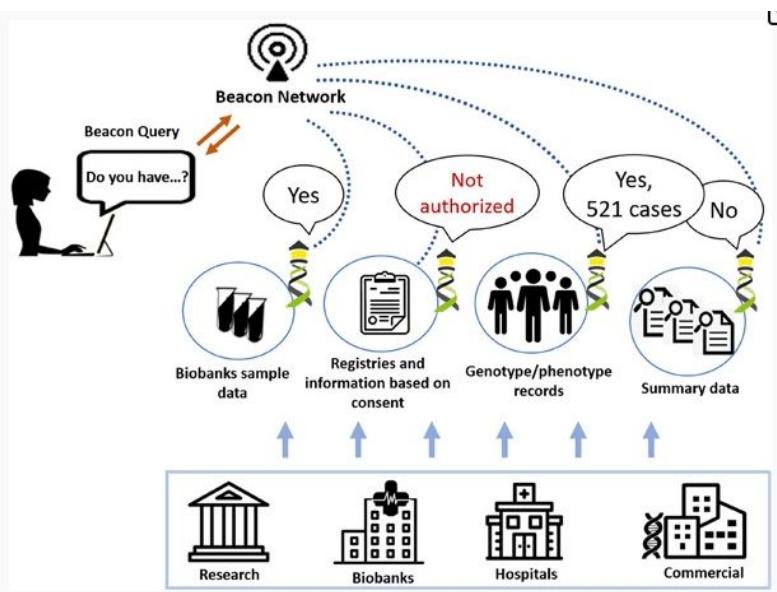


INFORMATICS | Open Access | CC BY SA

## 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



DATABASES | Open Access | CC BY SA

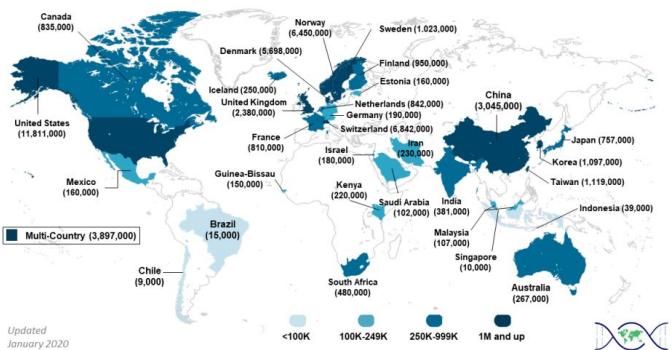
## Variant-level matching for diagnosis and discovery: Challenges and opportunities

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 ↴

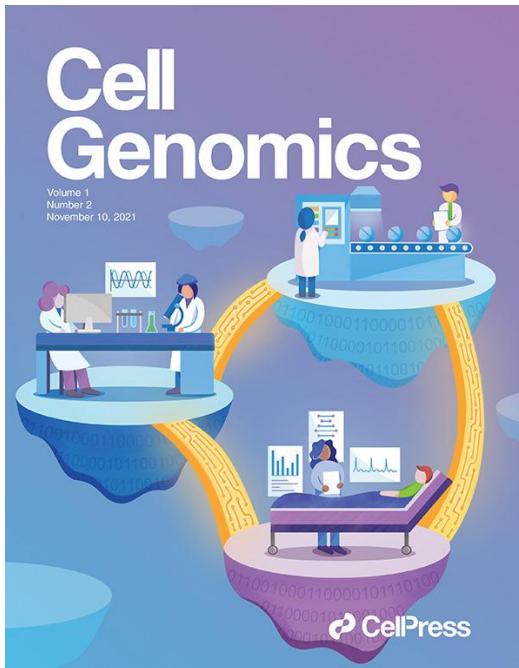
First published: 22 February 2022 | <https://doi.org/10.1002/humu.24359> | Citations: 1

MyGene2, Geno2MP, VariantMatcher, Franklin

### IHCC Member Cohorts across the World



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



<https://www.cell.com/cell-genomics>

## Cell Genomics

CellPress  
OPEN ACCESS

### Perspective

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

Heidi L. Rehm,<sup>1,2,47</sup> Angela J.H. Page,<sup>1,3,\*</sup> Lindsay Smith,<sup>3,4</sup> Jeremy B. Adams,<sup>3,4</sup> Gil Alterovitz,<sup>5,47</sup> Lawrence J. Babb,<sup>1</sup> Maximilian P. Barkley,<sup>6</sup> Michael Baudis,<sup>7,8</sup> Michael J.S. Beauvais,<sup>3,9</sup> Tim Beck,<sup>10</sup> Jacques S. Beckmann,<sup>11</sup> Sergi Beltran,<sup>12,13,14</sup> David Bernick,<sup>1</sup> Alexander Bernier,<sup>9</sup> James K. Bonfield,<sup>15</sup> Tiffany F. Boughtwood,<sup>16,17</sup> Guillaume Bourque,<sup>9,18</sup> Sarior R. Bowers,<sup>19</sup> Anthony J. Brookes,<sup>10</sup> Michael Brudno,<sup>18,19,20,21,38</sup> Matthew H. Brush,<sup>22</sup> David Bujold,<sup>9,18,38</sup> Tony Burdett,<sup>23</sup> Orion J. Buske,<sup>24</sup> Moran N. Cabili,<sup>1</sup> Daniel L. Cameron,<sup>25,26</sup> Robert J. Carroll,<sup>27</sup> Esmeralda Casas-Silva,<sup>123</sup> Debyani Chakravarty,<sup>29</sup> Bimal P. Chaudhari,<sup>30,31</sup> Shu Hui Chen,<sup>32</sup> J. Michael Cherry,<sup>33</sup> Justina Chung,<sup>3,4</sup> Melissa Cline,<sup>34</sup> Hayley L. Clissold,<sup>19</sup> Robert M. Cook-Deegan,<sup>35</sup> Mélanie Courtoot,<sup>23</sup> Fiona Cunningham,<sup>23</sup> Miro Cupak,<sup>6</sup> Robert M. Davies,<sup>36</sup> Danielle Denisko,<sup>19</sup> Megan J. Doerr,<sup>31</sup> Lena I. Dolman,<sup>19</sup> Edward S. Dove,<sup>38</sup> L. Jonathan Dursi,<sup>20,39</sup> Stephanie O.M. Dyke,<sup>8</sup> James A. Eddy,<sup>37</sup> Karen Elbbeck,<sup>40</sup> Kyle P. Elliott,<sup>22</sup> Susan Fairley,<sup>3,23</sup> Khalid A. Fakhro,<sup>41,42</sup> Helen V. Firth,<sup>15,43</sup> Michael S. Fitzsimons,<sup>44</sup> Marc Fiume,<sup>8</sup> Paul Fleck,<sup>23</sup> Ian M. Fore,<sup>23</sup> Mallory A. Freeberg,<sup>23</sup> Robert R. Freimuth,<sup>45</sup> Lauren A. Fromont,<sup>52</sup> Jonathan Fuert,<sup>4</sup> Clara L. Gaff,<sup>18,17</sup> Weiniu Gan,<sup>33</sup> Elena M. Ghanaim,<sup>46</sup> David Glazer,<sup>47</sup> Robert C. Green,<sup>1,48,49</sup> Malachi Griffith,<sup>50</sup> Obi L. Griffith,<sup>50</sup> Robert L. Grossman,<sup>44</sup> Tudor Groza,<sup>51</sup> Jaime M. Guidry Avil,<sup>29</sup> Roderic Guigó,<sup>13,52</sup> Dipayan Gupta,<sup>25</sup> Melissa A. Haendel,<sup>53</sup> Adia Hamosh,<sup>54</sup> David P. Hansen,<sup>18,83</sup> Reece K. Hart,<sup>1,100,124</sup> Dean Mitchell Hartley,<sup>55</sup> David Haussler,<sup>35</sup> Rachelle M. Hendricks-Sturup,<sup>56</sup> Calvin W.L. Ho,<sup>57</sup> Ashley E. Hobbs,<sup>8</sup> Michael M. Hoffman,<sup>19,20,21</sup> Oliver M. Hofmann,<sup>26</sup> Petr Holub,<sup>58,59</sup> Jacob Shujui Hsu,<sup>60</sup> Jean-Pierre Hubaux,<sup>61</sup> Sarah E. Hunt,<sup>23</sup> Ammar Husami,<sup>62</sup> Julius O. Jacobsen,<sup>63</sup> Samuya S. Jamuar,<sup>64,65</sup> Elizabeth L. Janes,<sup>3,66</sup> Francis Jeanson,<sup>128</sup> Aina Jene,<sup>32</sup> Amber L. Johns,<sup>37,68</sup> Yann Joly,<sup>71</sup> Steven J.M. Jones,<sup>20</sup> Alexander Kanitz,<sup>8,70</sup> Kazuto Kato,<sup>71</sup> Thomas M. Keane,<sup>23,72</sup> Kristina Kekesi-Lafraunce,<sup>3,9</sup> Jerome Kalleher,<sup>73</sup> Giselle Kerr,<sup>23</sup> Seik-Soon Khor,<sup>74,75</sup> Bartha M. Knoppers,<sup>9</sup> Melissa A. Konopko,<sup>76</sup> Kenjiro Kosaki,<sup>77</sup> Martin Kubo,<sup>59</sup> Jonathan Lawson,<sup>1</sup> Rasko Leinonen,<sup>23</sup> Stephanie Li,<sup>1,3</sup> Michael F. Lin,<sup>78</sup> Mikael Linden,<sup>79,80</sup> Xianglin Liu,<sup>60</sup> Isuru Udara Liyanage,<sup>23</sup> Javier Lopez,<sup>101</sup> Anneke M. Lucassen,<sup>81</sup> Michael Lukowski,<sup>44</sup> Alice L. Mann,<sup>3,15</sup> John Marshall,<sup>58</sup> Michele Mattioni,<sup>82</sup> Alejandro Metke-Jimenez,<sup>83</sup> Anna Middleton,<sup>64,85</sup> Richard J. Milne,<sup>64,85</sup> Fruzsina Molnar-Gabor,<sup>65</sup> Nicola Mulder,<sup>57</sup> Monica C. Munoz-Torres,<sup>53</sup> Rishi Nag,<sup>23</sup> Hidekawa Nakagawa,<sup>68,69</sup> Jamar Nasir,<sup>60</sup> Arcadi Navarro,<sup>52,91,92,93</sup> Tristan H. Nelson,<sup>64</sup> Ania Niewielska,<sup>23</sup> Amy Nisselle,<sup>17,26,95</sup> Jeffrey Niu,<sup>20</sup> Tommi H. Nyström,<sup>79,80</sup> Brian D. O'Connor,<sup>1</sup> Sabine Oesterle,<sup>8</sup> Soichi Ogishima,<sup>98</sup> Laura A.D. Paglione,<sup>97,98</sup> Emilie Palumbo,<sup>13,52</sup> Helen E. Parkinson,<sup>23</sup> Anthony A. Philippakis,<sup>1</sup> Angel D. Pizarro,<sup>99</sup> Andreas Prlic,<sup>100</sup> Jordi Rambla,<sup>13,52</sup> Augusto Rendon,<sup>101</sup> Renee A. Rider,<sup>65</sup> Peter N. Robinson,<sup>102,103</sup> Kurt W. Rodamer,<sup>104</sup> Laura Lyman Rodriguez,<sup>105</sup> Alan F. Rubin,<sup>25,26</sup> Manuel Rueda,<sup>52</sup> Gregory A. Rushton,<sup>1</sup> Rosalyn S. Ryan,<sup>106</sup> Gary I. Saunders,<sup>76</sup> Helen Schuilenburg,<sup>23</sup> Torsten Schwede,<sup>8,70</sup> Serena Scollen,<sup>76</sup> Alexander Sem,<sup>107</sup> Nathan C. Sheffield,<sup>108</sup> Neerjah Skantharajah,<sup>3,4</sup> Albert V. Smith,<sup>109</sup> Heidi J. Sofia,<sup>46</sup> Dylan Spalding,<sup>79,80</sup> Amanda B. Spurlock,<sup>110</sup> Zornitza Stark,<sup>15,17,28</sup> Lincoln D. Stein,<sup>1,73</sup> Kaito Suematsu,<sup>77</sup> Patrick Tan,<sup>84,111,112</sup> Jonathan A. Tedds,<sup>78</sup> Alastair A. Thomson,<sup>33</sup> Adrian Thorogood,<sup>9,113</sup> Timothy L. Tickle,<sup>1</sup> Katsushi Tokunaga,<sup>75,114</sup> Juha Törnroos,<sup>74,80</sup> David Torrents,<sup>92,116</sup> Sean Upchurch,<sup>113</sup> Alfonso Valencia,<sup>92,118</sup> Roman Valls Guimera,<sup>25</sup> Jessica Vamathevan,<sup>23</sup> Susheel Varma,<sup>23,117</sup> Danya F. Years,<sup>17,28,95,111</sup> Coby Viner,<sup>10,20</sup> Craig Voisin,<sup>119</sup> Alex H. Wagner,<sup>31,32</sup> Susan E. Wallace,<sup>10</sup> Brian P. Walsh,<sup>22</sup> Vivian Ota Wang,<sup>29</sup> Marc S. Williams,<sup>94</sup> Eva C. Winkler,<sup>120</sup> Barbara J. Wold,<sup>115</sup> Grant M. Wood,<sup>1</sup> J. Patrick Woolley,<sup>75</sup> Chisato Yamasaki,<sup>71</sup> Andrew D. Yates,<sup>23</sup> Christina K. Yung,<sup>4,121</sup> Lyndon J. Zass,<sup>87</sup> Ksenia Zaytseva,<sup>9,122</sup> Junjun Zhang,<sup>8</sup> Peter Goodhand,<sup>4,5</sup> Kathryn North,<sup>17,28</sup> and Ewan Birney<sup>23,123</sup>

ga4gh.org

# Get Involved! Visit **GA4GH.ORG**

## Join a Work Stream!

Contact [secretariat@ga4gh.org](mailto:secretariat@ga4gh.org)



**Become an Organizational Member**  
[ga4gh.org/members](http://ga4gh.org/members)



**Subscribe to GA4GH Updates**  
[ga4gh.org/subscribe](http://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 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



Google Cloud

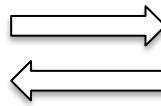
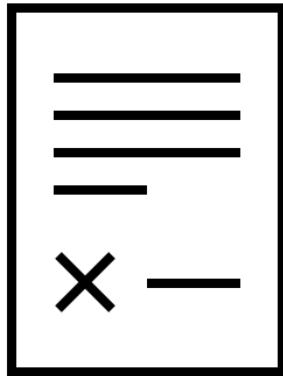


Microsoft Azure

# 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

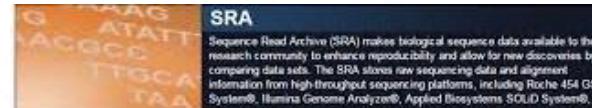
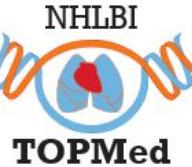


Example: U-Pitt enrolled in STRIDES. NIH-funded PIs supported by NIGMS (U24), NIDDK (U01), & NIDCD (R44) benefit from STRIDES discounts using the cloud to support their award/research activity

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



NATIONAL CANCER INSTITUTE  
GENOMIC DATA COMMONS



# 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



National Institutes of Health  
Turning Discovery Into Health

# 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](http://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

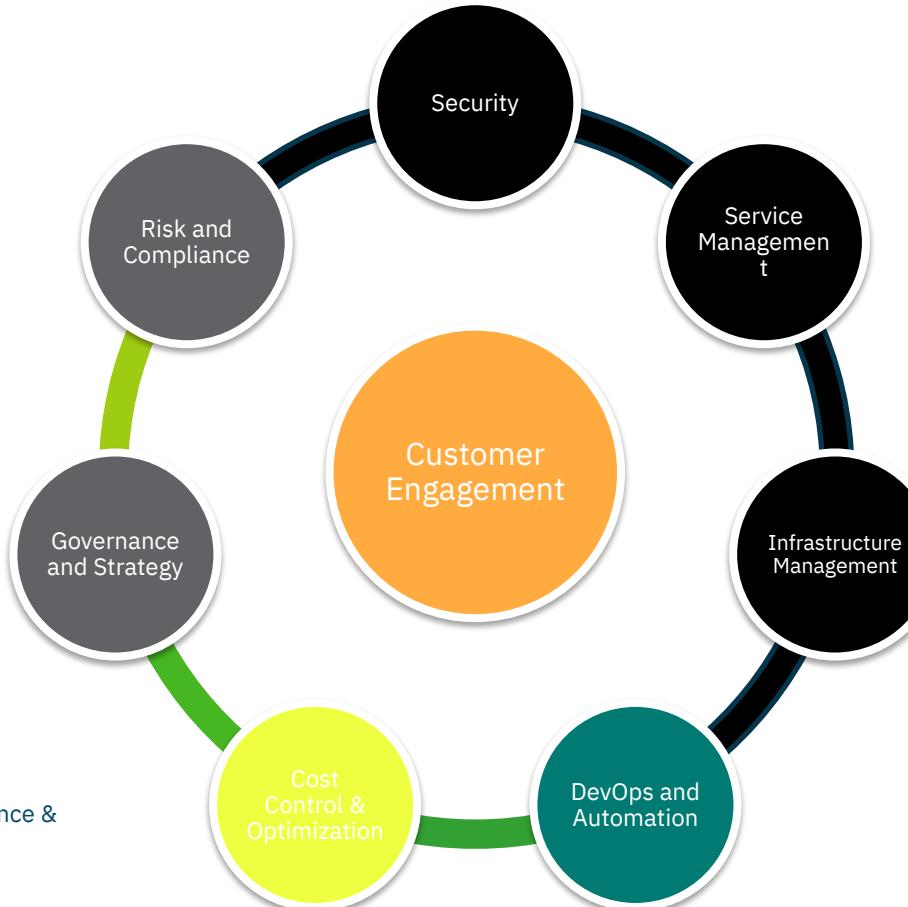
- FISMA, FedRAMP, & CSF
- NIST 800-37, -53, & -171
- Continuous monitoring

## 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

<i>Topic 1: Bringing researchers to cloud computing</i>	Tiffany Miller
<i>Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses</i>	Jack DiGiovanna
<i>Topic 3: What technologies and data types are missing across platforms?</i>	Ken Wiley
<i>Topic 4: Diversifying genomic data science</i>	Asiyah Lin   Kim Albero
<i>Topic 5: Flagship use cases for interoperability</i>	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	<ul style="list-style-type: none"><li>- Educate Policy people and program officers and include in development</li><li>- 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</li></ul>
"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/1NnYE84dRLSRtCBtVc2j8aOskQfD  
AEIXcT-nPsDan3XQ/edit#](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)

	JD	Jack DiGiovanna (me)
	NK	Natalie Kucher (Co-host)
	MF	Michael Feolo
	BG	Bruno Grande
	AH	Allison Heath
	BV	Ben Vizzier - UCSC GI (he/him)
	BS	Beth Sheets
	MB	Michael Baumann
	mc	mike conway
	TB	Teresa Barsanti

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)



In the first image, it is assumed that everyone will benefit from the same supports. They are being treated equally.



In the second image, individuals are given different supports to make it possible for them to have equal access to the game. They are being treated equitably.



In the third image, all three can see the game without any supports or accommodations because the cause of the inequity was addressed. The systemic barrier has been removed.

[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](mailto: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)