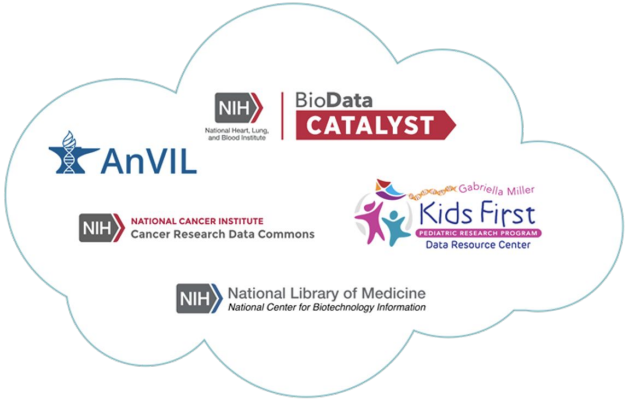
June 22-23, 2022

Welcome to Day 2
We will begin shortly…

NIH Cloud Platform Interoperability Spring 2022 Virtual Workshop



Today's Agenda

Day 2: Thursday, June 23, 2022

11:00 AM - 11:05 AM – Welcome and start of Day 2
Stephen Mosher (Johns Hopkins University)

Interoperability Driven Science

Cloud platform interoperability enables scientific discovery. Here we will learn of the latest advances in NCPI demonstration projects and related cloud platforms.

11:05 AM - 11:20 AM – The ELIXIR Cloud for European Life Sciences
Jonathan Tedds (ELIXIR)

11:20 AM - 11:35 AM – Sex chromosome complement aware alignments
Melissa Wilson (ASU)

11:35 AM - 11:50 AM – Genome-wide Sequencing Analysis to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors.
Nara Sobreira (JHU)

11:50 AM - 1:05 PM – Working Group Updates

15 min - Community/Governance WG
Bob Grossman (University of Chicago)
Stanley Ahalt (University of North Carolina at Chapel Hill)

15 min - Systems Interoperation WG
Jack DiGiovanna (SevenBridges)

15 min - FHIR WG
Robert Carroll (Vanderbilt University Medical Center)

15 min - NCPI Outreach WG
Stephen Mosher (Johns Hopkins University)

15 min - Search WG
Dave Rogers (Clever Canary)
Kathy Reinold (Broad Institute)

1:05 PM - 1:35 PM – Break

Technical Aspects of Interoperability

Technologies that enable interoperability are important to develop with stakeholders involved to promote the usability of the technical standards and products. In this session, we will hear about technologies enabling interoperability and their successful implementations in research.

1:35 PM - 1:50 PM – The Texas Advanced Computing Center (TACC) as an Interoperable Cloud Resource for Biomedical Research
Dan Stanzione (TACC)

1:50 PM - 2:05 PM – FHIR for Genomics: The Path Forward
Mullai Murugan (Baylor College of Medicine)

2:05 PM - 2:20 PM – Supporting Genomic Data Sharing through the Global Alliance for Genomics and Health
Heidi Rehm (Broad Institute)

2:20 PM - 2:35 PM – Interoperability Opportunities & Challenges with the Cloud and STRIDES
Nick Weber (NIH STRIDES)

2:35 PM - 3:10 PM – Concurrent Breakouts

Topic 1: Bringing researchers to cloud computing
Topic 2: Reproducibility and Interoperability of batch and ad hoc analyses
Topic 3: What technologies and data types are missing across platforms?
Topic 4: Diversifying genomic data science
Topic 5: Flagship use cases for interoperability

Day 2 Breakout Moderators

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

3:10 PM - 3:50 PM – Report Back
5 minutes for report prep; 5 minute report per group; 10 minutes open discussion

3:50 PM - 4:00 PM – Summary, Future Directions, & Meeting close
Michael Schatz (Johns Hopkins University)

4:00 PM – Meeting close

Interoperability Driven Science

11:05 AM - 11:50 AM EDT

The ELIXIR Cloud for European Life Sciences

Jonathan Tedds (ELIXIR)

The ELIXIR Cloud for European Life Sciences
NCPI Meeting, 23 June 2022

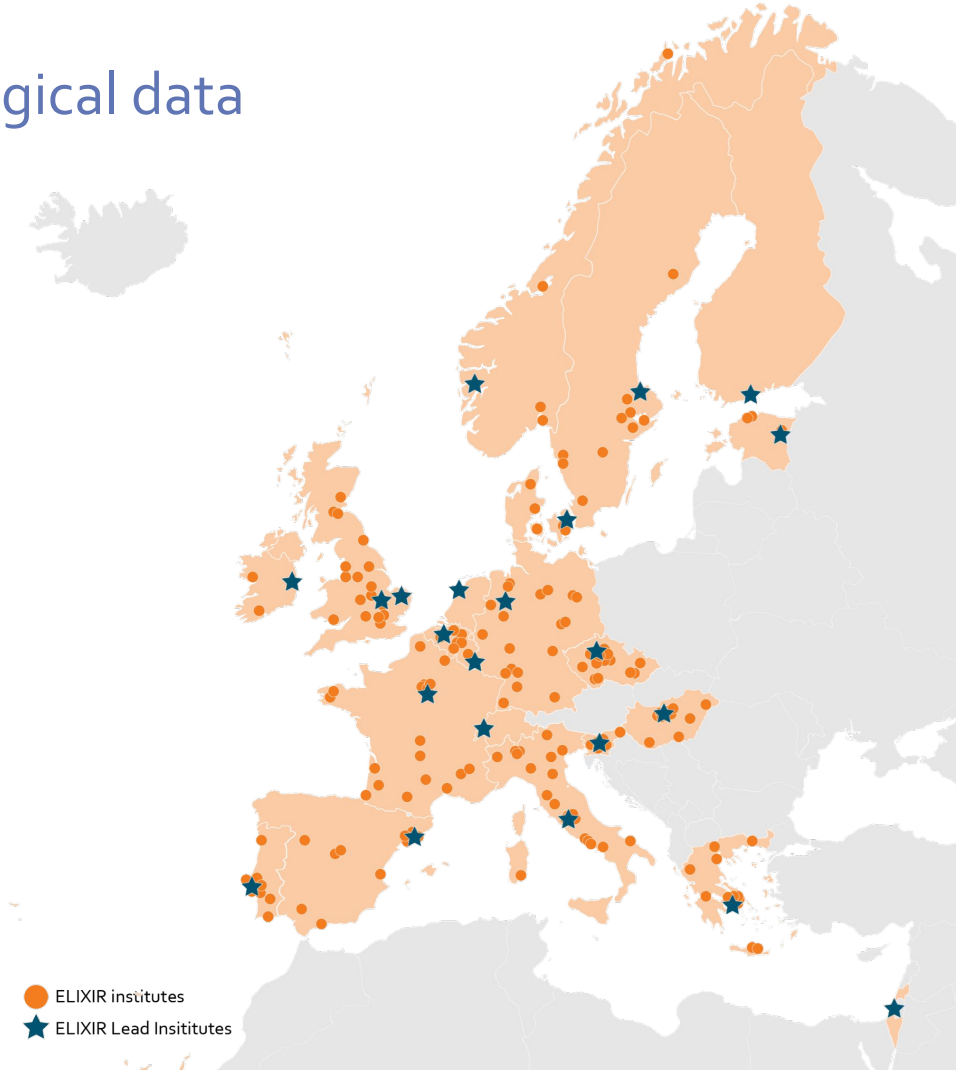
Jonathan Tedds (Compute, Tools Platform & EOSC Coordinator)

www.elixir-europe.org

A sustainable infrastructure for biological data

ELIXIR Members

ELIXIR Observers

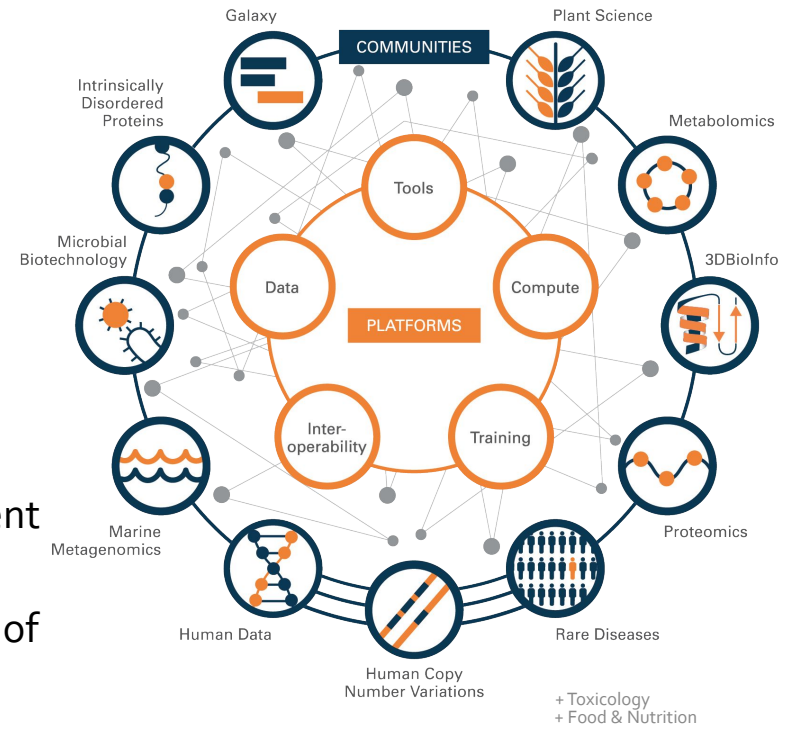


ELIXIR Services for all domains of life sciences

The ELIXIR Nodes **collectively run hundreds of bioinformatics services**, where:

- 5 Platforms coordinate services across all scientific domains and all the Nodes
- 13 Communities work in a particular domain and give feedback on platform services
- 12 Focus groups bring together people with an interest in a particular topic
- EU projects & internal projects drive development of services and knowledge exchange

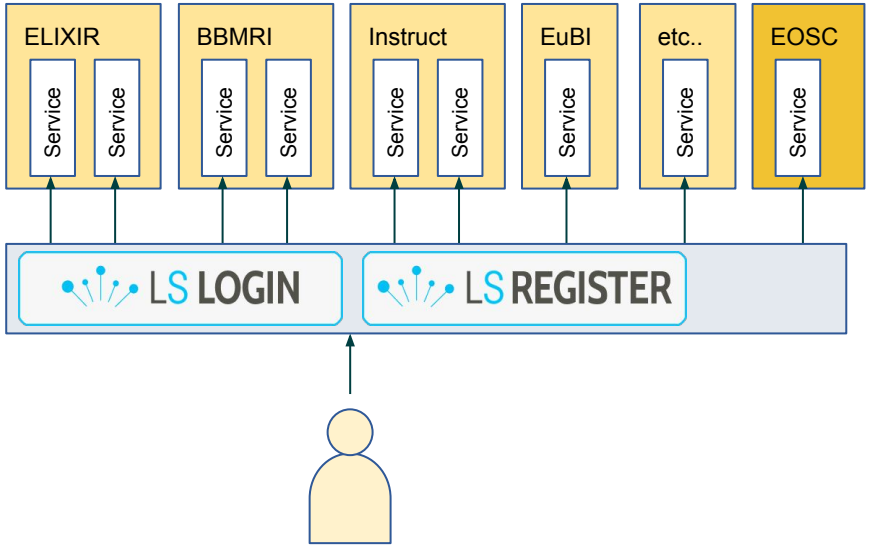
The vast majority of ELIXIR services are available free of charge and accessible globally by anyone interested



More: elixir-europe.org/how-we-work

Accessing ELIXIR Cloud and beyond: Life Science Login

- Common AAI for 13 European life science research infrastructures
- ELIXIR a major contributor
- Uses common internet standards
- Successful ELIXIR AAI migration to LS Login for users, April 2022
  - Services to follow
- Sustainable post-project service model
  - *Community driven*



https://lifescience-ri.eu/ls-login.html

EOSC-Life

elixir

Services & Solutions

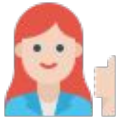
Galaxy PROJECT	WorkflowHub	ELIXIR::GA4GH Cloud
Web-based platform for reproducible computational analysis	Registry for describing, sharing and publishing scientific computational workflows	Federated, interoperable network of workflow engines and compute nodes based on GA4GH standards
ELIXIR Community	EOSC-Life resource	GA4GH Driver Project
APIs & (third-party) GUIs	API & GUI	APIs & third-party GUIs

Maturity

How we work

Represent ELIXIR stakeholders in GA4GH & **promote** GA4GH standards within ELIXIR

Prototype real-world use cases with ELIXIR stakeholders, **develop** PoCs & **deploy** at ELIXIR nodes



Consult on integrating GA4GH standards into existing solutions and provide **technical support**

Interoperability testing with third party GA4GH-powered solutions

Relevant GA4GH APIs

Passport

Grant access to data & compute

TRS: Tool Registry Service API

Access workflows and container images

DRS: Data Repository Service API

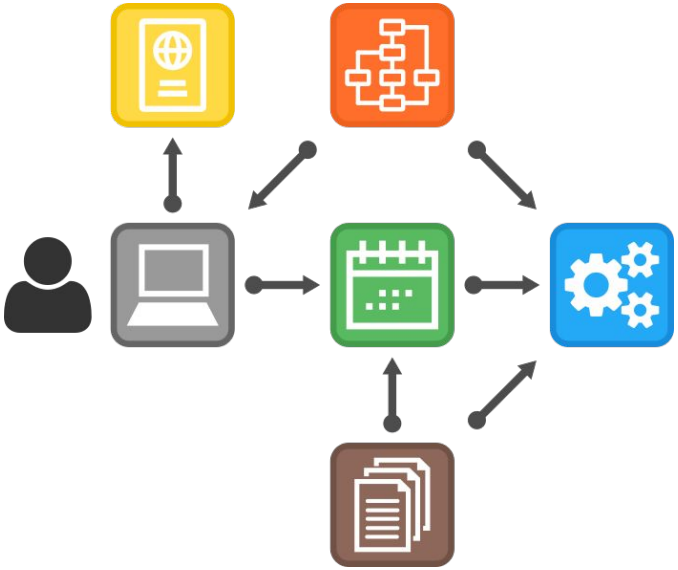
Access to data sets

WES: Workflow Execution Service API

Interpret workflows & schedule task execution

TES: Task Execution Service API

Execute tasks

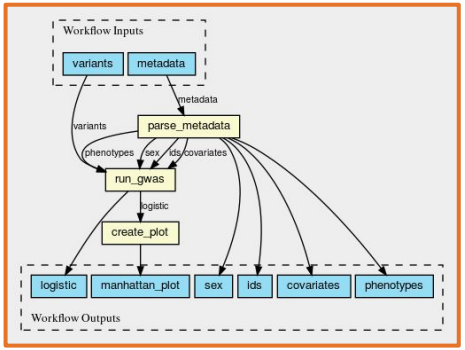


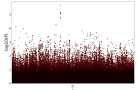
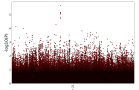
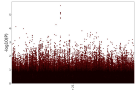
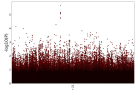
Moonshot demonstrator (8th GA4GH Plenary)

Goal: *Showcase* ***reproducibility of GA4GH Cloud*** *implementations*

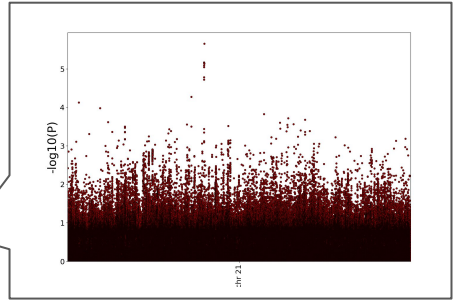
VCF,
1000 Genomes project

Simple GWAS analysis workflow



Platform	DNASTACK	Terra	elixir AAI	SevenBridges
GA4GH Cloud APIs	WES DNAstack	TRS Dockstore DRS Anvil DRS	TRS TRS-Filer / Biocontainers WES cwl-WES DRS RDSDS TES TESK	WES Seven Bridges DRS Seven Bridges
Results				

Identical results!



ELIXIR Cloud resources for COVID-19 response

Find computing resources to help you analyse datasets

ELIXIR runs computing services that can be accessed by research projects. Many additional computing resources have been made available to support COVID-19 research projects and a number offer access to Docker Orchestrators, including Mesos and OpenStack access, Kubernetes/OKD and potentially GPUs where needed. For assistance please contact jonathan.tedds@elixir-europe.org, ELIXIR's Compute Platform Coordinator. Examples of compute resources include:

- de.NBI cloud (ELIXIR Germany) provides priority access for projects relating to COVID-19.
- CSC (ELIXIR Finland) has prioritised access to its cloud services for COVID-19 research.
- e-INFRA CZ (ELIXIR Czech Republic) offers relaxed access conditions to supercomputer resources, storage services and distributed compute resources.
- EMBL-EBI is contributing EMBASSY Cloud resources as detailed on the European Open Science Cloud, EOSC Marketplace.
- A specific Galaxy COVID-19 instance for genomic analysis is available through Laniakea, ELIXIR Italy's on-demand platform.
- The European Galaxy server is an open, web-based platform for data intensive research and provides access to compute and storage resources. There are more than 2,500 different scientific tools, specific COVID-19 training materials, and workflows to guide users through COVID-19 data analysis.
- SIB (ELIXIR Switzerland) is providing a ready-to-use slurm workload manager with a scientific software stack via the ExPASy SIB Portal.
- IFB (ELIXIR France) is providing a federated set of high performance compute and cloud resources including national and regional servers.

Implementation Example

de.NBI – Deutsches Netzwerk für Bioinformatik Infrastruktur

de.NBI consortium

- 42 project partners
- 32 institutions
- 8 service centers
- designated national German node in ELIXIR

de.NBI mission

- Provision of comprehensive first-class bioinformatics services to users in basic and applied life sciences research
- Bioinformatics training in Germany and Europe through a wide range of workshops and courses
- Cooperation of the German bioinformatics community with international bioinformatics network structures

de.NBI Administration Office (AO)

de.NBI Service Centers

- **HD- HuB - Heidelberg Center for Human Bioinformatics**
  - DKFZ Heidelberg
  - EMBL Heidelberg
  - Universität Heidelberg
  - Universität des Saarlandes
  - Charité Berlin
  - Coordinator: P. Bork, Heidelberg
- **BiGi - Bielefeld-Gießen Resource Center for Microbial Bioinformatics**
  - Universität Bielefeld
  - Universität Gießen
  - Universität Magdeburg
  - Coordinator: A. Goesmann, Gießen
- **BioInfra.Prot - Bioinformatics for Proteomics**
  - Medizinisches Proteom-Center der Universität Bochum
  - Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V. Dortmund
  - Forschungszentrum Borstel
  - Max-Planck-Institut Molekulare Zellbiologie und Genetik, Dresden
  - Coordinator: M. Eisenacher, Bochum
- **CIBI - Center for Integrative Bioinformatics**
  - Freie Universität Berlin
  - Universität Konstanz
  - Universität Tübingen
  - Max-Planck-Institut für Molekulare Zellbiologie und Genetik, Dresden
  - Leibniz-Institut für Pflanzenbiochemie Halle
  - Coordinator: O. Kohlbacher, Tübingen
- **RBC - RNA Bioinformatics Center**
  - Universität Freiburg
  - Universität Leipzig
  - Max-Delbrück-Centrum für Molekulare Medizin Berlin
  - Universität Rostock
  - Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut e.V., Jena
  - Coordinator: R. Backofen, Freiburg
- **GCBN - German Crop BioGreenformatics**
  - Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung Gatersleben
  - Helmholtz Zentrum München
  - Forschungszentrum Jülich
  - Coordinator: U. Scholz, Gatersleben
- **BioData - Center for Biological Data**
  - Jacobs University Bremen - SILVA
  - Universität Bremen - PANGAEA
  - Leibniz-Institut DSMZ Braunschweig - BacDive
  - TU Braunschweig - BRENDA
  - Universität Hamburg - EnzymeStructures
  - Coordinator: F. O. Glöckner, Bremen
- **de.NBI-SysBio - de.NBI Systems Biology Service Center**
  - HITS Heidelberger Institut für Theoretische Studien
  - Universität Heidelberg
  - Max-Planck-Institut für Dynamik Komplexer Technischer Systeme, Magdeburg
  - Coordinator: W. Müller, Heidelberg
- **Associated Partners** – without de.NBI funding
  - Universität Kiel
  - Universität Jena
  - DKFZ Heidelberg

➡ www.denbi.de

de.NBI Cloud Federation

- fully **academic cloud** federation
- Established 2016
- provides **storage and computing resources** for the life sciences community
- **free of charge** for academic use
- federation is **maintained by the six German cloud centers** located in Bielefeld, Heidelberg, Berlin, Freiburg, Giessen and Tübingen
- de.NBI Cloud offers a solution to enable **integrative analyses,** the **efficient use of data** in research, and computational **capacities for bioinformatics training**.

https://cloud.denbi.de

de.NBI Cloud Infrastructure

Largest scientific cloud in Germany and one of the leading European academic clouds in life sciences

Computing Hardware

Focus on compute power
Specialized hardware (additionally GPU, FPGA)

~56,000

>529 GPUs

up to 4 TB

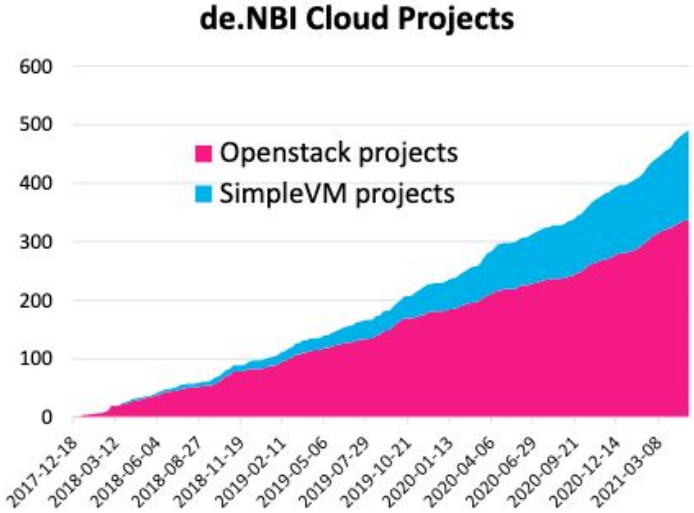
Storage capacity

Focus on reference data
Data and storage via different file storage protocols

80 PB

330 TB

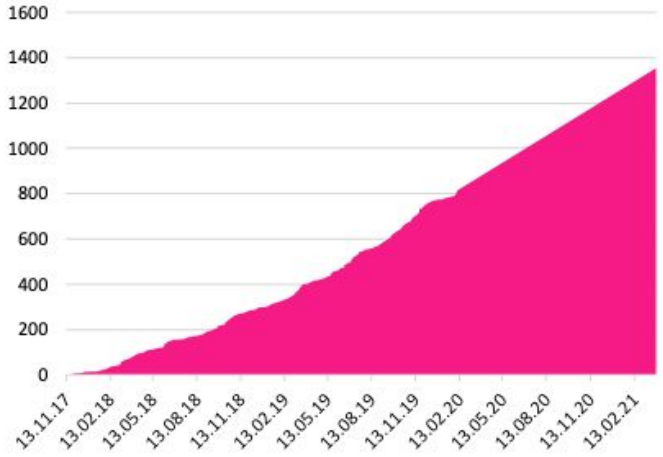
Project Numbers



Q1 2021: 323 OpenStack projects, 137 SimpleVM projects

- Full OpenStack Environment per Project
- For fully customizable provisioning and deployment of VMs and Services / Clusters

- Custom project-type based on OpenStack
- For simple deployment of VMs and Services / Clusters and integration of e.g. Bioconda



Q1 2021: 1355 registered users

+ 1000's of users of:

Galaxy EUROPE · BIIGLE · PhenoMeNal · EGGNOG-MAPPER · silva ngs

Global Alliance for Genomics & Health

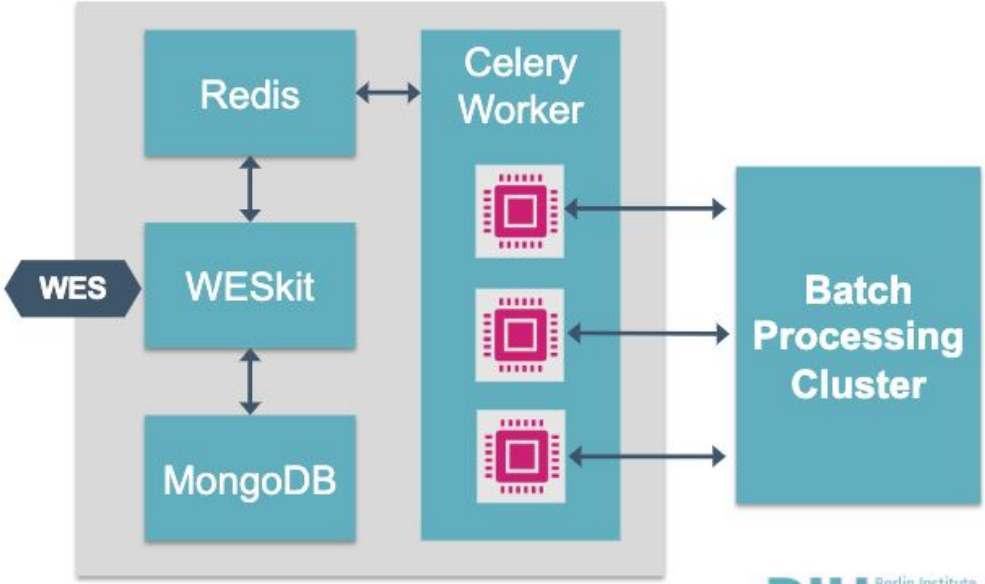
WESKIT

GA4GH WES implementation

https://gitlab.com/one-touch-pipeline/weskit

Features

- WES for Snakemake and Nextflow
- Developed for high data throughput usage at Charité Universitätsmedizin Berlin and DKFZ
- HPC and Cloud deployment supported



Containerized deployment in Docker Swarm

ELIXIR Cloud: Gap analysis

- Interoperable cost transfer / payment system
  - Okay for commercial clouds, but how about academia?
  - Science credits, credit cards, crypto? Not easy…
- Access control
  - Concrete vision of access control via Passport only shaping up now - planning for European Genomic Data Infrastructure project 2022+
  - But only for data so far, can ELIXIR spearhead compute access?
- Sensitive data
  - How to secure data beyond access control
  - Crypt4GH, multi-party homomorphic encryption: how to integrate with Cloud APIs?
- Technical implementation support
  - COVID-19 response illustrated the importance of skilled technical support

Sex chromosome complement aware alignments

Melissa Wilson (ASU)

Sex chromosome complement aware alignment

Brendan Pinto and Melissa Wilson

Many Thanks

Brian O'Connor

@boconnor

Michael Schatz

@mike_schatz

Samantha Zarate

@sz_genomics

Who are we?

Brendan Pinto

@drpintothe2nd

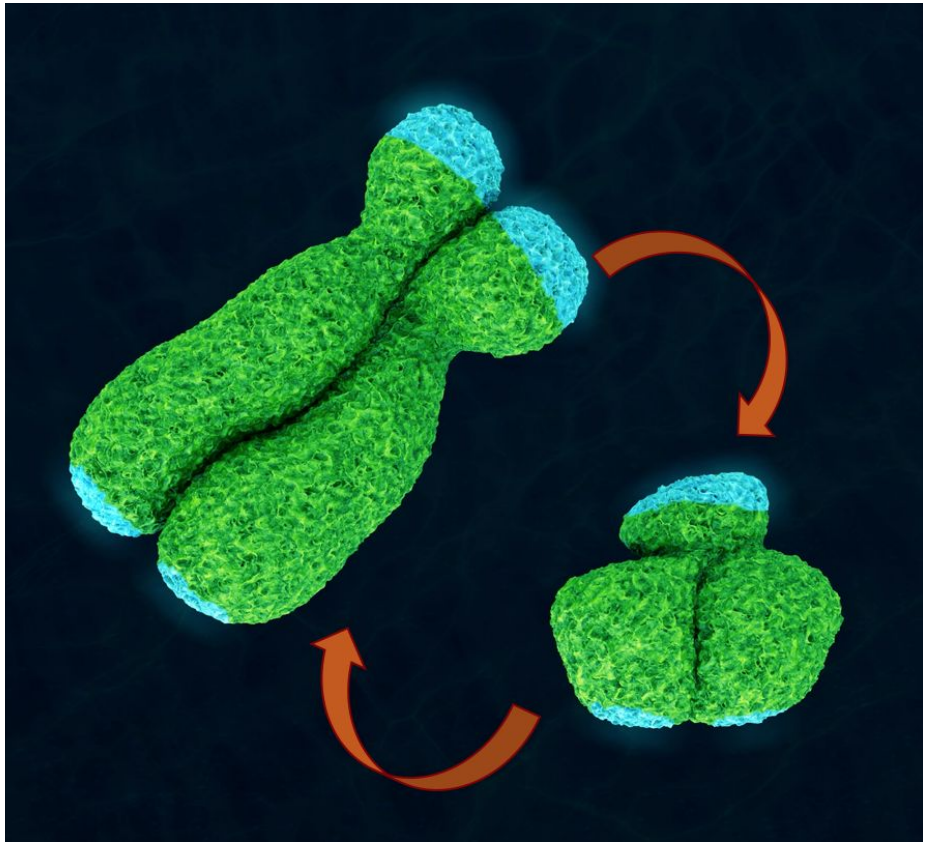
Melissa Wilson

@sexchrlab

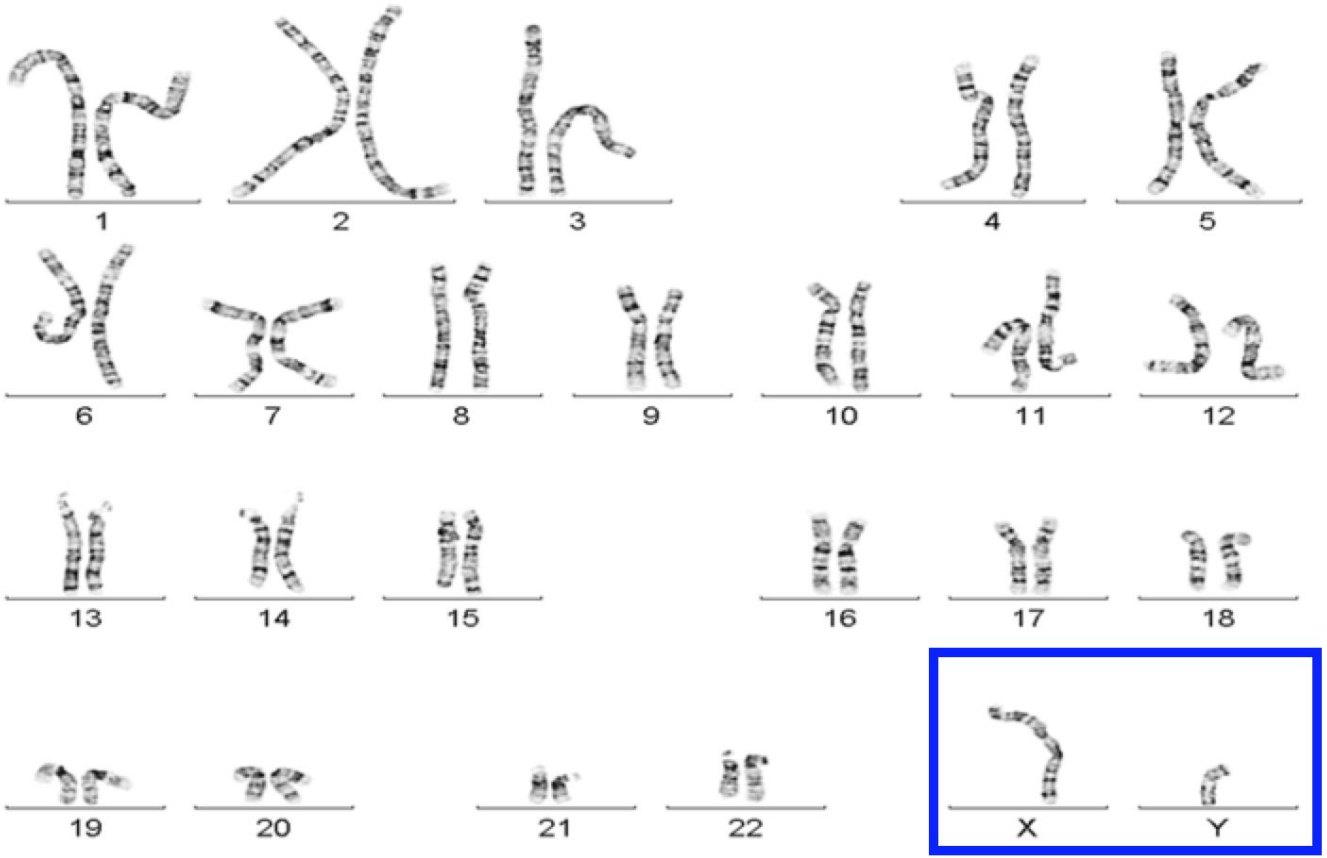
ANALYZE ALL THE GENOMES

Sex chromosomes share sequence similarity

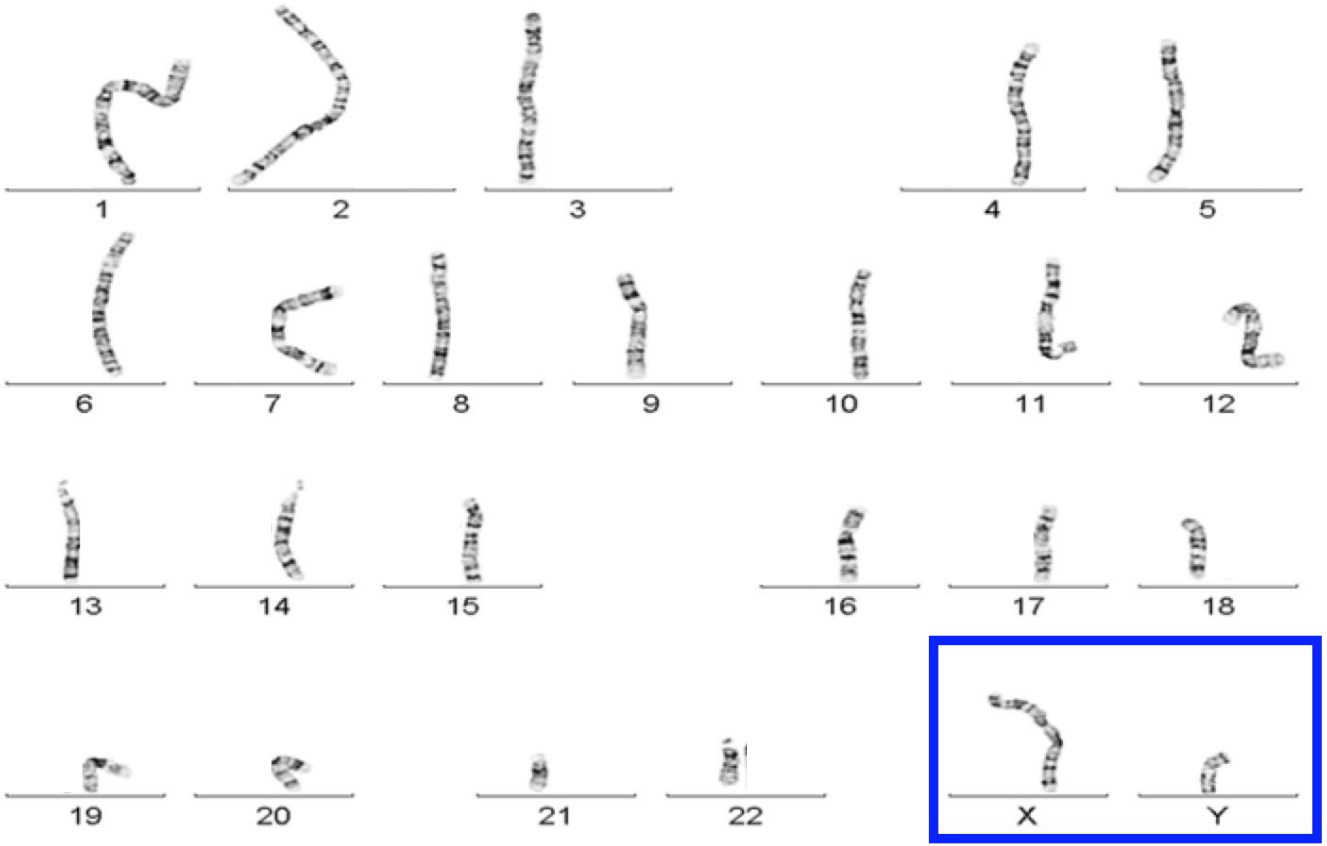
- The X and Y chromosomes share sequence similarity due to shared evolutionary ancestry that affects alignments and quantification of NGS data
- PARs share 100% homology



Human karyotype



Human reference genome



github.com/SexChrLab/XYalign

Realign with appropriate sex chromosome masks

XX samples: hard mask chrY

XY samples: hard mask PARs on chrY

Workflow overview

Data: 15 female (XX) samples (GTEX)

1. Convert CRAM to BAM format (samtools)
2. Strip reads from GRCh38 BAM files (samtools/bbmap)

   4.1. Trim reads + FastQC (Trim Galore!)
3. Re-map reads to CHM13v2.0 (bwa/samtools)
   a. Karyotype aware (Y hard-masked)
   b. Karyotype unaware (default)
4. Call haplotypes (GATK)
5. Call variants - GenotypeVCFs (GATK)

Called SNPs overview: "X vs. Autosome"

Total numbers of quality-filtered, biallelic SNPs called:

Chromosome	Unaware (GenBank*)	Aware (XYalign)	% change (A/U)
chr8	567,459	566,549	-0.17%
chrX	363,652	418,786	+15.2%

Called SNPs overview: X chromosome breakdown

Total numbers of quality-filtered, biallelic SNPs called:

chrX Region	Unaware (GenBank*)	Aware (XYalign)	% change (A/U)
PAR (2.8 Mbp)	34	1,118	+3,188.2%
XTR (4.7 Mbp)	15,103	19,140	+26.7%
non-PAR (151 Mbp)	348,515	398,528	+14.4%

Called SNPs overview: X chromosome breakdown

Total numbers of quality-filtered, biallelic SNPs called:

chrX (intragenic) regions	Unaware (GenBank*)	Aware (XYalign)	% change (A/U)
PAR (1.3 Mbp)	7	410	+5,757.1%
XTR (1.0 Mbp)	2,863	3,841	+34.2%
non-PAR (59.3 Mbp)	120,317	140,683	+16.9%

ANALYZE ALL THE GENOMES?

Consistent issues

Most issues that we ran into can be binned into two categories:

1. Unhelpful WOMtool validation errors (specifically when porting to Terra), e.g.

   a. Error message: "`ERROR: Unexpected symbol (line 6, col 5) when parsing 'setter'. Expected equal, got "String". String bam_to_reads_mem_size ^ $setter = :equal $e -> $1`"

   b. Translation: "WDL missing a dedicated inputs section."

   c. Why is this an issue? Unhelpful error messages inhibit forward progress.

Issues continued

2. Data localization during analysis, e.g.

a. Error message (GATK): "`A USER ERROR has occurred: … Cannot read non-existent file: <PATH-TO-**VERY**-EXISTENT-FILE.txt>`"

b. Translation: "GATK cannot stream data from your Google Bucket, try something else."

c. Work-around: Copy all inputs into the working directory for each WDL task — call input as a String instead of a File..

d. Why is this an issue? As nearly every program gets caught by this issue, the documentation on this is exceptionally poor. Only found 2 reports of this on 2 different forums (Terra and GATK) after weeks(!) of searching. 😭😭

(Many) fatal errors, but not new errors!

David Heiman
2 years ago

Hi Beri, there was no fix, only a hack - I wrote a WDL to copy the files to the workspace, then ran on those.

1) The error was:

```
A USER ERROR has occurred: Couldn't read file. Error was: drs:/dataguids.org/76cc4177-cf95-4
```

The issue is that the drs:// file paths are not being resolved to gs:// paths. My suspicion is that the WDL workflow defining the inputs bams as Array[String] rather than Array[File] may be causing the

file localization not working

P

Philipp Hahnel
7 months ago· 18 comments

Follow

Hi, I've checked the other related articles on issues with file localization, and my problem doesn't seem to be amongst those. I've written a WDL to use samtools on a bam and a ref fasta.

1. Problem: The bai does not localize, all other files are localized:

```
2021/11/22 19:12:43 Starting container setup.
```

In summary…

- We can do really incredible things with sex chromosome complement aware alignments to improve variant calling

- We can do this at scale on Terra

- It's going to take us a while longer to figure out how to do this at scale on Terra
  - Getting started on Terra – adding odd Terra-specific quirks for beginners?

Genome-wide Sequencing Analysis to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors

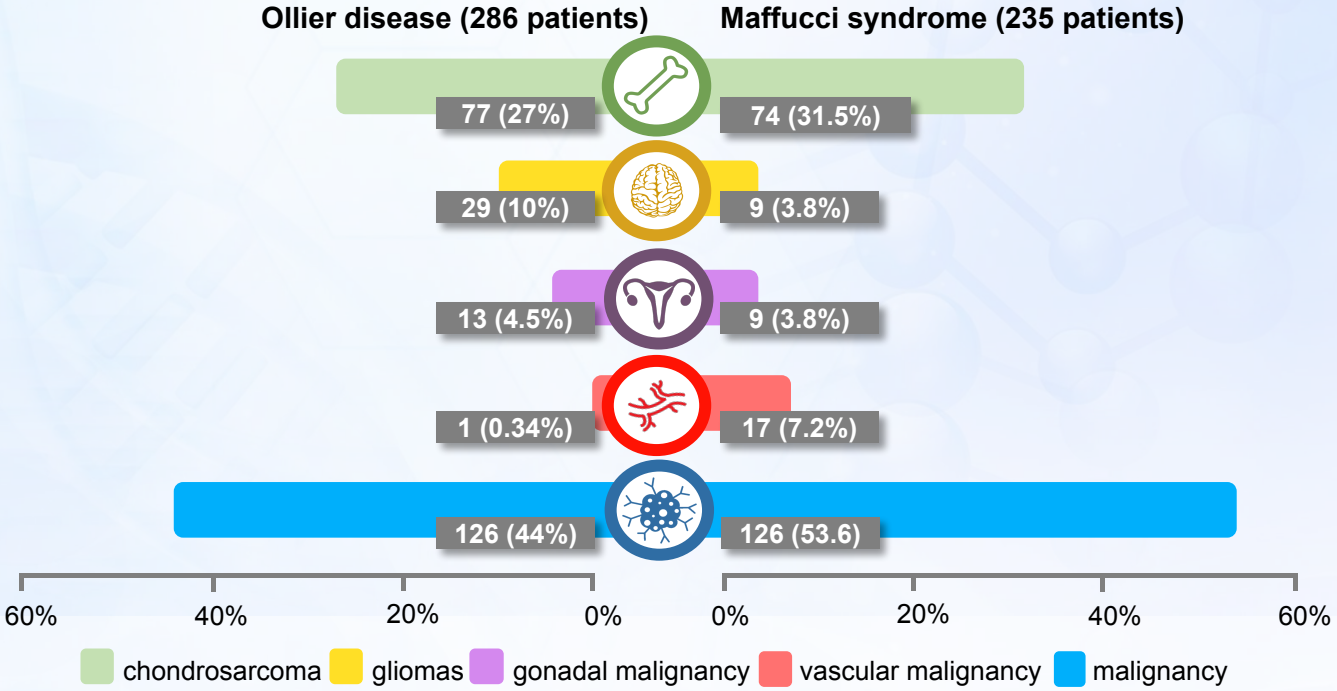
Nara Sobreira (Johns Hopkins University)

Genome-wide Sequencing Analysis to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors

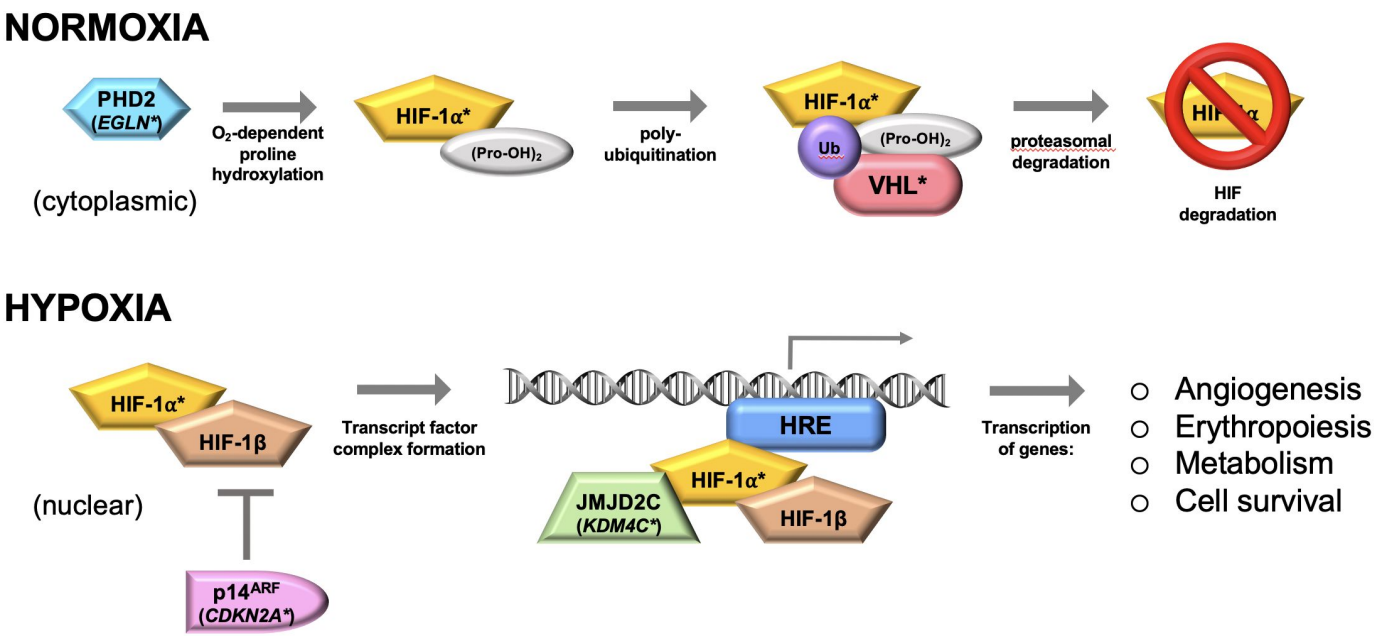
Renan Martin
Nara Sobreira
Johns Hopkins University School of Medicine

Scientific question

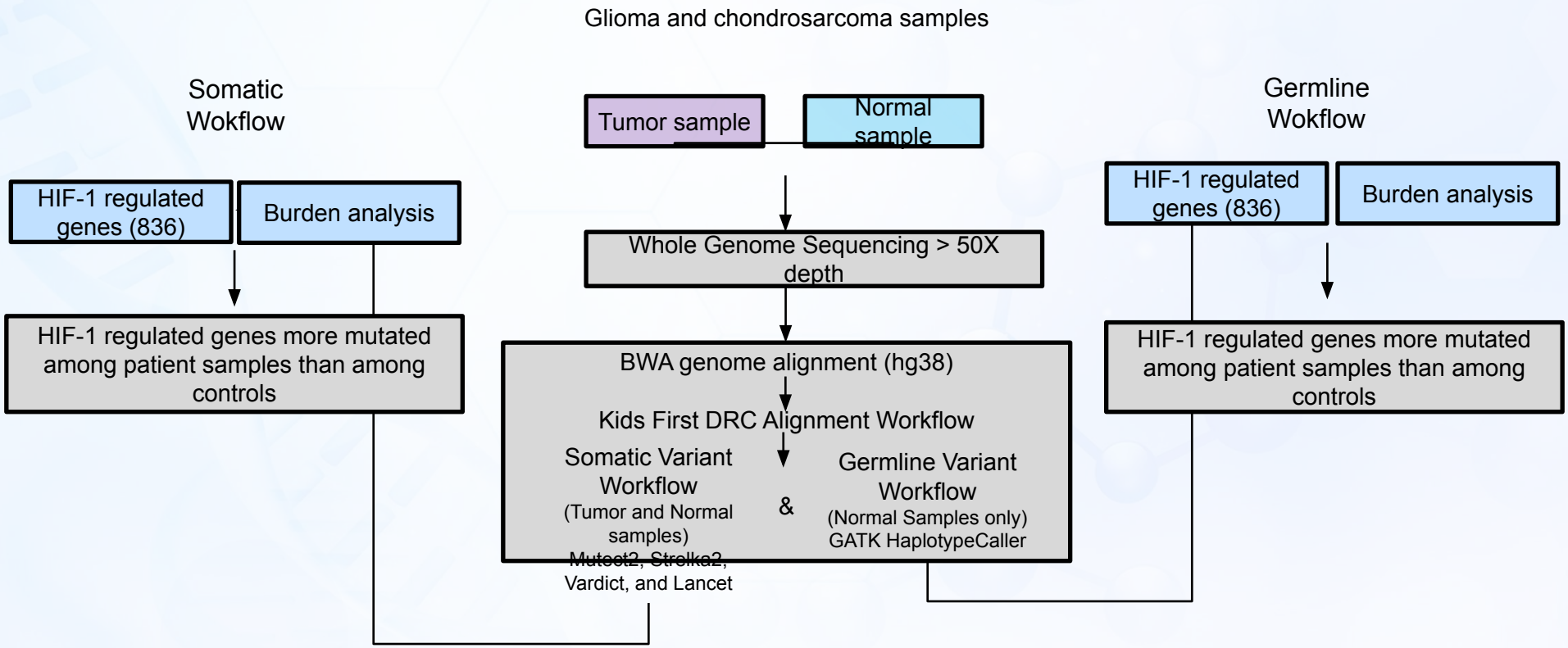
- Are pathogenic variants in genes related to HIF-1 pathway mutated in patients with Ollier disease and Maffucci syndrome and in patients with isolated forms of gliomas and chondrosarcomas?



25% of the patients have variants in one of 7 genes related to the HIF-1 pathway



Regulation of HIF-1α degradation at normoxia and hypoxia. * Genes found mutated in patients with OD or MS.


Glioma and chondrosarcoma samples
Somatic Wokflow
HIF-1 regulated genes (836)
Burden analysis
HIF-1 regulated genes more mutated among patient samples than among controls
Tumor sample
Normal sample
Whole Genome Sequencing > 50X depth
BWA genome alignment (hg38)
Kids First DRC Alignment Workflow
Somatic Variant Workflow
(Tumor and Normal samples)
Mutect2, Strelka2, Vardict, and Lancet
&
Germline Variant Workflow
(Normal Samples only)
GATK HaplotypeCaller
Germline Wokflow
HIF-1 regulated genes (836)
Burden analysis
HIF-1 regulated genes more mutated among patient samples than among controls

Interoperability plan

- ❑ Access germline WGS data from 61 probands (trios) with Ollier disease and Maffucci syndrome sequenced as part of the Gabriella Miller Kids First Pediatric Research Program and stored in CAVATICA
- ❑ Access germline WES data from 33 probands with Ollier disease and Maffucci syndrome sequenced as part of the BHCMG-CMG Program and stored in AnVIL
- ❑ Access tumor (and corresponding non-tumor tissue) WGS data from 816 patients from the Pediatric Brain Tumor Atlas (CBTN and PNOC)
  - ✔ Data will be accessed through the Kids First Program Data Resource Center and CAVATICA
- ❑ Access tumor WGS data from 878 patients with chondrosarcoma (PNOC)
  - ✔ Data will be accessed through the National Cancer Institute's Cancer Research Data Commons (NCI CRDC)

Pediatric Brain Tumor Atlas Datasets

CBTN
CRDC dataset (within CCDI)

- 998 probands
- 783 with VCF (harmonized pipeline)

PNOC
Kids First Collaborator dataset

- 79 probands
- 33 with VCF (harmonized pipeline)

Status

- Already accessible through CAVATICA

Aligned Reads Individual gVCFs
This dataset includes genomic data that are c...

Aligned Reads Individual gVCFs

Aligned Reads Individual gVCFs Family-Base
This dataset includes genomic data that are c...

Kids First: Familial Leukemia
NIH X01 Project Abstract - Charles Mullighan, PI
phs001738 dbGaP Study Page
VCFs Aligned Reads Unaligned Reads

Kids First: Orofacial Cleft - African and Asian Ancestry
NIH X01 Project Abstract - Azeez Butali and Te...
phs001997 dbGaP Study Page
Aligned Reads Individual gVCFs
This dataset includes genomic data that are c...

Kids First: Novel Cancer Susceptibility in Families (from BASIC3)
NIH X01 Project Abstract - Sharon Plon, PI
phs001878 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: Osteosarcoma
NIH X01 Project Abstract - Kenan Onel, PI
phs001714 dbGaP Study Page
Aligned Reads

Kids First: Craniofacial Microsomia
NIH X01 Project Abstract - Daniela Luquetti, PI
phs002130 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base
This dataset includes genomic data that are c...

Kids First: Kidney and Urinary Tract Defects
NIH X01 Project Abstract - Ali Gharavi, PI
phs002162 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: Microtia - Hispanic
NIH X01 Project Abstract - Jonathan Seidman, ...
phs002172 dbGaP Study Page
Aligned Reads

Kids First: Intersections of Cancer & SBD
NIH X01 Project Abstract - Hakon Hakonarson,...
phs001846 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: Esophageal Atresia and Tracheoesophageal Fistulas
NIH X01 Project Abstract - Wendy Chung, PI
phs002161 dbGaP Study Page
Aligned Reads

Kid First: Hemangiomas (PHACE)
NIH X01 Project Abstract - Dawn Siegel, PI
phs001785 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base
This dataset includes genomic data that are c...

Kids First: Nonsyndromic Craniosynostosis
NIH X01 Project Abstract - Simeon Boyd, PI
phs001806 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: Myeloid Malignancies
NIH X01 Project Abstract - Soheil Meshinchi, PI
phs002187 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: Leukemia & Heart Defects in Down Syndrome
NIH X01 Project Abstract - Philip Lupo and Ste...
phs002330 dbGaP Study Page
Aligned Reads Individual gVCFs Family-Base

Kids First: T-Cell ALL
NIH X01 Project Abstract - David Teachey, PI
phs002276 dbGaP Study Page
Aligned Reads VCFs

Gallery View | Table View

Available Collaborator Datasets

Pediatric Brain Tumor Atlas: CBTTC
CBTTC Website
CBTTC Data Access Form
Aligned Reads VCFs

TARGET: Acute Myeloid Leukemia
phs000465 dbGaP Study Page
Aligned Reads

TARGET: Neuroblastoma
phs000467 dbGaP Study Page
Aligned Reads

Pediatric Brain Tumor Atlas: PNOC
CBTTC Website
CBTTC Data Access Form
Aligned Reads Gene Expression TSVs Ann

Open DIPG ICR London
Open Access
No Application Necessary
IDAT GPR

Open as page

Pediatric Brain Tumor Atlas: CBTTC

First Portal Releas...	June 18, 2018
Data Types Availa...	Aligned Reads VCFs
Sequencing Center	Multiple
About the Study	CBTTC Website
Applying for Acce...	CBTTC Data Access Form
Data Access Com...	CBTTC Data Access Committee
Known Data Issues	CBTTC clinical event data is collected in a way that associates a diagnosis to a biospecimen, most often a tumor. A participant can have multiple tumors over time that have different diagnoses. Currently, this data in the Kids First Data Resource Portal is being presented as a diagnosis being attached to the participant and the association between tumor and diagnosis is not being displayed. This issue is being worked on. In the meantime, a list of diagnoses and directly associated clinical events is available by emailing support@kidsfirstdrc.org.
Note	Empty

Returning?

A A A

Children's Brain Tumor Network

Until every child is cured

CBTN Request Form

NOTE: Sample processing at the Operations Center and sample shipments may be delayed due to limited on-site personnel. Once you submitted your request and it is approved, we will provide the timeline by which we would deliver your specimens. We thank you for your patience and understanding during this time.

Please complete the Specimen/Data Use Request Form below.

Please keep in mind the following timeline after the submission of your request. All time is in business days.

Specimen Requests:

A primary reviewer reviews specimen requests within two weeks, and then the CBTN scientific committee has two weeks for any additional questions/comments.
Cell line requests will be reviewed within a week of submission by the Operations Center and Scientific co-Chair(s)

Data Use Requests:

CBTN Institutions: **Raw Genomic Data, Clinical Data, Imaging**

1. The request is reviewed for completeness by the CBTN Operations Center (1 day)
2. Access to the data is granted

Non-CBTN Institutions: **Clinical Data, Imaging**

1. The request is reviewed for completeness by the CBTN Operations Center (1 day)
2. Access to the data is granted.

Non-CBTN Institutions: **Raw Genomic Data,**

1. The request is reviewed for completeness by the CBTN Operations Center (1 day)
2. The request is submitted to the CBTN Data Use Committee for review. The committee has one week for review/questions/comments.
3. The investigator is responsible for providing executed DUA per NIH GDS requirements for the release of data.

If you have any questions or concerns regarding either process, please email research@cbtn.org. For additional information about CBTN, please visit CBTN.org.

What are you requesting:
* must provide value

- [ ] Specimens
- [ ] Data

Kids First Data Resource Center

Dashboard | Studies | Explore Data | Variant | File Repository | Members | Resources (New) | Nara

Studies

Search Studies

KF-DSD, Neuroblastoma...

Domain

Select All | None

- [ ] Cancer 2

Program

Select All | None

- [x] Pediatric Brain Tumor Atlas 2
- [ ] Kids First 24
- [ ] TARGET 2
- [ ] CARING 1
- [ ] ICR 1

Family Data

- False 1

Program = Pediatric Brain Tumor Atlas ×

2

+ New query

Showing 2 studies

Code	Name	Program	Domain	dbGap	Participants	Available participants per Data Category								
						Families	Seq	Snv	Cnv	Exp	Sv	Pat	Rad	C
PBTA-PNOC	Pediatric Brain Tumor Atlas: PNOC	Pediatric Brain Tumor Atlas	Cancer		79	0	66	59		30				
PBTA-CBTN	Pediatric Brain Tumor Atlas: CBTTC	Pediatric Brain Tumor Atlas	Cancer		5944	4512	992	744		1		901	248	8
					6023	**4512**	**1058**	**803**	**0**	**31**	**0**	**901**	**248**	**8**

Kids First Data Resource Center

Dashboard | Studies | Explore Data | Variant | File Repository | Members | Resources (New) | Nara

Search Studies

KF-DSD, Neuroblastoma...

Domain

Select All | None

- Birth Defect 16
- Cancer 10

Program

Select All | None

- [x] Kids First 24
- Pediatric Brain Tumor Atlas 2
- TARGET 2
- CARING 1
- ICR 1

KF-ED	Kids First: Enchondromatoses	Kids First	Cancer	phs001987	82	28	82	82
KF-OCEA	Kids First: Orofacial Cleft - European Ancestry	Kids First	Birth Defect	phs001168	1414	474	1295	1295
KF-TALL	Kids First: T Cell ALL	Kids First	Cancer	phs002276	1133	0	1133	1133
KF-GMHP	Kids First: Microtia - Hispanic	Kids First	Birth Defect	phs002172	334	182	334	334
KF-GNINT	Kids First: Intersections of Cancer & SBD	Kids First	Cancer, Birth Defect	phs001846	1777	1467	1776	1776
KF-OFCLA	Kids First: Orofacial Cleft - Latin American	Kids First	Birth Defect	phs001420	804	271	804	804
KF-FALL	Kids First: Familial Leukemia	Kids First	Cancer	phs001738	365	56	365	365
KF-CM	Kids First: Craniofacial Microsomia	Kids First	Birth Defect	phs002130	245	81	222	222
KF-SCD	Kids First: Syndromic Cranial Dysinnervation	Kids First	Birth Defect	phs001247	801	248	801	801
KF-KUT	Kids First: Kidney and Urinary Tract Defects	Kids First	Birth Defect	phs002162	132	44	132	132

portal.kidsfirstdrc.org/dashboard

Kids First

Dashboard | Studies | Explore Data | Variant | File Repository | Members | Resources (New)

My Dashboard

My Saved Queries

Cohort Queries 0 | File Queries 1

Explore Data and save virtual studies!

Authorized Studies 5

Kids First: Neuroblastoma — Authorized: 5,625 / 18,054 files

Data Use Groups: Open Access

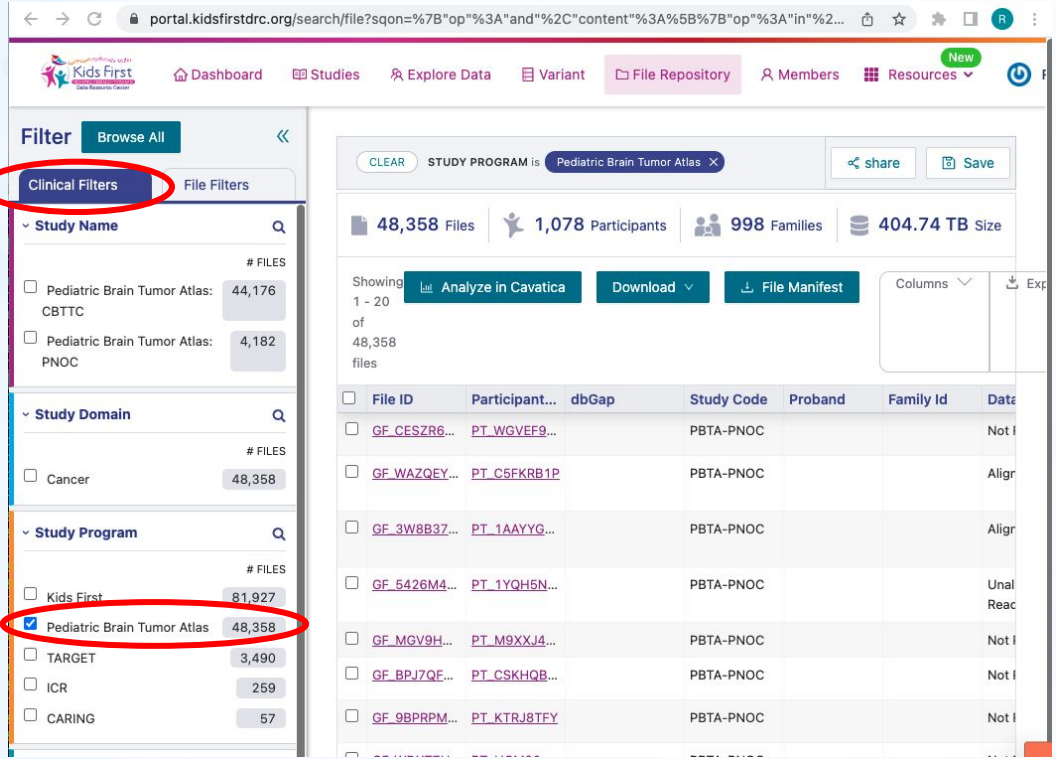
Kids First: Leukemia & Heart Defects in Down Syndrome — Authorized: 2,400 / 11,149 files

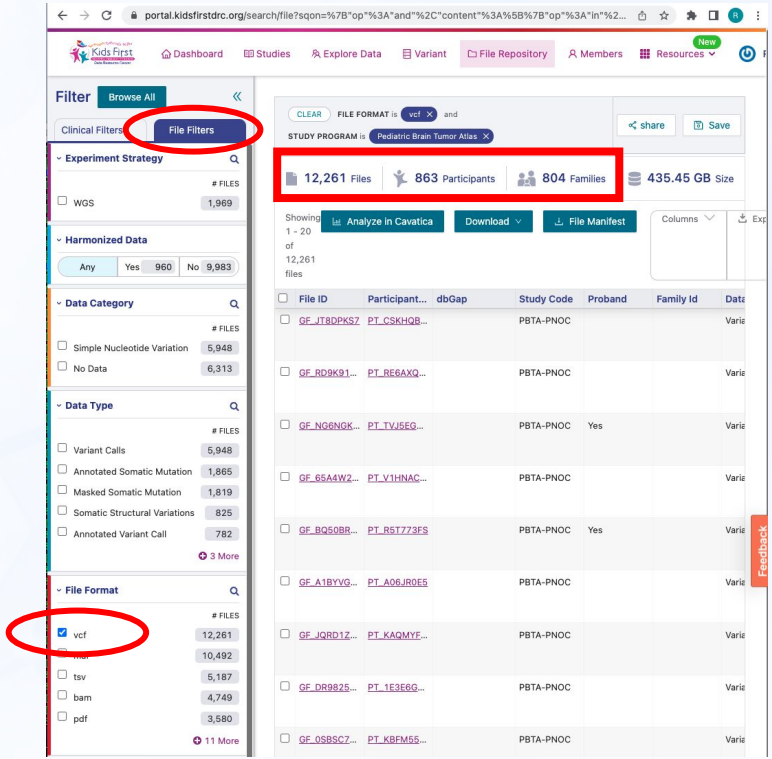
Data Use Groups: Open Access

Pediatric Brain Tumor Atlas: PNOC — Authorized: 1,763 / 4,182 files

Data Use Groups: Open Access

OpenDIPG: ICR London — Authorized: 259 / 259 files


portal.kidsfirstdrc.org/search/file?sqon=%7B"op"%3A"and"%2C"content"%3A%5B%7B"op"%3A"in"%2...
Kids First
Dashboard
Studies
Explore Data
Variant
File Repository
Members
Resources
New
Filter
Browse All
Clinical Filters
File Filters
Study Name
FILES
Pediatric Brain Tumor Atlas: CBTTC 44,176
Pediatric Brain Tumor Atlas: PNOC 4,182
Study Domain
FILES
Cancer 48,358
Study Program
FILES
Kids First 81,927
Pediatric Brain Tumor Atlas 48,358
TARGET 3,490
ICR 259
CARING 57
CLEAR
STUDY PROGRAM is Pediatric Brain Tumor Atlas
share
Save
48,358 Files
1,078 Participants
998 Families
404.74 TB Size
Showing 1 - 20 of 48,358 files
Analyze in Cavatica
Download
File Manifest
Columns
File ID
Participant...
dbGap
Study Code
Proband
Family Id
GF_CESZR6...
PT_WGVEF9...
PBTA-PNOC
GF_WAZQEY...
PT_C5FKRB1P
PBTA-PNOC
GF_3W8B37...
PT_1AAYYG...
PBTA-PNOC
GF_5426M4...
PT_1YQH5N...
PBTA-PNOC
GF_MGV9H...
PT_M9XXJ4...
PBTA-PNOC
GF_BPJ7QF...
PT_CSKHQB...
PBTA-PNOC
GF_9BPRPM...
PT_KTRJ8TFY
PBTA-PNOC


portal.kidsfirstdrc.org/search/file?sqon=%7B"op"%3A"and"%2C"content"%3A%5B%7B"op"%3A"in"%2...
Kids First
Dashboard
Studies
Explore Data
Variant
File Repository
Members
Resources
New
Filter
Browse All
Clinical Filters
File Filters
Experiment Strategy
FILES
WGS 1,969
Harmonized Data
Any
Yes 960
No 9,983
Data Category
FILES
Simple Nucleotide Variation 5,948
No Data 6,313
Data Type
FILES
Variant Calls 5,948
Annotated Somatic Mutation 1,865
Masked Somatic Mutation 1,819
Somatic Structural Variations 825
Annotated Variant Call 782
3 More
File Format
FILES
vcf 12,261
10,492
tsv 5,187
bam 4,749
pdf 3,580
11 More
CLEAR
FILE FORMAT is vcf and
STUDY PROGRAM is Pediatric Brain Tumor Atlas
share
Save
12,261 Files
863 Participants
804 Families
435.45 GB Size
Showing 1 - 20 of 12,261 files
Analyze in Cavatica
Download
File Manifest
Columns
File ID
Participant...
dbGap
Study Code
Proband
Family Id
GF_JT8DPKS7
PT_CSKHQB...
PBTA-PNOC
GF_RD9K91...
PT_RE6AXQ...
PBTA-PNOC
GF_NG6NGK...
PT_TVJ5EG...
PBTA-PNOC
Yes
GF_65A4W2...
PT_V1HNAC...
PBTA-PNOC
GF_BQ50BR...
PT_R5T773FS
PBTA-PNOC
Yes
GF_A1BYVG...
PT_A06JR0E5
PBTA-PNOC
GF_JQRD1Z...
PT_KAQMYF...
PBTA-PNOC
GF_DR9825...
PT_1E3E6G...
PBTA-PNOC
GF_0SBSC7...
PT_KBFM55...
PBTA-PNOC
Feedback

portal.kidsfirstdrc.org/search/file?sqon=%7B"op"%3A"and"%2C"content"%3A%5B%7B"op"%3A"in"%2C"content"%3A%7B"field"%3A"file_format"%2C"value"%3A%5B"vcf"%5D%7D%7D%2C%7B"op"%3A"in"%2C"content"%3A...

Kids First | Dashboard | Studies | Explore Data | Variant | File Repository | Members | Resources (New) | Renan

Filter Browse All

Clinical Filters | File Filters

Data Type

- Variant Calls 4,609
- Annotated Somatic Mutation 1,865
- Masked Somatic Mutation 1,819
- Annotated Variant Call 782
- Somatic Structural Variations 782

3 More

File Format

FILES

- vcf 10,879
- maf 9,180
- tsv 5,027
- bam 4,484
- pdf 3,548

9 More

Family Shared Data Types

FILES

- Aligned Reads 10,773
- Genome Aligned Read 10,649
- Annotated Variant Call 10,638
- Genomic Variant 10,637

CLEAR | FILE FORMAT is vcf | and STUDY NAME is Pediatric Brain Tumor Atlas... | share | Save

10,879 Files | **804** Participants | **804** Families | **433.66 GB** Size

Showing 1 - 20 of 10,879 files

Analyze in Cavatica | Download | File Manifest | Columns | Export TSV

File ID	Participants ID	dbGap	Study Name	Study Code	Proband	Family Id	Data Type	Data Category	File Format	File Size	Actions
GF_VKQWPF16	PT_VAJN5QP8		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_RDZVYSJH	Variant Calls	Simple Nucleotide Variation	vcf	889.54 KB	
GF_3JT2W7K5	PT_HMF8J4EG		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_96PFTQRJ	Annotated Variant Call		vcf	533.37 MB	
GF_HKM5MD0V	PT_FFRHWB74		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_GG0WCQY2	Masked Somatic Mutation		vcf	835.64 KB	
GF_7VW938AK	PT_XTVQB9S4		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_8YZ1P1GY	Annotated Somatic Mutation		vcf	1.52 MB	
GF_NQEAYVY2	PT_HJMP6PH2		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_Y1SF4EM1	Variant Calls	Simple Nucleotide Variation	vcf	409.33 KB	
GF_JH3P3V5W	PT_WWME595X		Pediatric Brain Tumor Atlas: CBTTC	PBTA-CBTN	Yes	FM_Z9689EQQ	Variant Calls	Simple Nucleotide Variation	vcf	861.62 KB	

Show 20 rows

<< < 1 2 3 4 5 > >>

Feedback

Export TSV with file metadata for selected samples
to further select files to be analyzed in CAVATICA

cavatica.sbgenomics.com/home

CAVATICA

Projects | Data | Public Apps | Public Projects | Developer | Controlled projects

- Data Browser
- Public Reference Files
- Public Test Files
- Volumes
- Data Tools
- Datasets

Projects

Search

PhenoDB Dev Project
Created by:d3b-bixu · May 20, 2022, 15:3

1000g_test
Created by:renan.martin · Apr 28, 2022, 8

R03
Created by:renan.martin · Mar 2, 2022, 16:52

KF X01 ODMS_BEEC_PHACE
Created by:cavatica · Jun 16, 2021, 13:57

KFDRC Sobreira Strelka2 Collab
Created by:kids-first-drc · Dec 18, 2020, 11:57

Genome-wide Sequencing to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors
Created by:kids-first-drc · May 4, 2020, 13:40

cavatica.sbgenomics.com/p/datasets

CAVATICA | Projects | Data | Public Apps | Public Projects | Developer | Controlled projects | renan.martin

Datasets

All Member Admin

Search

- M PBTA-PNOC
- M PBTA-CBTN
- MARIS_NB_XE_01
- MARIS_NB_CL_01
- SU2C_MB_PA_01
- MCGILL_DIPG_PA_01
- Chordoma Foundation Dataset

Browse datasets from your left side.
Marked with A and M are the ones that you can copy.

All Member Admin

Search

- M PBTA-PNOC
- M PBTA-CBTN
- MARIS_NB_XE_01
- MARIS_NB_CL_01
- SU2C_MB_PA_01
- MCGILL_DIPG_PA_01
- Chordoma Foundation Dataset

MEMBER **PBTA-PNOC**

DESCRIPTION

PNOC is an international consortium with study sites within the United States, Canada, Europe and Australia dedicated to bring new therapies to children and young adults with brain tumors.The Pacific Pediatric Neuro-Oncology Consortium (PNOC) is a network of over 22 children's hospitals that conduct clinical trials of new therapies for children with brain tumors. Our goal is to improve outcomes by translating the latest findings in cancer biology into better treatments for these children.

Patients with brain tumors that cannot be treated with standard therapy, or that have recurred following standard therapy, are often eligible for clinical trials. Clinical trials provide access to promising new treatments that may not be available outside specialized centers.

At PNOC, our focus is personalized medicine – testing new therapies that are specific to the biology of each patient's tumor to maximize their effectiveness. Our goal is to improve overall outcome for children with brain tumors.

Controlled Data Access

For access to the BAM, FASTQ, CRAM files and Called Germline Variants, a data access request will need to be submitted at https://redcap.chop.edu/surveys/?s=A7M873HMN8 and a signed Data Use Agreement(included on the Redcap form) will be required. Please email reserach@cbtn.org for additional details.

MEMBERS

- cavatica OWNER — Write, Copy, Admin
- yuankun — Write, Copy
- zhangb1 — Write, Copy
- gaonkark — Write, Copy
- victorlu247 REQUESTED — Read
- ennisb — Write, Copy

Leave dataset

Files

Search

Case ID: All | Sample ID: All | Experimental strategy: All | +

Copy

Name	Case...	Sample ID	Sample ty...	Primary s...	Gender	Experimental
harmonized-data	-	-	-	-	-	-
source-data	-	-	-	-	-	-

Files

Search

Case ID: All | Sample ID: All | Experimental strategy: All | +

Copy

Name	Case...	Sample ID	Sample ty...	Primary s...	Gender	Experimental
harmonized-data	-	-	-	-	-	-
source-data	-	-	-	-	-	-

Files

Search

Case ID: All | Sample ID: All | Experimental strategy: All | +

Copy

Search projects

Projects

- PhenoDB Dev Project
- 1000g_test
- R03
- KF X01 ODMS_BEEC_PHACE
- KFDRC Sobreira Strelka2 Collab
- Single gene pathogenic variants associated with BEEC (Bladder extrophy, Epispadias, Complex)

Ollier disease and Maffucci syndrome BHCMG-CMG Program - AnVIL

Access germline WES data from 33 probands with Ollier disease and Maffucci syndrome

Status

- To be accessed

gen3.theanvil.io

Submit Data | Documentation | rmart120 | Logout

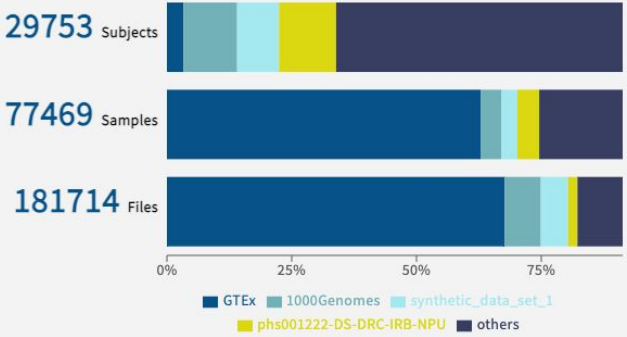
The AnVIL

Dictionary | Exploration | Workspace | Profile

The AnVIL

The AnVIL supports the management, analysis and sharing of human disease data for the research community and aims to advance basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cancer. The data commons supports cross-project analyses by harmonizing data from different projects through the collaborative development of a data dictionary, providing an API for data queries and download, and providing a cloud-based analysis workspace with rich tools and resources.

Submit Data



Define Data Field

Explore Data

Analyze Data

gen3.theanvil.io/explorer

Submit Data | Documentation | rmart120 | Logout

The AnVIL

Dictionary | Exploration | Workspace | Profile

Data | File | Downloadable

Explorer Filters | Data Tools | Summary Statistics | Table of Records

Data Access

- Data with Access
- Data without Access
- All Data

Export to Seven Bridges | Export All to Terra | Export to PFB | Export to Workspace

Subjects
3,202

Projects
1

Filters

Projects | Subject | Sample | Sequencing

Collapse all

Project Id 1 selected

- open_access-1000Genomes 3,202
- tutorial-synthetic_data_set_1 2,504

Anvil Project Id

- no data 3,202

Project dbGaP Accession Number

- open 3,202

Sex

Female
1,271
(39.7%)

Male
1,233
(38.5%)

no data
698
(21.8%)

Ancestry

no data

Showing 1 - 20 of 3,202 subjects

Show Empty Columns

Project Id	Sex	Samples Count	Sequencings Count

Export to Seven Bridges | Export All to Terra

- Export to CGC
- Export to CAVATICA
- Export to BDC (Seven Bridges)

Subjects
3,202

Sex

Female

cavatica.sbgenomics.com/import/pfb?URL=https:%2F%2Fgen3-theanvil-io-pfb-export.s3.amazonaws.com...

CAVATICA | Projects | Data | Public Apps | Public Projects | Developer | Controlled projects | renan.martin

Importing data

You are about to import data from Gen3 **anvil** as DRS files with associated metadata. The data will be imported via PFB file.
Learn more

Destination project

No project selected

Or Create new project

Resolve naming conflicts

Skip

Add tags

Type to search...

☑ I understand that data accessible via DRS, including but not limited to controlled-access data, may be subject to terms and conditions of acceptable use, and I confirm that I am only importing data in accordance with any applicable terms of use, including but not limited to my obligations under any applicable Data Use Agreements. Furthermore, I understand that I am importing a PFB file which may contain controlled access data and I confirm that I am solely responsible for managing access to this file since no other mechanisms protect this file in any way and the data could be accessed by other users in this project.

Import data

Terms | Privacy | Copyright

Destination project

1000g_test

PhenoDB Dev Project

1000g_test

R03

KF X01 ODMS_BEEC_PHACE

KFDRC Sobreira Strelka2 Collab

Genome-wide Sequencing to Identify th

Single gene pathogenic variants associa

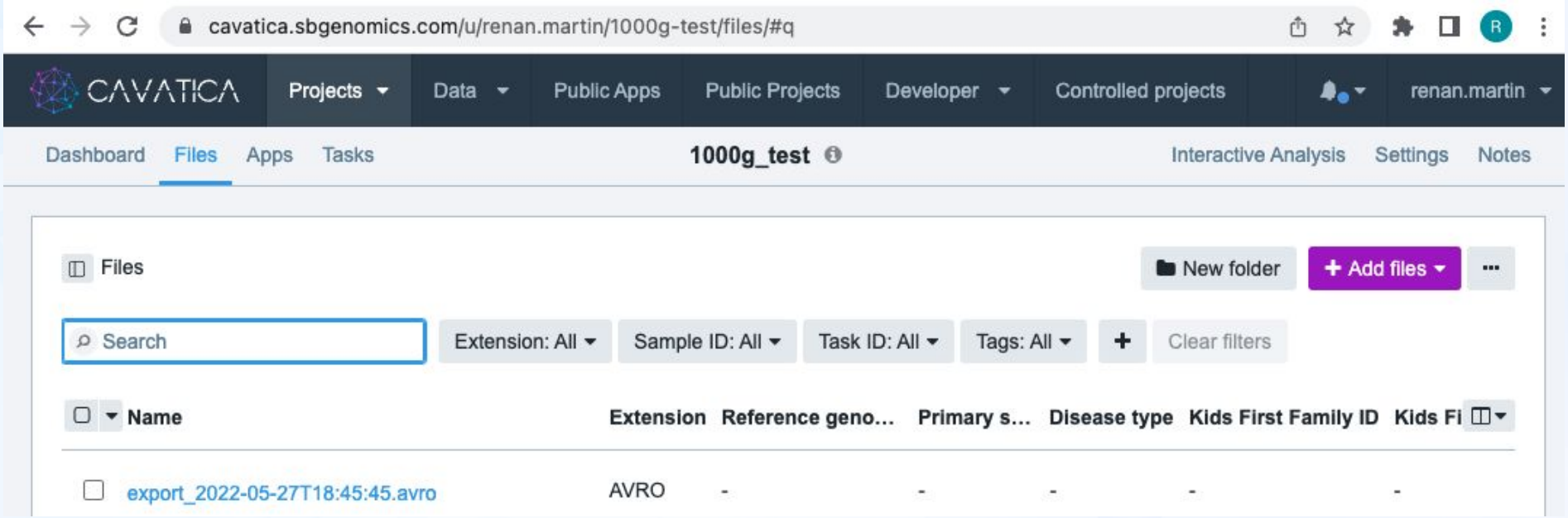
GMKF: Genomic Analysis of a Cohort wi

that I am only importing data in accordance

with any applicable terms of use, including

☑ I understand that data accessible via DRS, including but not limited to controlled-access data, may be subject to terms and conditions of acceptable use, and I confirm that I am only importing data in accordance with any applicable terms of use, including but not limited to my obligations under any applicable Data Use Agreements. Furthermore, I understand that I am importing a PFB file which may contain controlled access data and I confirm that I am solely responsible for managing access to this file since no other mechanisms protect this file in any way and the data could be accessed by other users in this project.

Import data



First the AVRO file will be displayed on Files Tab of the target Project

cavatica.sbgenomics.com/u/renan.martin/1000g-test/files/#q

CAVATICA | Projects | Data | Public Apps | Public Projects | Developer | Controlled projects | renan.martin

Dashboard | Files | Apps | Tasks | 1000g_test | Interactive Analysis | Settings | Notes

Files | New folder | + Add files

Search | Extension: All | Sample ID: All | Task ID: All | Tags: All | + | Clear filters

Name	Extension	Reference genome	Primary site	Disease type	Kids First Family ID	Kids First Biospecimen ID	Kids First Participant ID
DRS NA18567.haplotypeCalls.er.raw.g.vcf.gz	VCF.GZ	-	-	-	-	-	-
DRS HG02127.final.cram	CRAM	-	-	-	-	-	-
DRS HG04019.final.cram	CRAM	-	-	-	-	-	-
DRS HG01840.final.cram	CRAM	-	-	-	-	-	-
DRS HG01138.haplotypeCalls.er.raw.vcf.gz.tbi	TBI	-	-	-	-	-	-
DRS NA18876.final.cram.crai	CRAI	-	-	-	-	-	-
DRS HG00234.haplotypeCalls.er.raw.g.vcf.gz.tbi	TBI	-	-	-	-	-	-
DRS HG04061.final.cram.crai	CRAI	-	-	-	-	-	-
DRS HG00323.final.cram.crai	CRAI	-	-	-	-	-	-
DRS HG00332.haplotypeCalls.er.raw.g.vcf.gz.tbi	TBI	-	-	-	-	-	-
DRS HG01167.final.cram.crai	CRAI	-	-	-	-	-	-
DRS HG01523.haplotypeCalls.er.raw.vcf.gz	VCF.GZ	-	-	-	-	-	-

Refresh | Showing 1-100 of 13010

Then, the AVRO file will be replaced by imported files once import finishes

Next Steps

Access Ollier disease and Maffucci syndrome files from BHCMG with CAVATICA

- Once the access on AnVIL/Gen3 is granted, we will be able to export (access) to CAVATICA via Seven Bridges (function already tested with open datasets)

Access chondrosarcoma files from NCI GDC Portal with CAVATICA

Acknowledgments

- Nara Sobreira' lab
  - Renan Martin
  - Elizabeth Wohler
  - Eliete Rodrigues
  - Corina Antonescu
  - Carolina Montano
- Kim Doheny
  - Sean Griffith
  - Laura Vail
- NIH - NCPI
  - Asiyah Lin
- Seven Bridges
  - Jack Digiovanna
- NIH – NCI
  - Jay Ronquillo
  - Erika Kim
- Broad Institute
  - Ruchi Munshi
  - Rachel Liao
- Funding - NIH – NHGRI and NCI

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

NCPI Working Group Updates

11:50 AM - 1:05 PM EDT

Community Governance WG

Bob Grossman (University of Chicago)
Stanley Ahalt (University of North Carolina at Chapel Hill)

General Framework

- The NCPI Community / Governance Working Group is not charged with coming up with specific policies or recommendations.
- Instead, this group is charged with coming up with
  - associated use cases and questions that help frame the fundamental governance questions;
  - concepts and frameworks to support interoperability for the use cases;
  - Key questions for the community consensus.
- We summarize the key questions, associated frameworks, and community consensus in technical papers.

Phase 1 - Viewing NCPI Platforms following NIST 800-53 (or other approved frameworks) as Authorized Environments

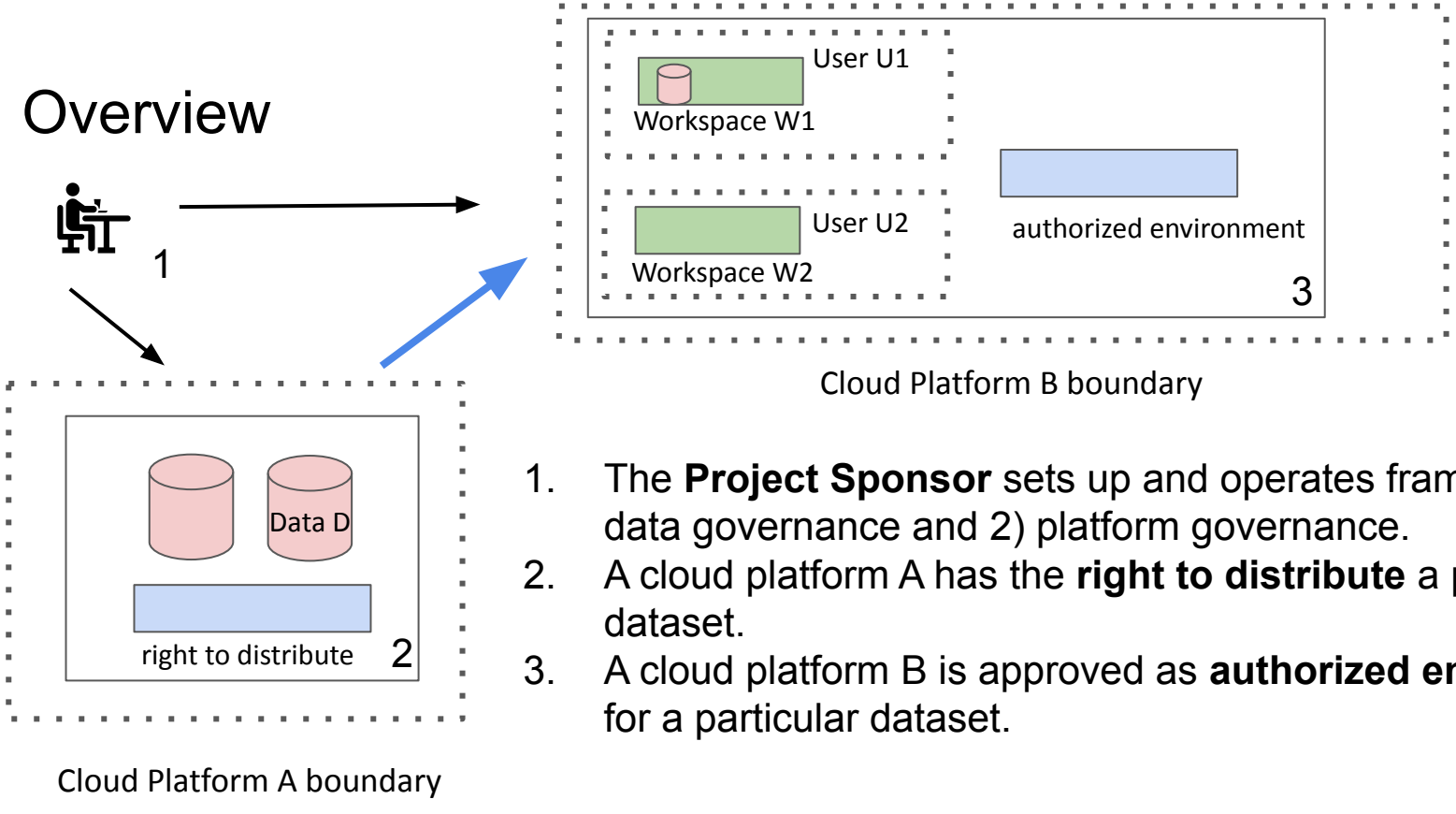
Key Concepts

Project Sponsor - Entity responsible for data and platform governance.

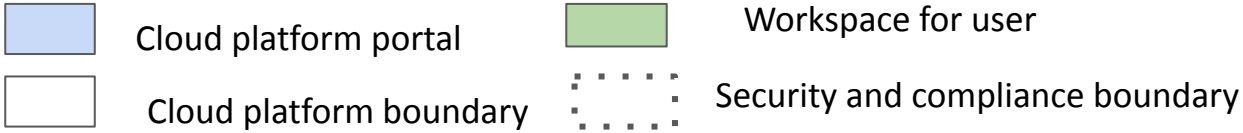
Right to distribute - the project sponsor determines whether the source cloud platform has the right to distribute a particular dataset

Authorized environment - the project sponsor determines whether the target cloud platform has appropriate security, compliance and governance to support the analysis of the data on the cloud platform by authorized researches

Overview



1. The **Project Sponsor** sets up and operates frameworks for 1) data governance and 2) platform governance.
2. A cloud platform A has the **right to distribute** a particular dataset.
3. A cloud platform B is approved as **authorized environment** for a particular dataset.



Computer Science > Distributed, Parallel, and Cluster Computing

[Submitted on 10 Mar 2022]

A Framework for the Interoperability of Cloud Platforms: Towards FAIR Data in SAFE Environments

Robert L. Grossman, Rebecca R. Boyles, Brandi N. Davis-Dusenbery, Amanda Haddock, Allison P. Heath, Brian D. O'Connor, Adam C. Resnick, Deanne M. Taylor, Stan Ahalt

As the number of cloud platforms supporting biomedical research grows, there is an increasing need to support interoperability between two or more cloud platforms. A well accepted core concept is to make data in cloud platforms findable, accessible, interoperable and reusable (FAIR). We introduce a companion concept that applies to cloud-based computing environments that we call a Secure and Authorized FAIR Environment (SAFE). SAFE environments require data and platform governance structures. A SAFE environment is a cloud platform that has been approved through a defined data and platform governance process as authorized to hold data from another cloud platform and exposes appropriate APIs for the two platforms to interoperate.

Comments: 11 pages with 1 figure and a 2 page appendix
Subjects: **Distributed, Parallel, and Cluster Computing (cs.DC)**
ACM classes: D.2.11; D.2.12; E.0
Cite as: arXiv:2203.05097 **[cs.DC]**
(or arXiv:2203.05097v1 **[cs.DC]** for this version)
https://doi.org/10.48550/arXiv.2203.05097

Status

- Community consensus and agreement on key concepts and framework
- Technical paper completed and published on arXiv
- Selected interoperability approved for selected datasets between pairs of NCPI Cloud Platforms
- No general guidelines yet about interoperability between 2 or more NCPI Platforms

Potential Next Steps

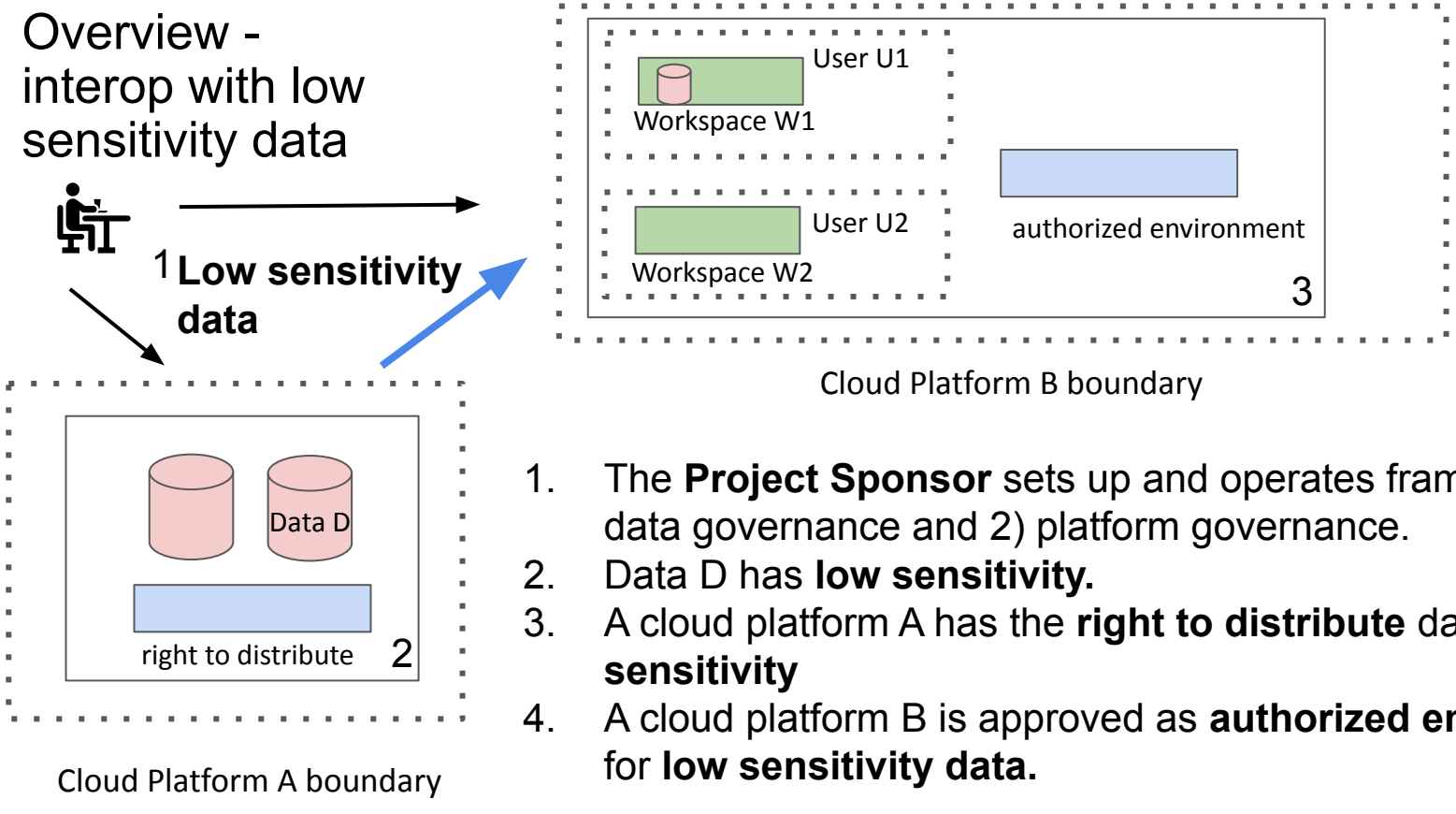
- Seek approval for the current NCPI Platforms as authorized environments for data from one of the other NCPI Platforms.
- Seek approval for selected other platforms that follow NIST 800-53 Moderate as authorized environments for one or more NCPI platforms.

Phase 2 - Interop for Low Sensitivity Data

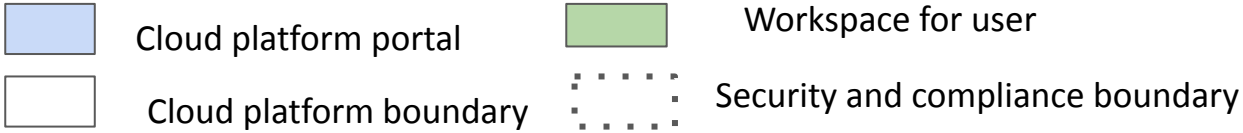
Basic Idea

- Not all data in current NCPI platforms are equally sensitive
- Today, controlled access genomic data is classified is usually housed in cloud platforms that FISMA Moderate.
- For less sensitive data, such as as certain aggregate or summary data level data, perhaps we can classify as less sensitive (call it low sensitivity) data and approved in cloud platforms that are are FISMA Low or approved for CUI, for example.

Overview - interop with low sensitivity data



1. The **Project Sponsor** sets up and operates frameworks for 1) data governance and 2) platform governance.
2. Data D has **low sensitivity.**
3. A cloud platform A has the **right to distribute** data that is **low sensitivity**
4. A cloud platform B is approved as **authorized environment** for **low sensitivity data.**



Controlled Unclassified Information (CUI)

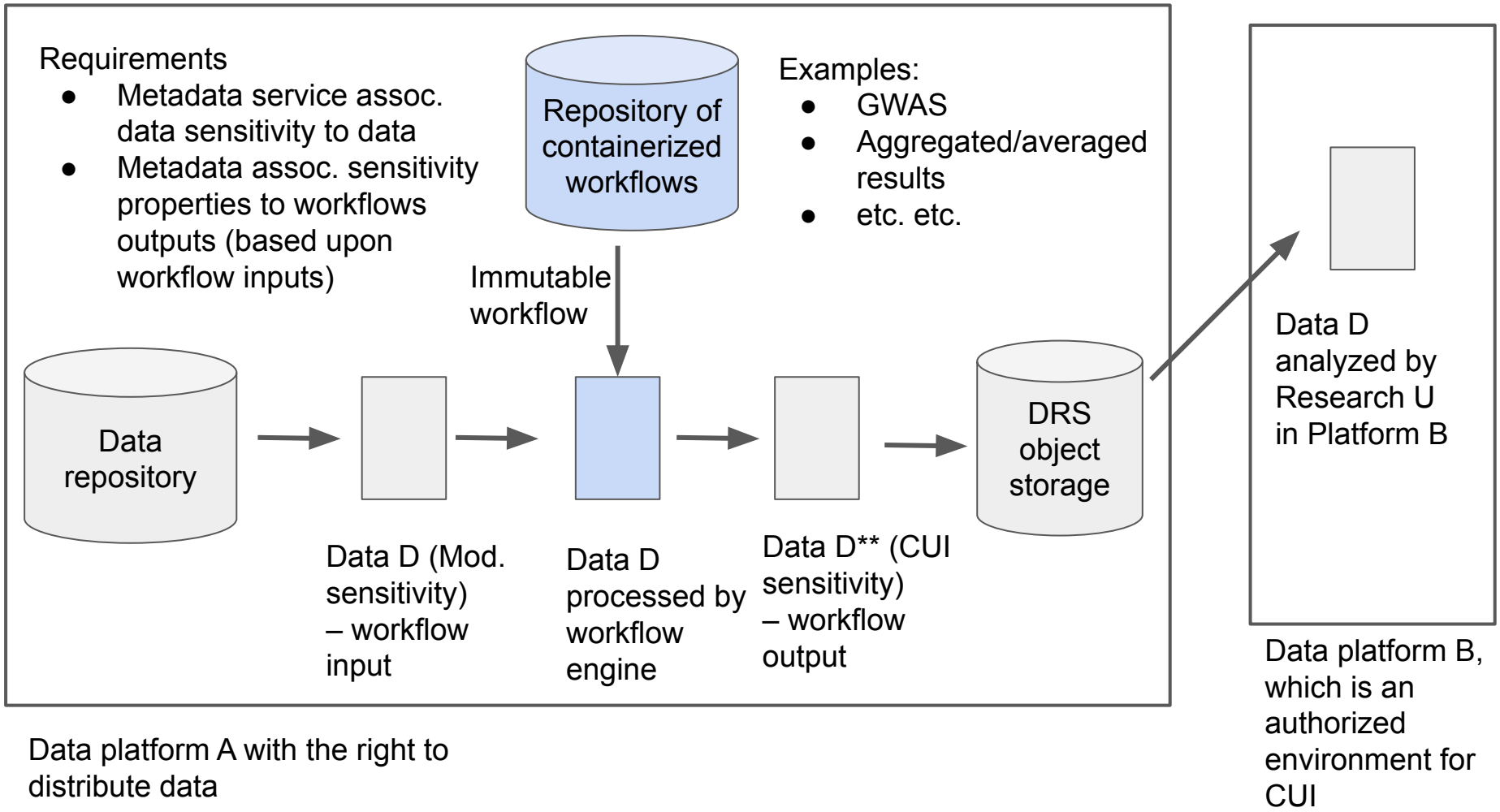
NIST Special Publication 800-171
Revision 2

Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations

RON ROSS
VICTORIA PILLITTERI
KELLEY DEMPSEY
MARK RIDDLE
GARY GUISSANIE

- CUI
- Follows NIST 800-171
- Can be used for less sensitive data

A very simple use case of low sensitivity data being generated by applying approved workflows to genomic data.


Requirements
Metadata service assoc. data sensitivity to data
Metadata assoc. sensitivity properties to workflows outputs (based upon workflow inputs)
Repository of containerized workflows
Examples:
GWAS
Aggregated/averaged results
etc. etc.
Immutable workflow
Data repository
Data D (Mod. sensitivity) – workflow input
Data D processed by workflow engine
Data D** (CUI sensitivity) – workflow output
DRS object storage
Data D analyzed by Research U in Platform B
Data platform A with the right to distribute data
Data platform B, which is an authorized environment for CUI

Questions

- If there is a data or security incident, when data is transferred from one cloud platform to another, who is responsible when there is a security or data management event or incident?
  - The platform that receives the data?
  - As determined by the platform sponsor?
  - As determined by the Interconnection Security Agreement?
  - The platform that sends the data?
  - It depends upon the specifics of the event or incident?
  - In practice, it depends upon whether the sponsor of the target platform is another Institute or Center?
  - Some combination of the above?
- Answering these questions conservatively, has essentially slowed down access to the data by the research community from cloud platforms, despite the fact that the current cloud platforms tend to operate under higher levels of security and compliance.

Evaluating Risks

OWASP PROJECTS CHAPTERS EVENTS ABOUT Search OWASP.org Store Donate Join

OWASP Risk Rating Methodology

Watch 120 Star 671

DISCLAIMER

Over the years there has be lots of debate about the OWASP Risk Rating Methodology and the weighting of Threat Actor Skill levels. There are other more mature, popular, or well established Risk Rating Methodologies that can be followed:

- NIST 800-30 - Guide for Conducting Risk Assessments
- Government of Canada - Harmonized TRA Methodology
- Mozilla resources:
  - Risk Assessment Summary
  - Rapid Risk Assessment (RRA)

Alternatively you may with the review information about Threat Modeling, as that may be a better fit for your app or organization:

- https://owasp.org/www-community/Threat_Modeling
- https://owasp.org/www-community/Application_Threat_Modeling
- OWASP Threat Dragon

The OWASP® Foundation works to improve the security of software through its community-led open source software projects, hundreds of chapters worldwide, tens of thousands of members, and by hosting local and global conferences.

Upcoming OWASP Global Events

OWASP April Webinar

- April 12-13, 2022 UTC/GMT +3 hours

OWASP 2022 Global AppSec Europe

- The Open Web Application Security Project (OWASP) is an online community that produces freely-available articles, methodologies, documentation, tools, and technologies in the field of web application security. The Open Web Application Security Project provides free and open resources.
- NIST 800-30 also provides framework
- and several others are widely used

Sources: https://owasp.org/www-community/OWASP_Risk_Rating_Methodology

Risk

risk = risk impact * likelihood of risk

- Impact (also called risk impact) defines ‘how bad’ things can get, the worst-case scenario. Impact is primarily based upon the data.
- Likelihood defines the probable frequency, or rate at which the impacts we assessed may occur. Likelihood on the other hand is primarily driven by the presence or absence of security controls in the service.

Sources: https://owasp.org/www-community/OWASP_Risk_Rating_Methodology

https://infosec.mozilla.org/guidelines/assessing_security_risk

Some Risks

1. Honest but curious person downloads the data and exposes it through unintentional misuse.

2. Uses unsigned code that's a "look alike" docker that exfils the data

3. Data is modified through a bug and not detected

4. Other risks….

Sources: David Bernick email, discussion in previous NCPI Community / Governance WG call

Risks in the Context of Use Case 1

#	Risk	Use Case 1	Comment
1	Honest but curious person downloads the data and exposes it through unintentional misuse.	Data is aggregated sufficiently that risk of re-identification is quite low	
2	Uses unsigned code that's a "look alike" docker (like what's happening with NPM libraries now and supply chains) that exfils the data	Workflow is signed and data platform service executes workflow (vs user executing workflow)	
3	Data is modified through a bug and not detected	Risk is present whether data is analyzed in Platform A or egressed to Platform B	
4	**Other risks**		

Questions / Discussion

Systems Interoperation WG

Jack DiGiovanna (Seven Bridges)

Why is interoperability important for NIH?

NIH NATIONAL CANCER INSTITUTE

1-800-4-CANCER Live Chat Publications Dictionary

ABOUT CANCER CANCER TYPES RESEARCH GRANTS & TRAINING NEWS & EVENTS ABOUT NCI search

Home > Research > Annual Plan & Budget Proposal > Highlighted Scientific Opportunities

ANNUAL PLAN & BUDGET PROPOSAL

- Director's Message
- Budget Proposal
- Stories of Cancer Research
- Driving Discovery
- Highlighted Scientific Opportunities
  - Clinical Trials
  - Computer-Based Drug Design
  - Precision Prevention
  - Tumor Dynamics

Precision Prevention: Predicting and Intercepting Your Cancer

Imagine if we were able to determine an individual's cancer risk by characterizing their genetic makeup, family history, environmental exposures, and behavioral factors and then tailor personalized prevention approaches based on these factors. To achieve this, we need

Image credit: https://www.cancer.gov/research/annual-plan/scientific-topics/precision-prevention

ELSEVIER

Biochimica et Biophysica Acta (BBA) - Reviews on Cancer

Volume 1876, Issue 1, August 2021, 188573

The potential of AI in cancer care and research

Norman E. Sharpless M.D., Anthony R. Kerlavage Ph.D.

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.bbcan.2021.188573 Get rights and content

Abstract

Current applications of artificial intelligence (AI), machine learning, and deep learning in cancer research and clinical care are highly diverse—from aiding radiologists in reading medical images to predicting oncoprotein folding and dynamics. The list of available AI-based tools is growing rapidly and will only continue to expand. With the immense potential for AI to advance cancer

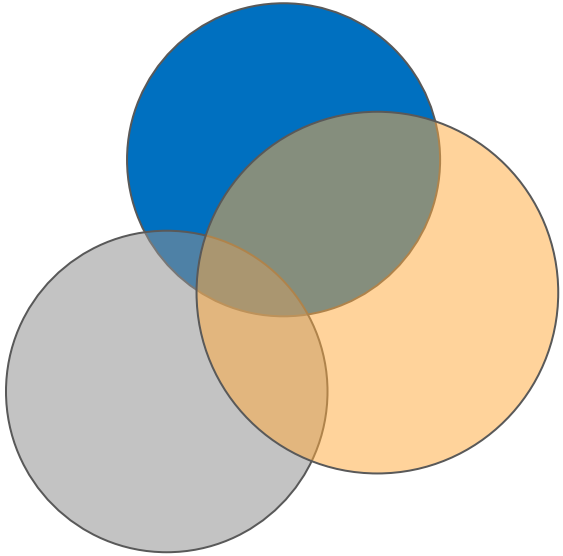
Image credit: https://www.sciencedirect.com/science/article/abs/pii/S0304419X21000706

System Interoperability mission to empower diverse researchers

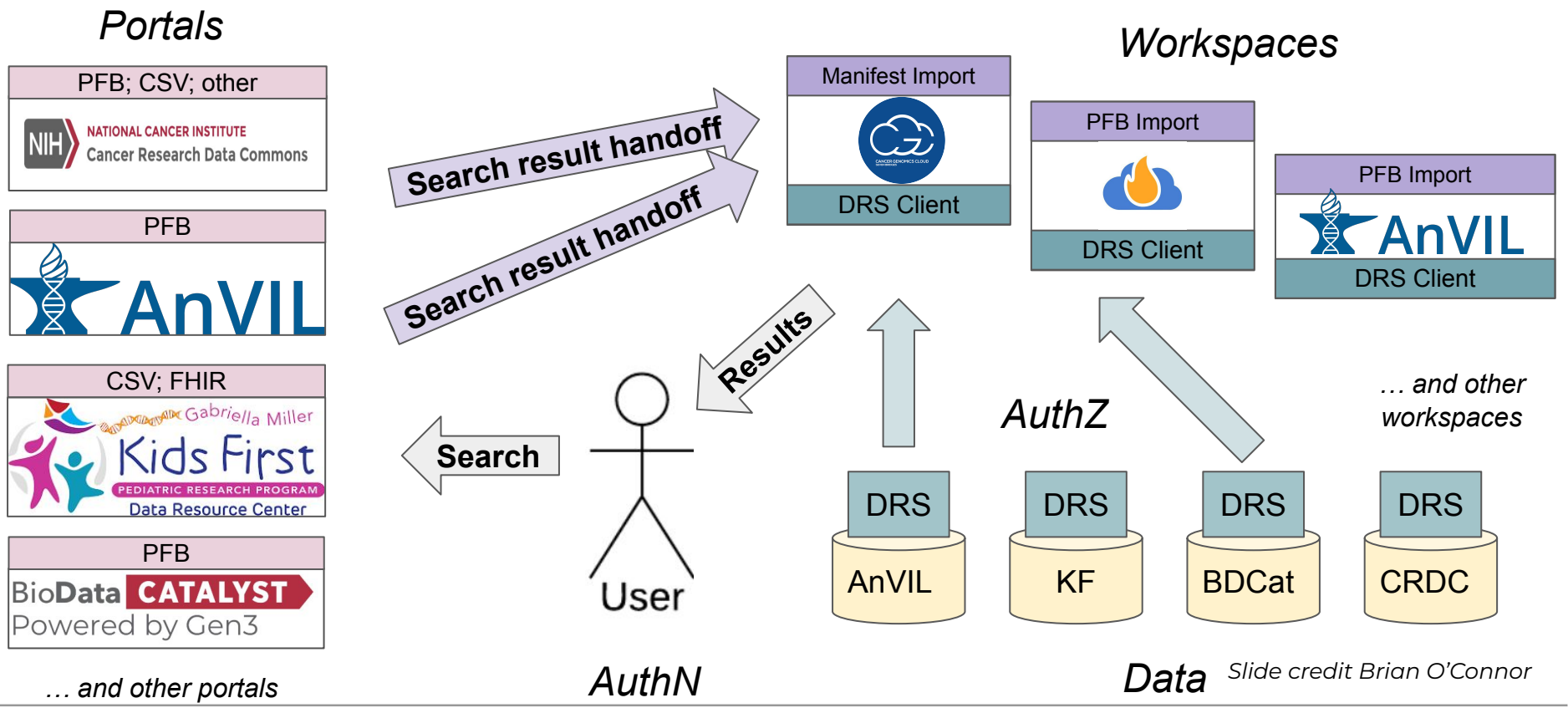
Empower **diverse researchers** to complete **scientific projects** across ICs by spearheading **technical improvements** across cloud "stacks”

Sys Interop is part of the researcher journey

Coordination	Valentina Di Francesco (NHGRI) & Ken Wiley (NHGRI)
Community Governance	Stanley Ahalt (RENCI) & Bob Grossman (UChicago)
Systems Interoperation	**Brian O'Connor** (Sage Bionetworks) & Jack DiGiovanna (Seven Bridges)
Outreach + Training	Stephen Mosher (JHU)
FHIR	Robert Carroll (Vanderbilt) & Allison Heath (CHOP)
Search	Dave Rogers (Clever Canary) & Kathy Reinold (Broad)



Helps users analyze scientifically-relevant data



Slide credit Brian O'Connor

Early CRDC-AnVIL “use-case” recently published in PNAS

Wilson McKerrow, David Fenyö, et al

Cloud costs funded via Collaborative Project

CRDC

CANCER GENOMICS CLOUD
SEVEN BRIDGES

AnVIL

GEN3 DATA COMMONS

Terra

PNAS

RESEARCH ARTICLE | SYSTEMS BIOLOGY | FULL ACCESS

LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint

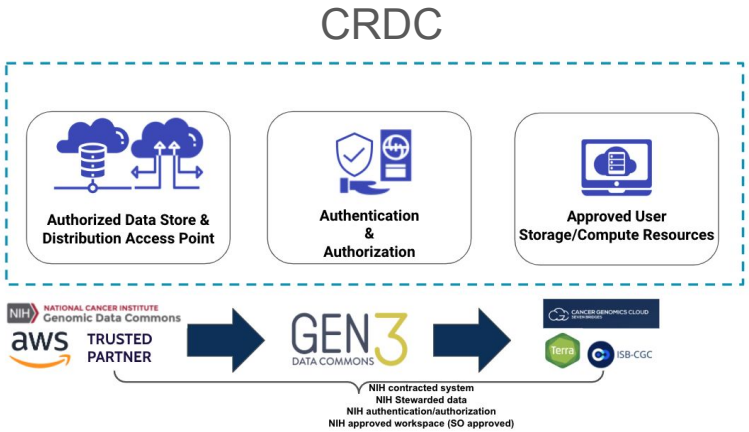
Wilson McKerrow, Xuya Wang, Carlos Mendez-Dorantes, +7, and David Fenyö Authors Info & Affiliations

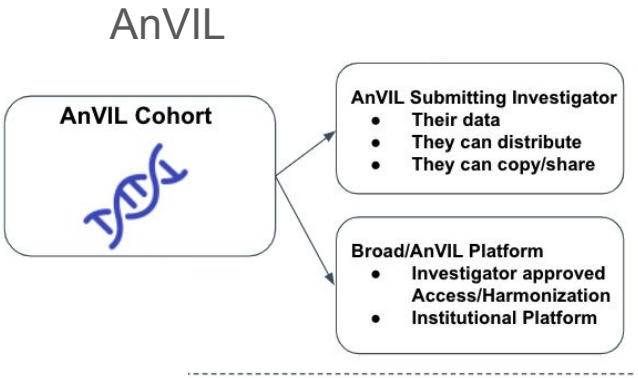
February 15, 2022 | 119 (8) e2115999119 | https://doi.org/10.1073/pnas.2115999119

Significance

In addition to canonical genes, our genomes encode repetitive copies of the LINE-1 retrotransposon. These elements duplicate themselves by cutting a single-strand break in genomic DNA and then reverse transcribing a new LINE-1 DNA copy into that breakpoint. In most contexts, LINE-1 elements are epigenetically repressed, but they are dramatically

NCPI is trailblazing interoperability policy as well

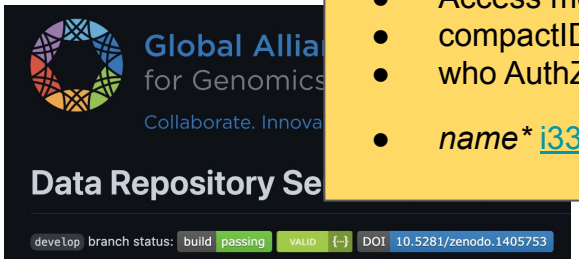




Together we've made it easier for the next researcher

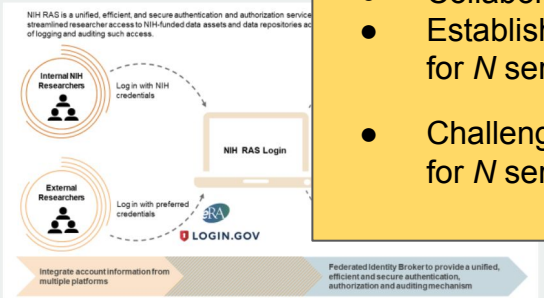
Agreed on a finite set of technical methods

Object access



- Access method i213
- compactIDs pr369
- who AuthZ pr381
- *name** i335

AuthN/Z



NIH Researcher Auth Service 1.0: Conceptual Overview

- Collaborating with NIH RAS
- Establishing *N* mTLS certs for *N* servers
- Challenge: *N* user passports for *N* servers

Manifests (PFB or CSV)

Mi data attributes

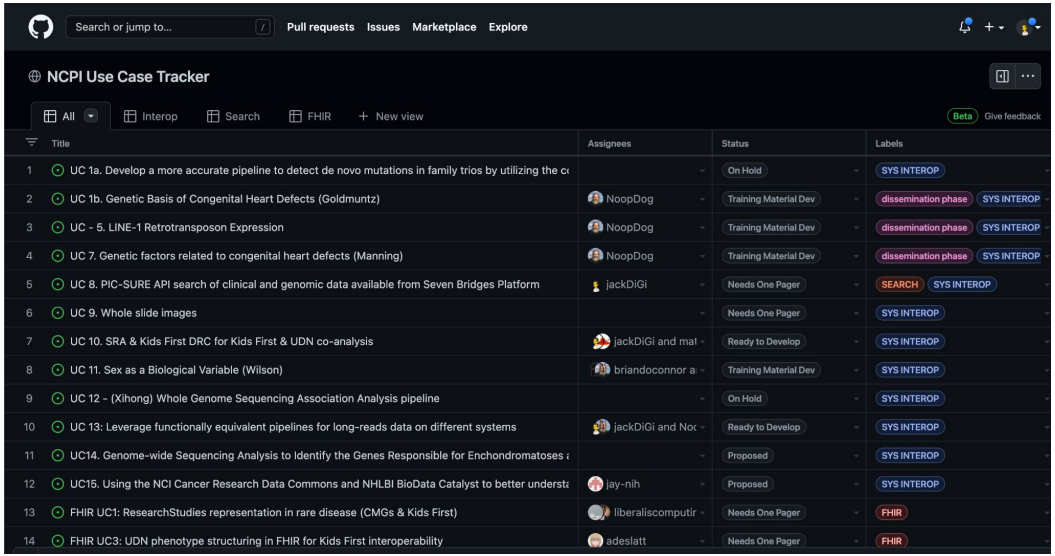
Attribute	Definition
drs_uri	DRS URI as defined by GA4GH DRS spec for pointers to file objects.
study_registration	External source from which the identifier included in study_id originates (answer can be dbGaP for example)
study_id	Unique identifier that can be used to retrieve more information for a study
participant_id	Unique identifier that can be used to retrieve more information for a participant
specimen_id	Unique identifier that can be used to retrieve more information for a specimen
experimental_strategy	The experimental strategy used to generate the data file referred to by the ga4gh_drs_uri. (Based on GDC definition)
file_format	The format of the data, see possible values from the data_format fields in GDC. Can use whatever values make sense for the particular implementation.
fhir_document_reference	optional fhir url pointing to the FHIR Document Reference, if metadata available on a FHIR Server
`file_name`	*The name of the file the DRS URI is pointing to.*

Defining minimal criteria has dramatically improved use cases

All use cases require a one-pager on a **public github repo**

Ensure that the this info is **agreed** upon:

- Platforms Involved
- Scientific question
- Science Lead & Platform Lead
- Interop/Tech Plan
- Funding Plan



Credit to Dave Rogers and Asiyah Lin

https://github.com/orgs/NIH-NCPI/projects/1/views/6

Two use cases presented earlier

Sex chromosome complement aware alignment

Brendan Pinto and Melissa Wilson

Genome-wide Sequencing Analysis to Identify the Genes Responsible for Enchondromatoses and Related Malignant Tumors

Renan Martin
Nara Sobreira
Johns Hopkins University School of Medicine

Happy to see how things have progressed

Early in this effort, our working group were *traveling salesmen* for interop, methods, etc

Funding & roadmap management was also *very challenging*

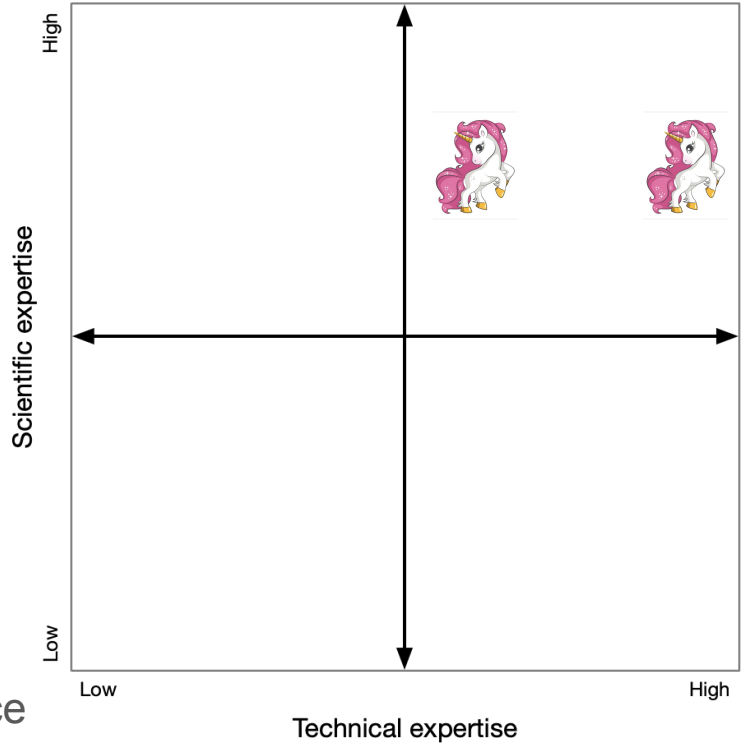
Use cases are now **publishing** manuscripts

More interop is happening

- **CFDE, RADx, INCLUDE**
- **Tools, Datasets**

Tech and policy are hardening to **reduce barriers** to science

User personas



Summary

Thank you for NIH ODSS's support and partnership for NCPI

Reusing developed components, *improving* the "use-case" process, and the *community helping* each other will **increase speed to results**

Researchers can analyze select **CRDC, TOPMed, Kids First**, and **AnVIL** data

Want to **build awareness & adoption to grow the ecosystem;** also need to optimize **strategy** - please connect us with the latest **researcher challenges**

Learn more @ https://anvilproject.org/ncpi

Lively Discussion

FHIR WG

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

Overview

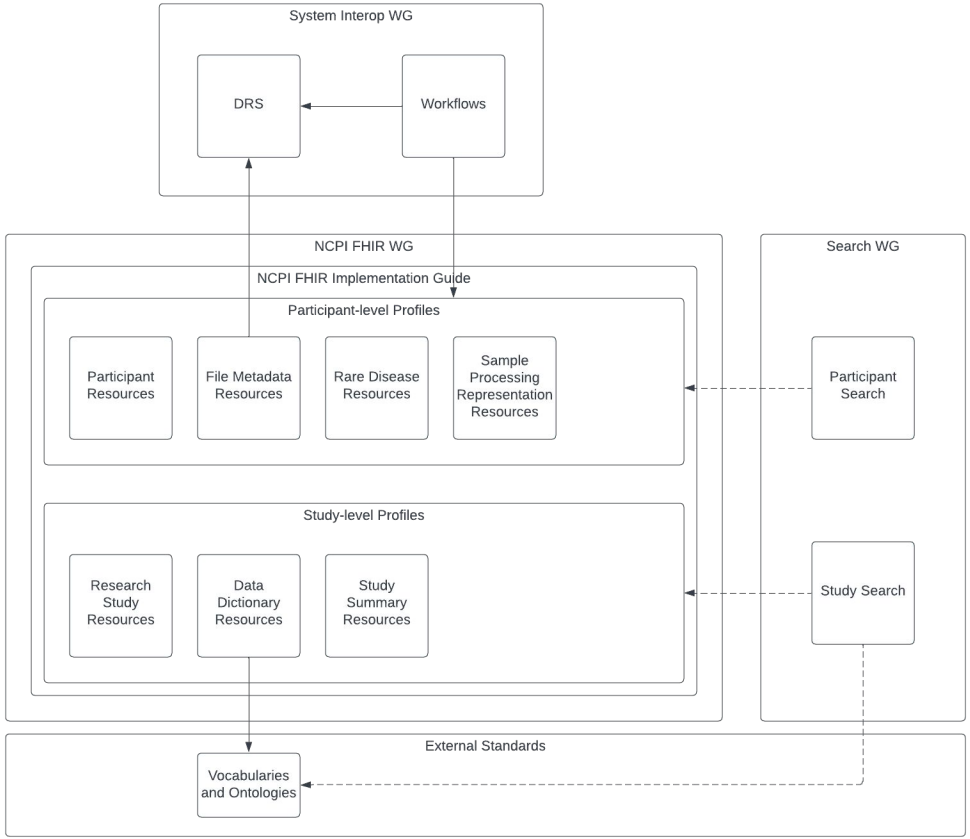
- Objectives for FHIR
- FHIR Service Deployment
- FHIR Implementation Guide v0.1 Complete
- Refactoring our approach- IG v0.2
- FHIR Code-a-thon next week!

Objectives of FHIR

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

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

Objectives of FHIR

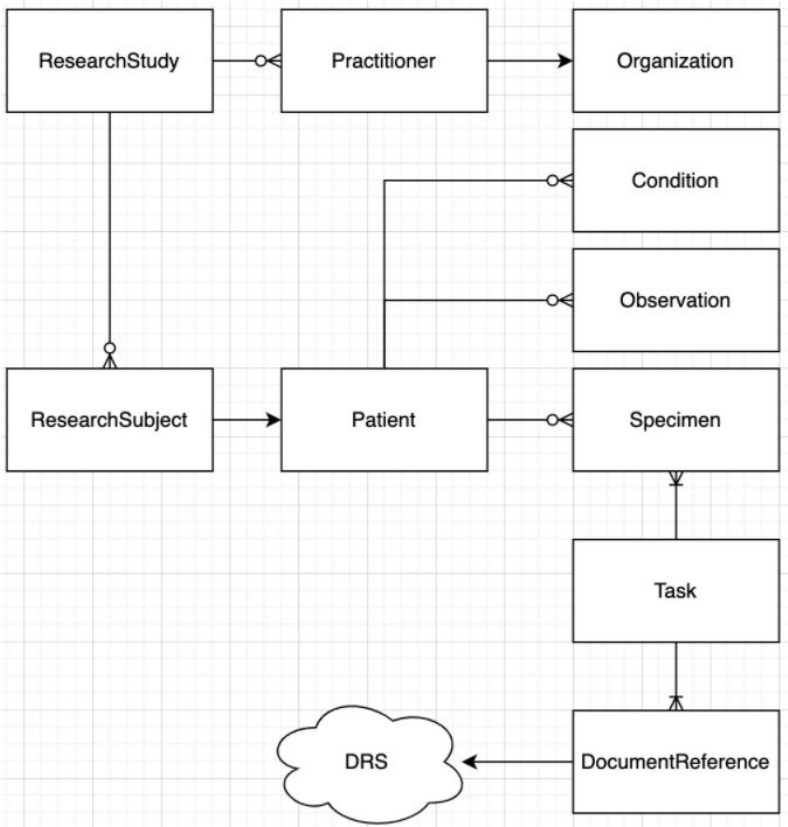


FHIR Service Deployment

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

Implementation Guide v0.1

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



Interoperable Data Services

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

Study Information

Diseases and Syndromes

Laboratory Measures

File Metadata

Participant Demographics

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.

Participant Details

Participant Demographics

Rare Disease

Diseases and Syndromes

Phenotypic Features

Research Study Metadata

Study Information

Data Dictionaries

Consent Groups

Variable Reports

‘Omics data

DRS References

Genomic Assays

File Metadata

IG v0.2

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

FHIR Code-a-thon

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

FHIR Code-a-thon

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

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

NCPI Outreach WG

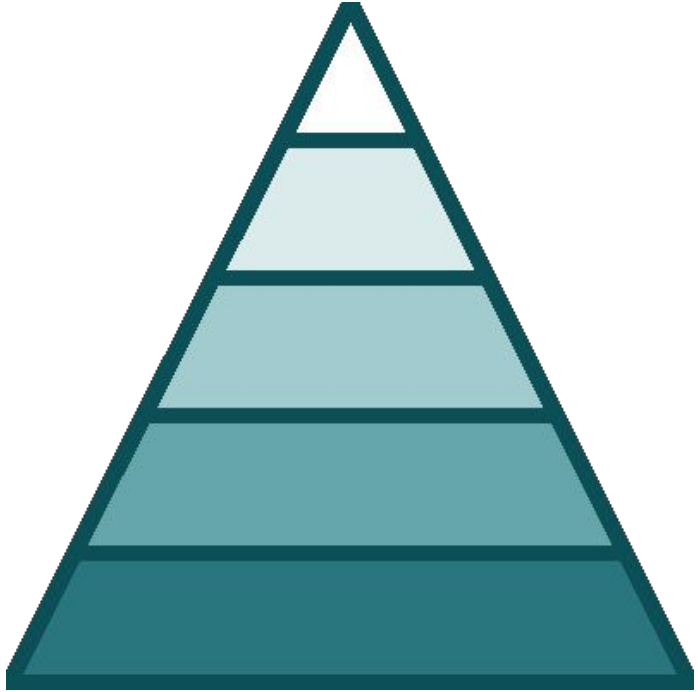
Stephen Mosher (Johns Hopkins University)

NCPI Outreach WG Mission

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

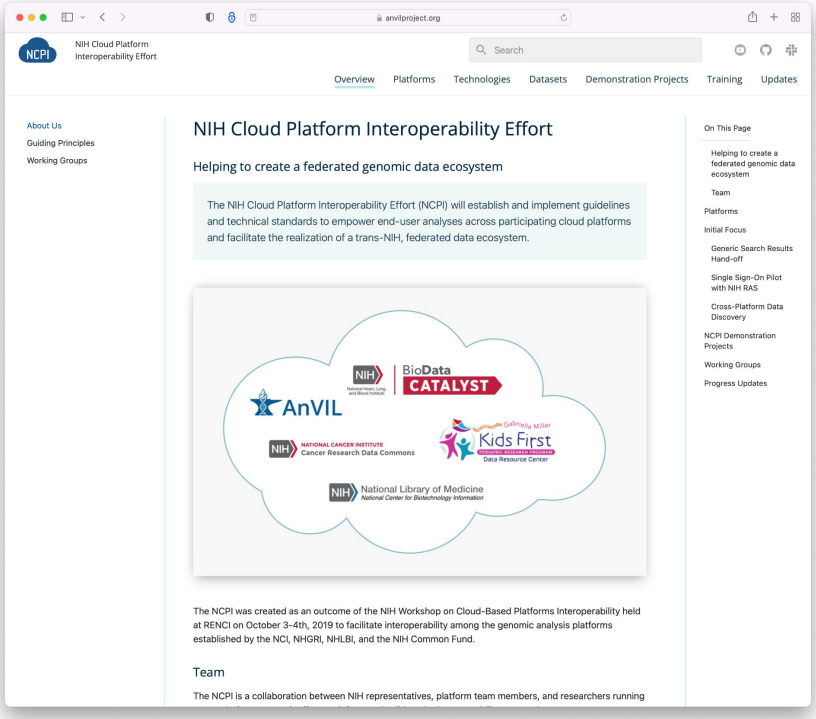
Goals

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

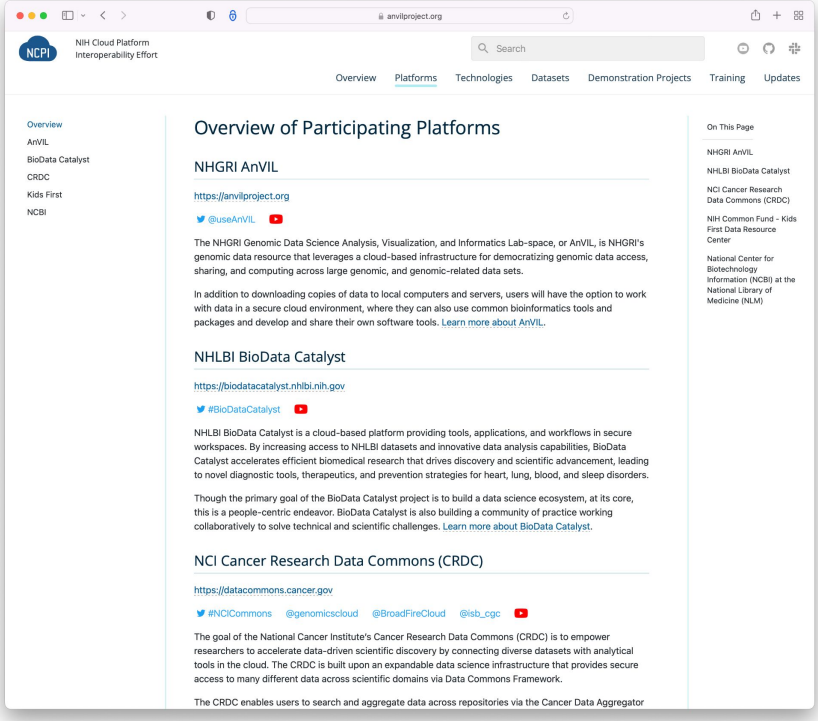


NCPI Portal

https://anvilproject.org/ncpi



Participating platforms



NCPI Portal

Technologies enabling science

anvilproject.org

NCPI — NIH Cloud Platform Interoperability Effort

Search

Overview | Platforms | Technologies | Datasets | Demonstration Projects | Training | Updates

Interoperation Technologies

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 eRA 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 informatoin on the most recent version of this API (1.1) see the Data Repository Service 1.1 Documentation.

On This Page

- Researcher Auth Service (RAS)
- Data Repository Service (GA4GH DRS)
- Fast Healthcare Interoperability Resources (FHIR)
- Portable Format for Bioinformatics (PFB)

The science driving the tech

anvilproject.org

NCPI — NIH Cloud Platform Interoperability Effort

Search

Overview | Platforms | Technologies | Datasets | Demonstration Projects | Training | Updates

NCPI Interop Demonstration Projects

- Genetic Basis of Congenital Heart Defects (Goldmuntz)
- LINE-1 Retrotransposon Expression (McKerrow)
- Genetic Factors Related to Congenital Heart Defects (Manning)
- Sex as Biological Variable (Wilson)

NCPI Interoperability 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 ...

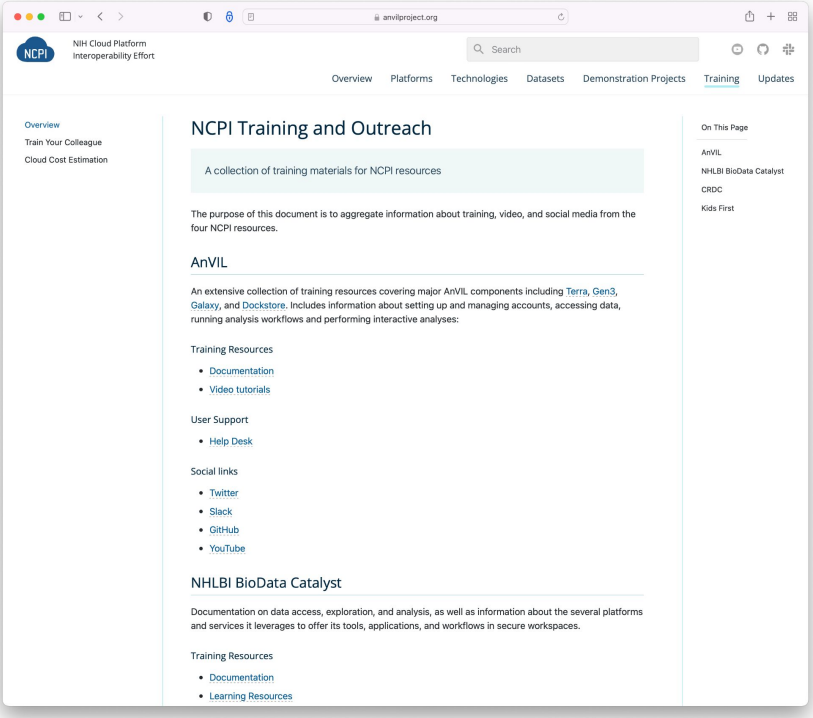
Sex as a Biological Variable (Wilson)

On This Page

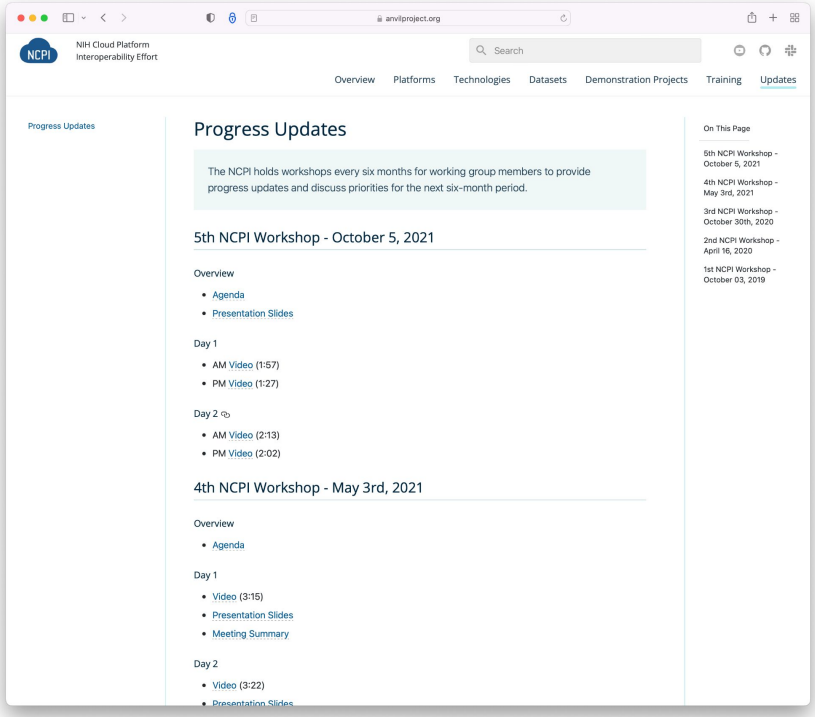
- Genetic Bases of Congenital Heart Defects (Goldmuntz)
- LINE-1 Retrotransposon Expression (McKerrow)
- Genetic Bases of Congenital Heart Defects (Manning)
- Sex as a Biological Variable (Wilson)

NCPI Portal

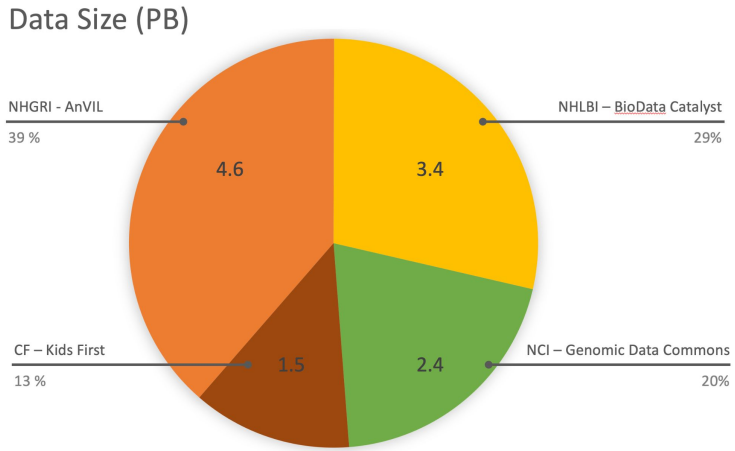
Aggregating outreach resources



Past workshop resources



NCPI Dataset Catalog



Researcher Auth Service

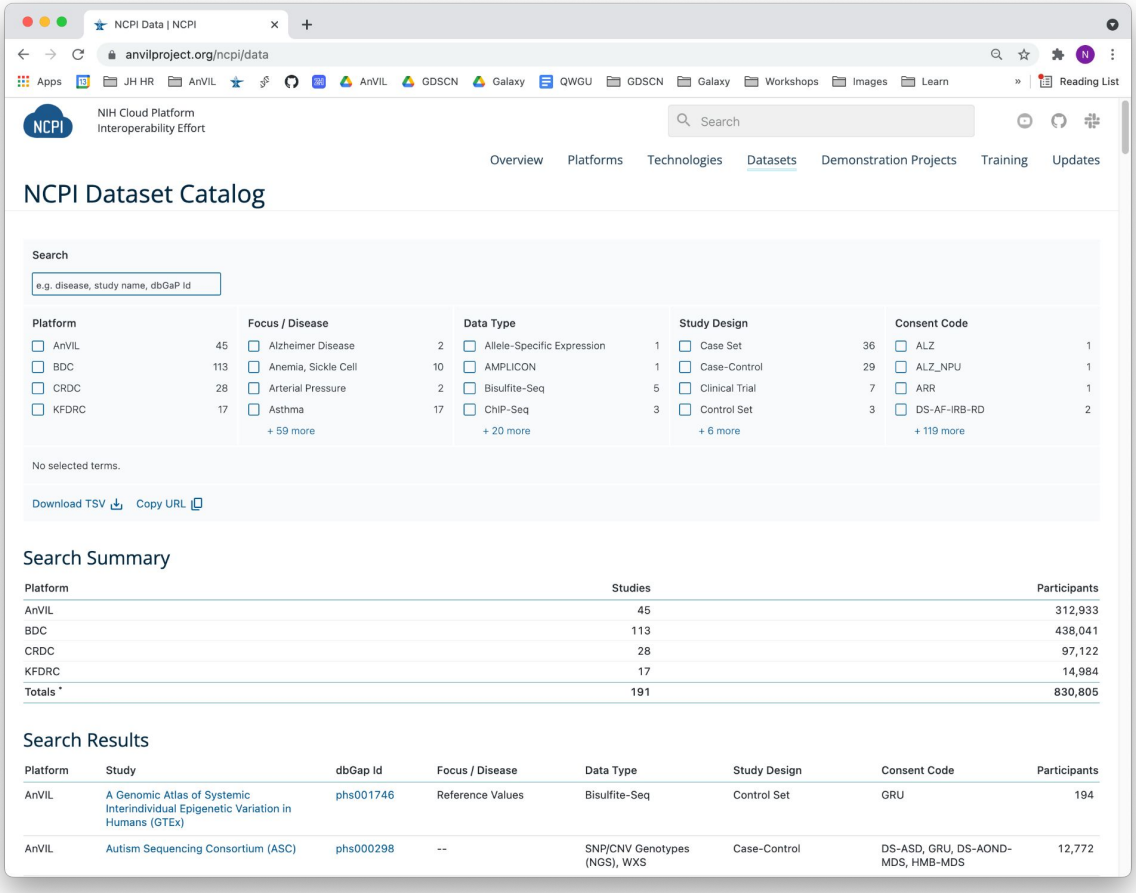
Data Repository Service

Fast Healthcare Interoperability Resources

12Pb / 830k participants and growing!

Cross-platform accessibility through several key technologies

Dataset Search (more details from Search WG)



Search by:

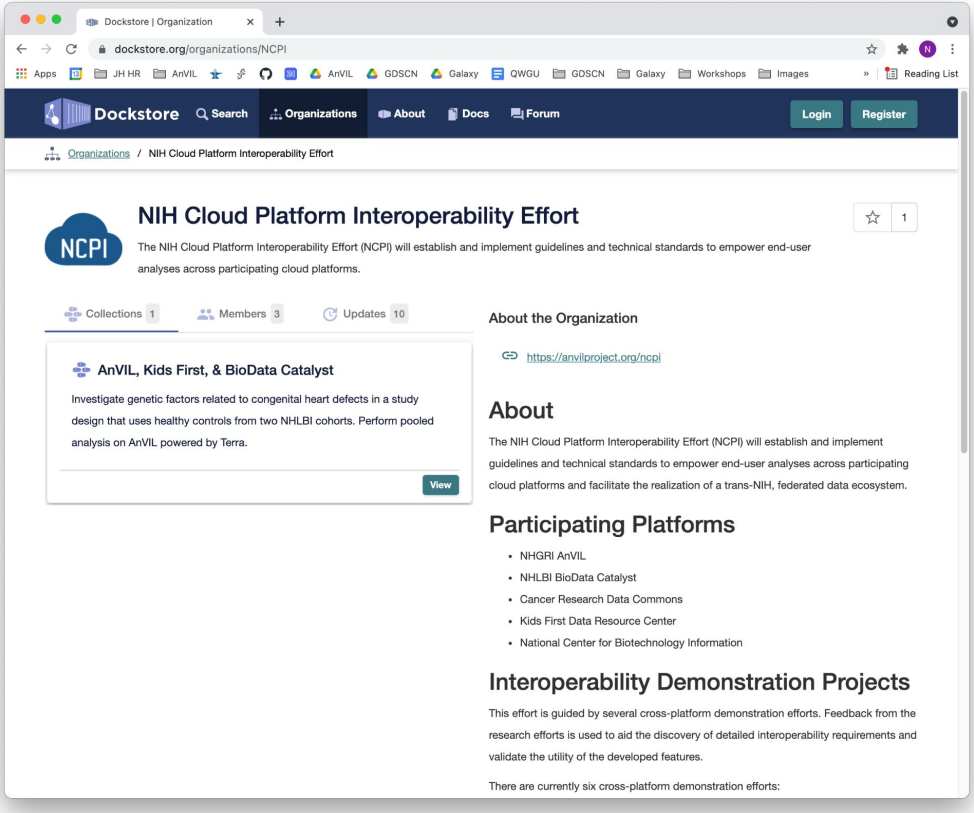
- Platform
- Focus or Disease
- Data type
- Study Design
- Consent Code

Budded off into new Search Working Group

Dockstore Organization for NCPI

Promoting FAIR practices in tool and workflow sharing

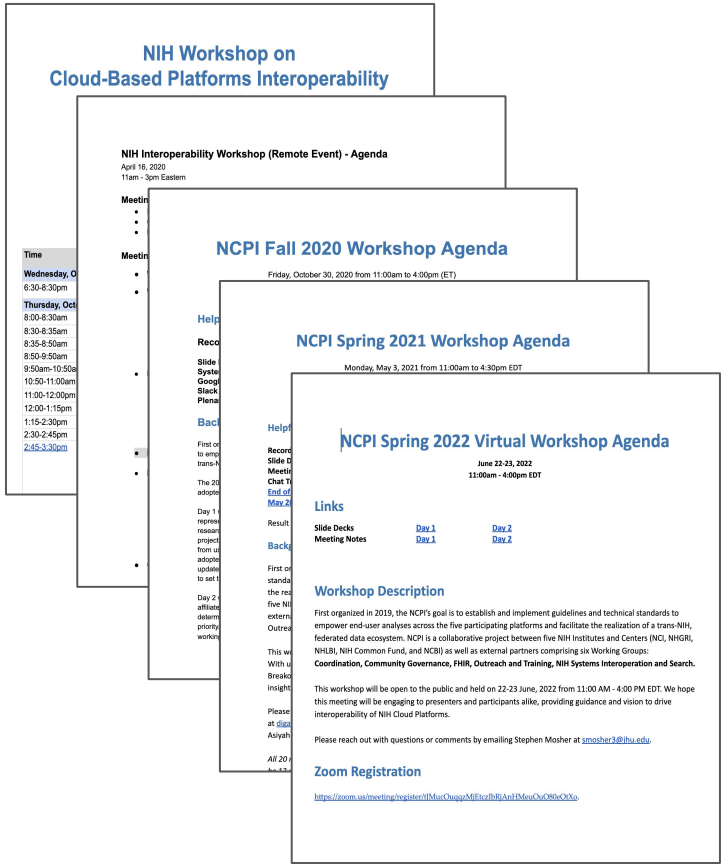
- Findable
- Accessible
- Interoperable
- Reusable



https://dockstore.org/organizations/NCPI

Supporting NCPI Workshops

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

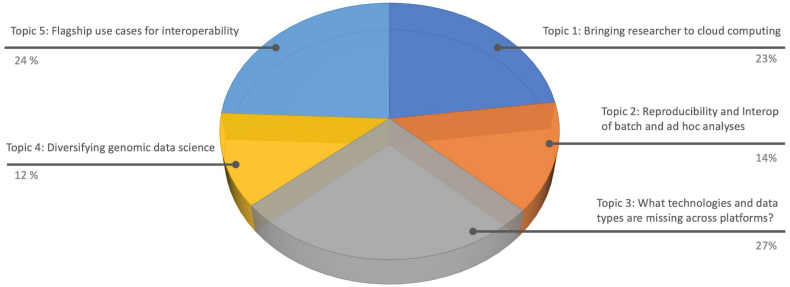


Today's Virtual Workshop

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

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

Breakout 2



Future: Administrative Coordinating Center (ACC)

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
Bethesda, Maryland 20892

www.nih.gov

March 16, 2022

Research Opportunity Announcement

Research Opportunity Title: NIH Cloud Platform Interoperability Administrative Coordinating Center

OTA-22-004

Participating Organization(s): National Institutes of Health

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

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

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

Search WG

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

Overview

- Mission, Vision, Strategy
- Search Use Cases

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

- Next Steps
- Discussion

Mission

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

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

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

See the NCPI Search Group Charter

Vision

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

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

Target Search Use Cases / Modalities

Support search of studies and datasets across platforms by:

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

Strategy

- Be driven by researcher scientific use-cases.

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

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

ODSS Search RFI Response Overview

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

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

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

RFI Response Overview

Specific recommendations included:

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

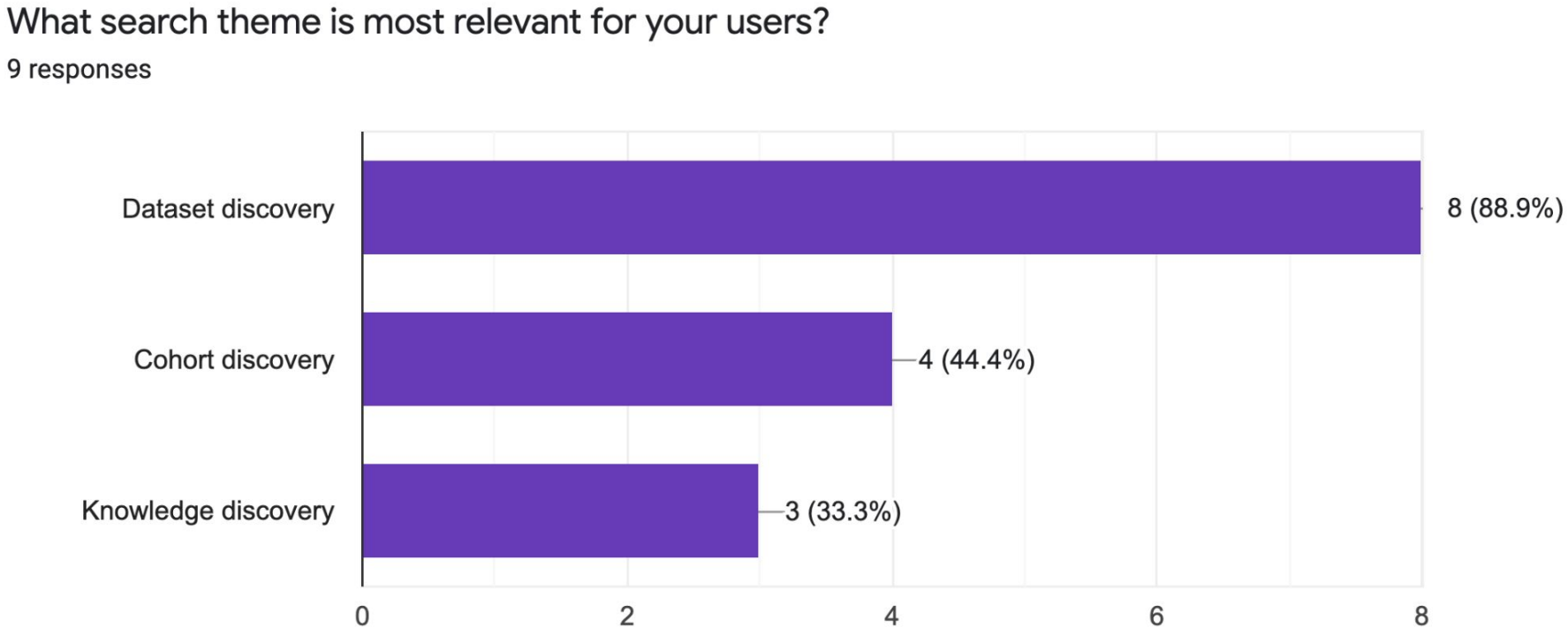
See the NCPI Search RFI Response.

Landscape Survey

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

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

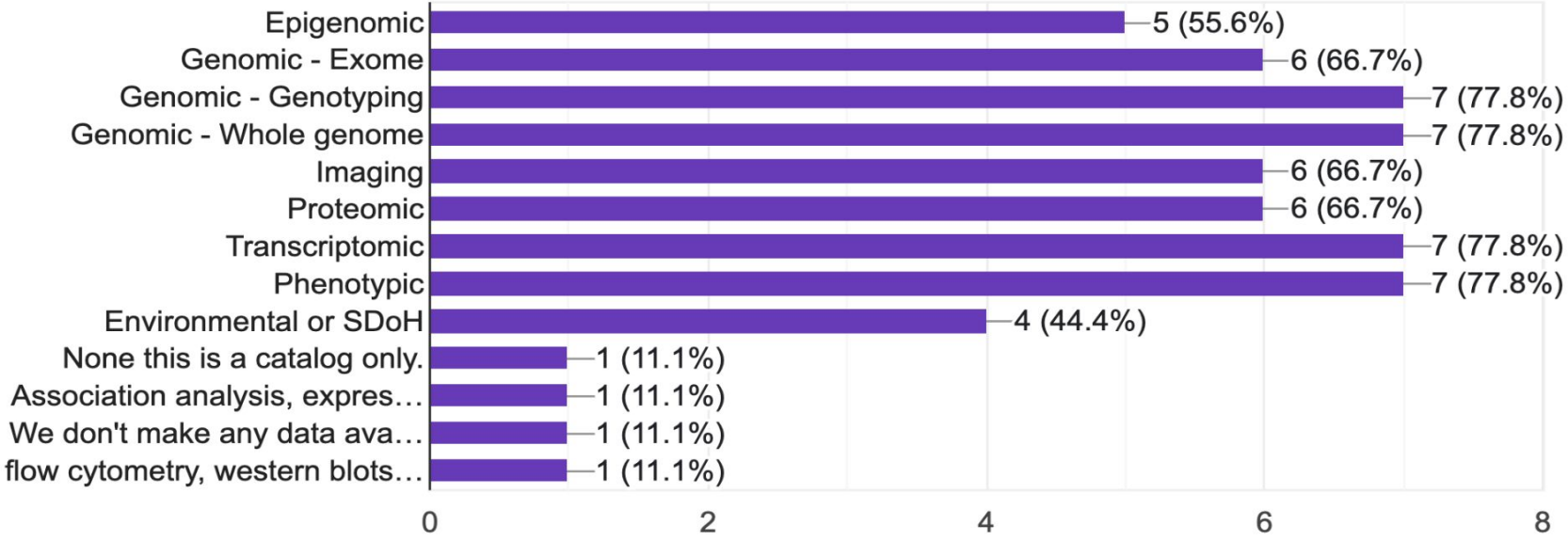
Landscape Survey - Theme



Landscape Survey - Data Modalities

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

9 responses



Landscape Survey - Phenotype Standards

- Summary
  - Most reference ontologies
  - Clearly some variation

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

Landscape Survey - Standards

Genotype Standards

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

Other Data Standards

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

Landscape Survey - Standards

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

Landscape Survey - Key Points

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

Landscape Survey - Challenges

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

Landscape Survey - Next Steps

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

- Publish the survey results on the NCPI Portal.

Demonstration Projects

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

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

See the NCPI Use case Tracker

Next Steps

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

Questions/Discussion?

Break

1:05 PM - 1:35 PM EDT

Technical Aspects of Interoperability

1:35 PM - 2:35 PM EDT

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

Dan Stanzione (TACC)

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

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

Cloud Platform Interoperability Workshop

June 2022

TACC - 2021

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

THE CHARGE FOR THIS TALK:

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

- (not necessarily in that order).

TACC AT A GLANCE - 2021

Personnel

185 Staff (~90 PhD)

Facilities

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

Systems and Services

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

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

Usage

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

WHAT WE DO

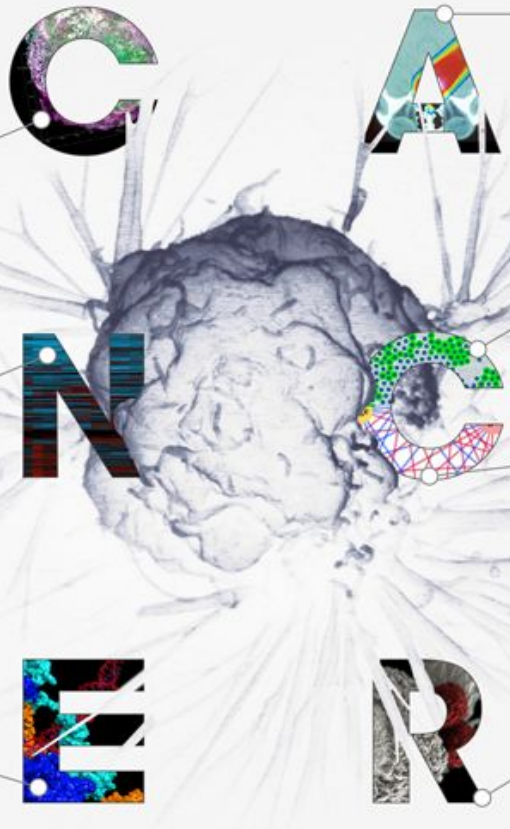
- Provide researchers with:
  - Computing, Data, AI , Software capabilities to support their research
  - The expert help to be able to use it!
  - In the ways they want to consume it
  - Help with grants/strategy
- Computation, AI, Data almost ubiquitous acro
  the sciences.

SYSTEMS UPDATES
A QUICK REMINDER ON OUR CURRENT MAJOR SYSTEMS

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

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

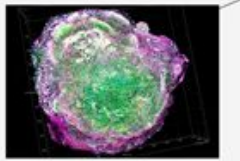
fighting



with supercomputers

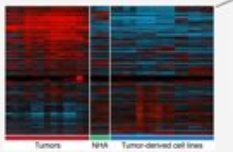
Computer Modeling

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



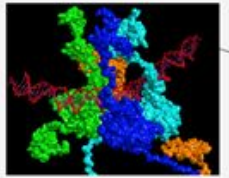
Big Data Analysis

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



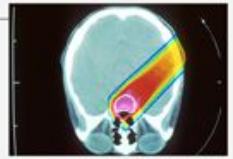
Molecular Dynamics Simulations

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



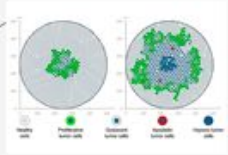
Quantum Calculations

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



Trial Design

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

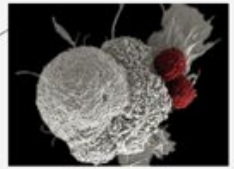


Clinical Planning

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

Artificial Intelligence

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

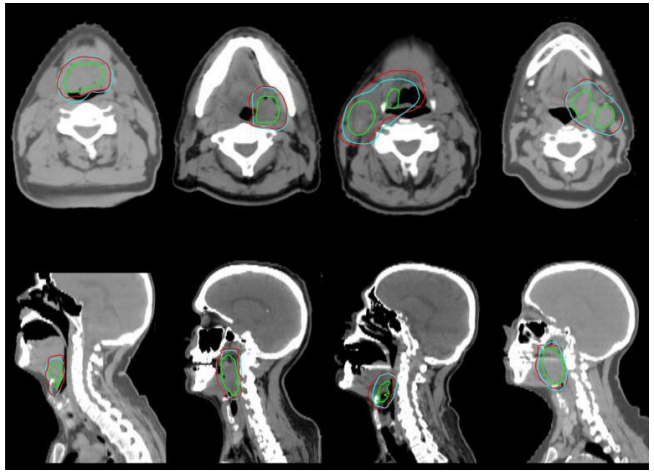


TACC
TEXAS ADVANCED COMPUTING CENTER

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

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

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



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

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

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

TACC

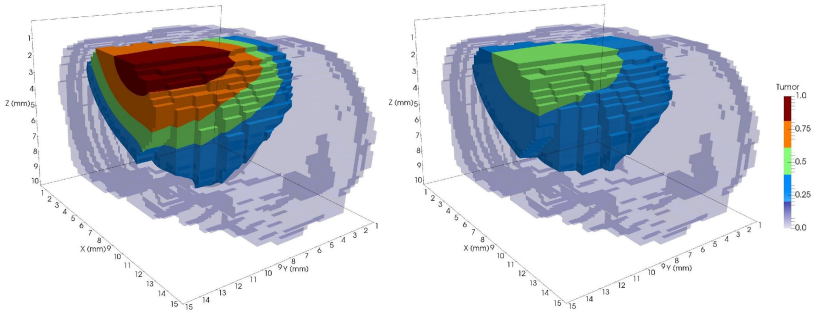
https://www.tacc.utexas.edu/-/an-ai-oncologist-to-help-cancer-patients-worldwide

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center

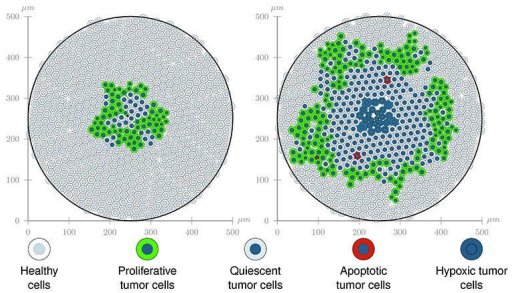
Complex **computer models** and **analytic tools** to predict how cancer will progress in a specific individual

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

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



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



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

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

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

TACC

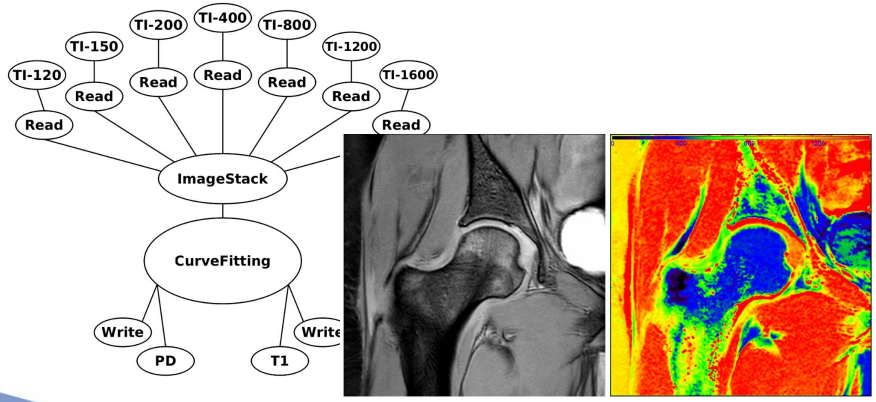
https://www.tacc.utexas.edu/-/tailoring-cancer-treatments-to-individual-patients

The University of Texas at Austin
Dell Medical School

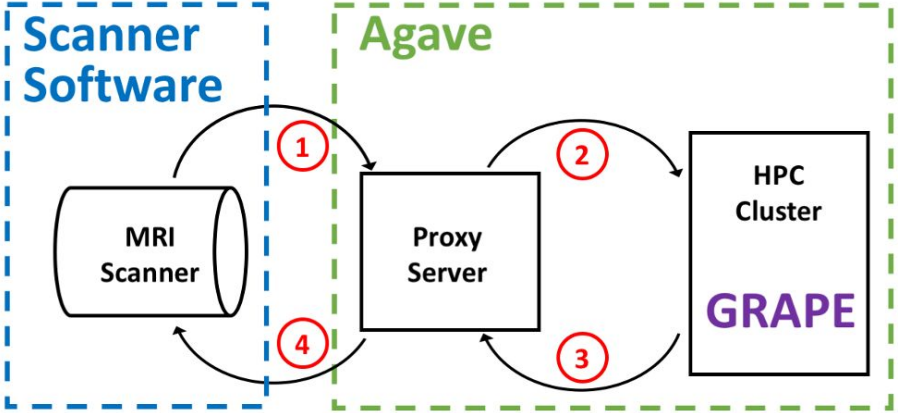
TAPIS and Jetstream enabled automated, real-time, quantitative magnetic resonance imaging

- **Problem:** Quantitative analysis of MR images is typically performed after the patient has left the scanner. Corrupted or poor quality images can result in patient call backs, delaying disease intervention.

- **Importance:** Real-time analytics of MRI scans can enable same-session quality control, reducing patient call backs, and it can enable precision medicine.



Quantitative calculations performed during scan session



Platform to automate analysis tied to HPC resources

- **Approach:** Dr. Refaat Gabr (UTHealth) and Dr. Joe Allen (TACC) used the CyVerse SDK and Agave to help develop an automated platform for real-time MRI,

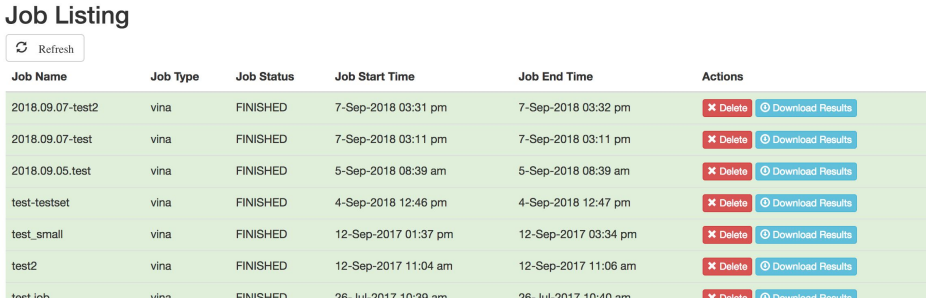
- **Result:** Scan data can now be automatically processed on high performance computing resources in real-time with no human intervention.

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

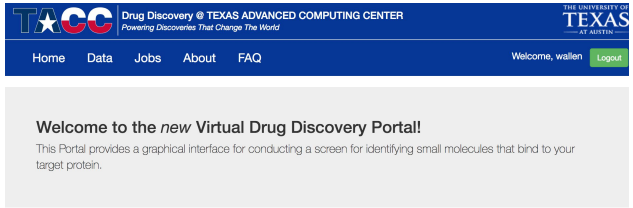
The Drug Discovery Portal empowers researchers worldwide to perform virtual screens on TACC HPC resources

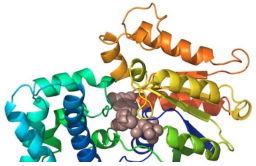
- **Problem:** While *virtual screening* has compelling advantages over experimental methods alone, it requires high-performance computational resources, software licenses, and technical expertise, which may be unattainable for small academic labs.

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



Job outputs are available for download in a web interface





The DrugDiscovery@TACC web portal

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

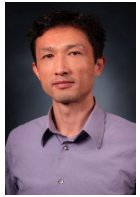
- **Result:** Users upload proteins of interest into a friendly web interface, choose a ZINC library to screen, and results are returned typically within 24 hours. The efforts have led to dozens of documented drug candidate hits.

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

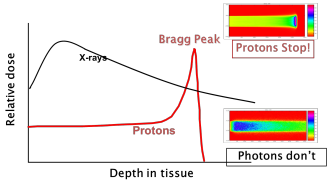
Particle/Proton Therapy Translational Research Platform

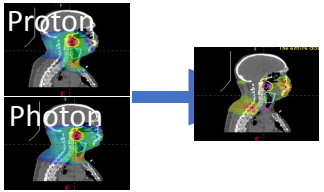
Xiaodong Zhang
(MDACC)

Hang Liu (TACC)



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





Photon vs Proton

25 GY unnecessary photon radiation

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

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

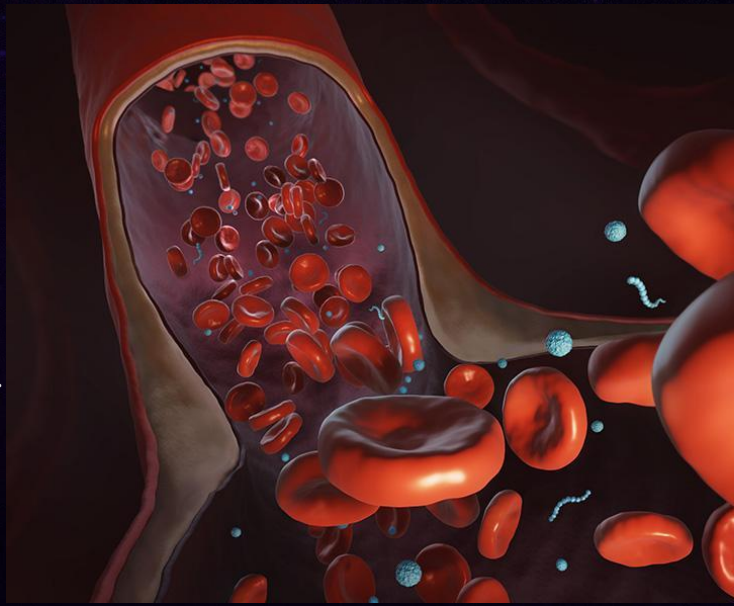
TEXAS The University of Texas at Austin
The University of Texas at Austin Oden Institute for Computational Engineering and Sciences
THE UNIVERSITY OF TEXAS MD Anderson Cancer Center
TACC TEXAS ADVANCED COMPUTING CENTER

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

TARGETING TUMORS WITH NANOWORMS

YING LI, UCONN

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



TECHNOLOGIES THAT HELP MOVE THINGS AROUND

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

TECHNOLOGIES THAT HELP MOVE THINGS AROUND

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

TECHNOLOGIES THAT HELP MOVE THINGS AROUND

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

- Plenty of our staff move back and forth ☺.

TH


FRONTERA
TACC
NSF
TEXAS

FHIR for Genomics: The Path Forward

Mullai Murugan (Baylor College of Medicine)

Overview - HL7 FHIR for Genomics

FHIR & CG Overview

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

Clinical Genomics Genomics Reporting IG

HL7 International

Genomics Reporting Implementation Guide
2.1.0-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 chang

Home Page

.1 Scope

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

his 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
- Germline and somatic variations
- Relevance of identified variations from the perspective of disease pathology, pharmacogenomics, transplant suitabi
- Full and partial DNA sequencing, including whole genome and exome studies

.2 How to Use this Guide

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

he remaining sections provide support for more specialized types of reporting. If your system is involved with genom f the implementation guide for further guidance.

ackground	Introduces some of the key genomics terms and relationships that should be understood b
General Genomic Reporting	Overall guidance in using the profiles and transactions defined in this guide. Guidance and report overall interpretations and how to report genotypes, haplotypes, and different types
Variant Reporting	Guidance on expressing information about variants gleaned from various sequencing appro array-based testing, etc.
harmacogenomic Reporting	Guidance and examples related to genomic testing done for the purpose of assessing gene for oncology and for general patient treatment.
Somatic Reporting	Guidance related to genomic testing done on somatic (non-germline) tissues, including as
Histocompatibility Reporting	Guidance related to genomic testing done for histocompatibility and immunogenomics ass

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-reportin

3 General Genomic Reporting

This page defines the core profiles and concepts that would be expected to be present in most genomic r
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.

Genomics Report	Groups together all the structured data being reported for a genomic testing.
Overall Interpretations	Reported when variant analysis (sequencing or targeted variants) is done. Provide reported.
Genomic Findings	These are observations about the specimen's genomic characteristics. For example haplotype, or variant that was detected.
Genomic Implications	These represent observations where the `Observation.subject` is typically the Pa refer to Genomic Findings. For example, "Patient may have increased susceptibility
Region Studied	These are observations describing the region or regions that were studied as part
Other Observations	The results of tests other than sequenced genomic variants may also be included t
Recommended Actions	Specific actions be taken, such as genomic counseling, re-testing, adjusting drug d
Contextual Resources	Other resources that provide contextual details.

19.0.3 Structures: Resource Profiles

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

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

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 varian

For each variant reporting pattern, different components MUST be used to properly define the varia 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 i relationships among human variations and phenotypes, and NCBI's Variation Services that relies).

4.2.1 Variants Defined by a Nomenclature Statement

This pattern describes the observed nucleotide sequence or configuration using HGVS or ISCN stat properly distinguish variants with the degree of precision needed for clinical use. Note that synonym normalization may be required.

Defining Component	Example Value
genomic-hgvs (LOINC 81290-9) OR coding-hgvs (LOINC 48004-6)	`{ "system" : "http://varnomen.hgvs.org", "code" : "NM_022787.3:c.769G>A" }`
cytogenomic-nomenclature (LOINC 81291-7)	`{ "system" : "urn:oid:2.16.840.1.113883.6.299", "code" : "46,XX,t(9;22)(q34;q4)" }`

4.2.2 Variants defined by multiple components (VCF-like)

This representation leverages multiple component slices to communicate an allele within the conte definitional representation in FHIR, but is limited to variations with known breakpoints, and alleles specify genome build and chromosome identifiers rather than explicit reference sequences. Build a

Defining Component	Example Value
genomic-ref-seq (LOINC 48013-7)	`{ "system" : "http://www.ncbi.nlm.nih.gov/nuccore", "code" : "NC_000010.10" }`

19.0.7 Terminology: Code Systems

These define new code systems used by systems conforming to this in

ClinVar Evidence Level Example Codes	ClinVar contains examples of evid https://www.ncbi.nlm.nih.gov/cli relevant source.
Coded Annotation Type Codes	Code System for specific types o
PharmGKB Evidence Level Example Codes	PharmGKB contains examples of https://www.pharmgkb.org/page
Sequence Phase Relationship Codes	Code System for specific types o
To Be Determined Codes	These codes are currently 'TBD' c validated.
Variant Confidence Status Codes	A code that represents the confic

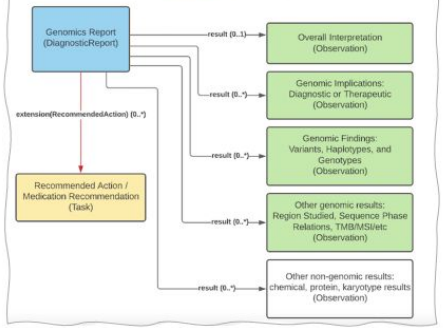
19.0.8 Terminology: Value Sets

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

Coded Annotation Types	Value Set for specific types of coded annotations
Condition Inheritance Patterns	Value Set for specific transmission patterns of a condition in a pedigree
DNA Change Type	DNA Change Type of a variant.
Evidence Level Examples	Example sources of values for Evidence Level
Functional Effect	The effect of a variant on downstream biological products or pathways.
Genetic Therapeutic Implications	Value Set for terms that describe a predicted ramification based on the presence
HUGO Gene Nomenclature Committee Gene Names (HGNC)	This value set includes all HGNC Codes, which includes multiple code systems. In Consortium/Locus, which "must be used with the HGNC gene ID including the pref

3.1 Genomics Report

The genomics report is the focus of all genomic reporting. It conveys metadata about the overall report (what kind etc.). It also typically includes a rendered version for review by a clinician. It also groups together all relevant infor will depend on the type of testing ordered, the reason for testing and the policies of the lab). Most of the structure recommendations that come with the report are expressed as FHIR Tasks. The report can be organized into sub- summaryOf., which is especially useful for later analysis steps. Additionally, an observation can be used to group with a specific LOINC panel code in the `Observation.code` for example).



19.0.5 Structures: Extension Definitions

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

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

New Implementers

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

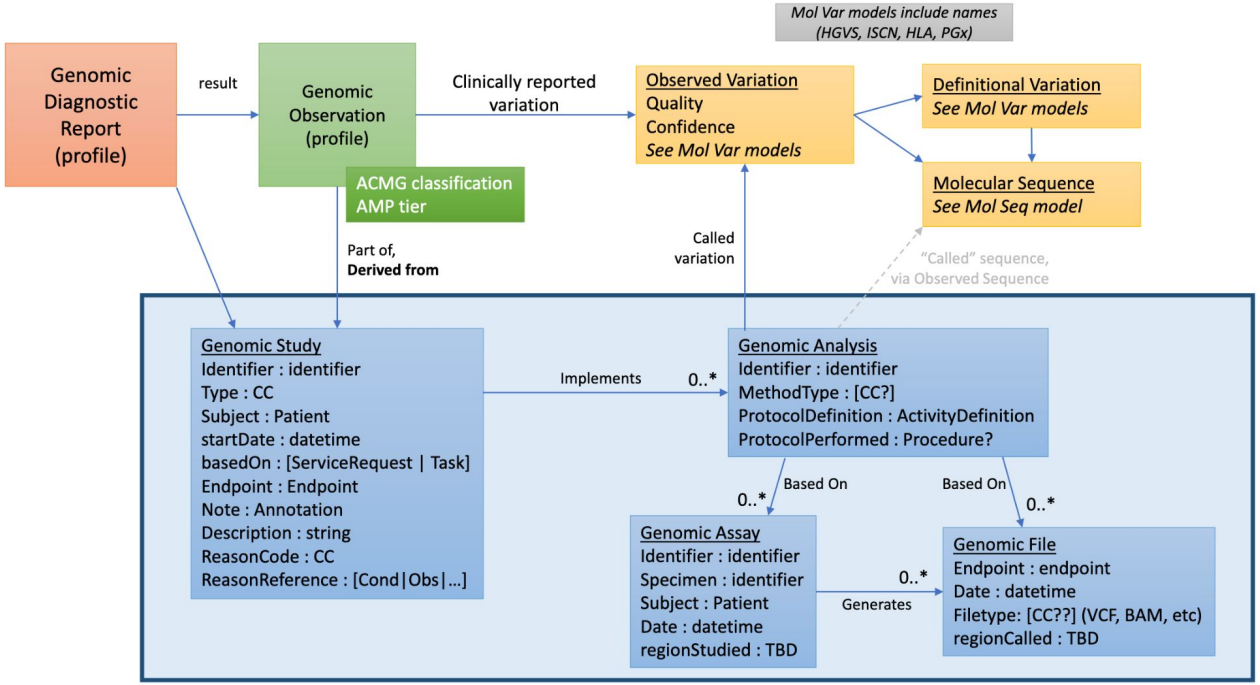
FHIR Genomics - New Initiatives & Ongoing Effort

Genomics FHIR Initiatives

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

Genomic Study

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



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

Not shown: implication, interpretation, etc

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

Challenges, and the path forward

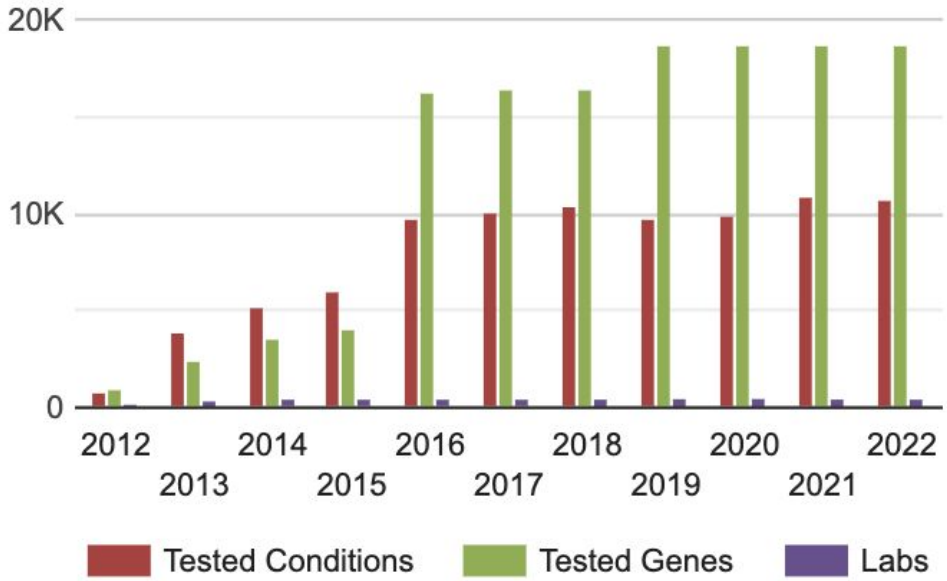
Challenges, and the path forward

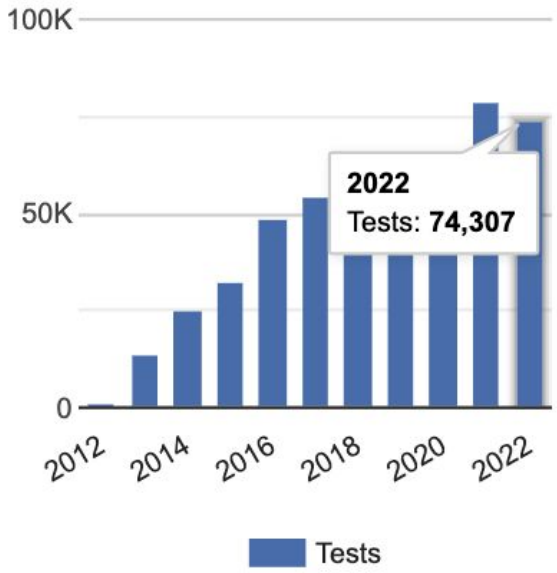
1. Learning curve
2. Ease of
3. Mul
4. Dive
5. Adc

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

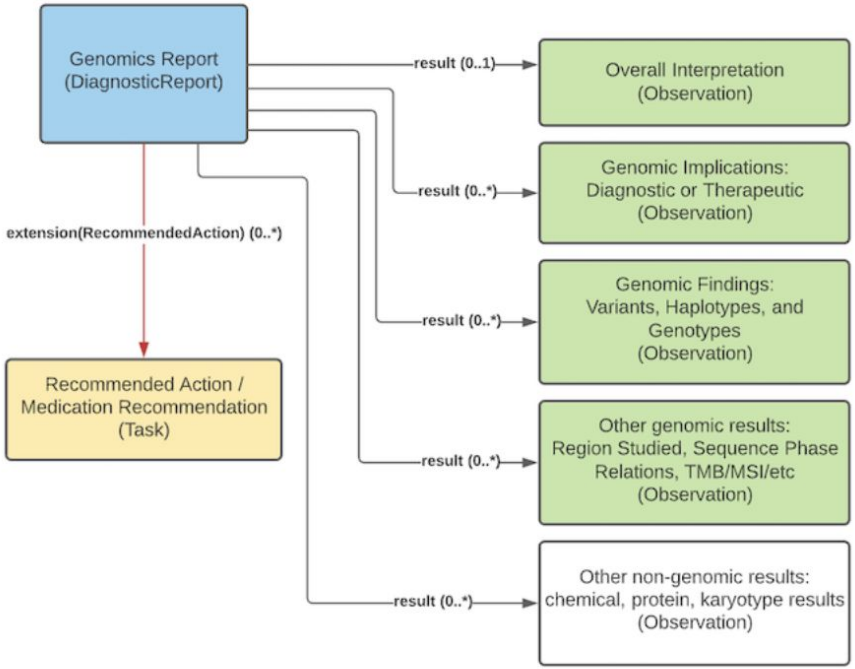
- **FMM0 (Draft)** – The resource is still in early development but has been accepted into the FHIR standard.

GTR Data

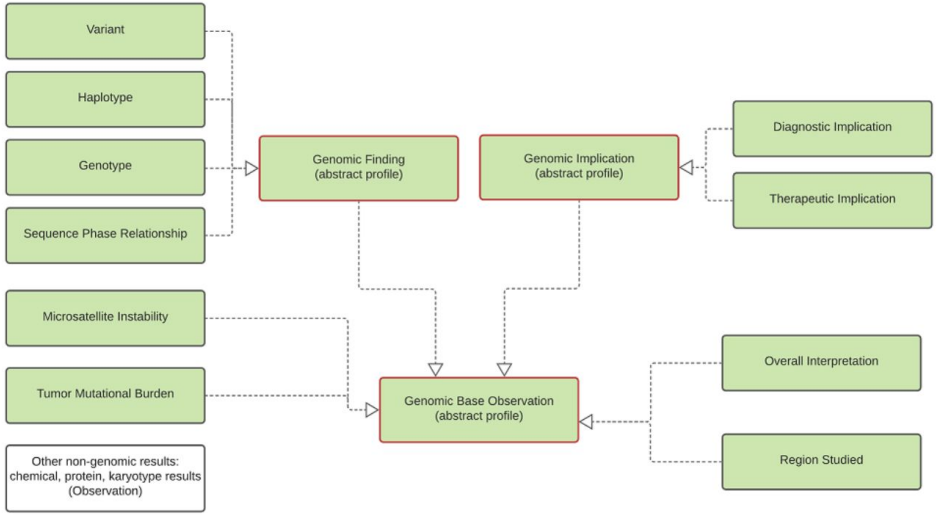




1. Clinical Genomics IG Learning Curve



Genomic Report Overview



Genomic Observations

4.2 Defining Variants

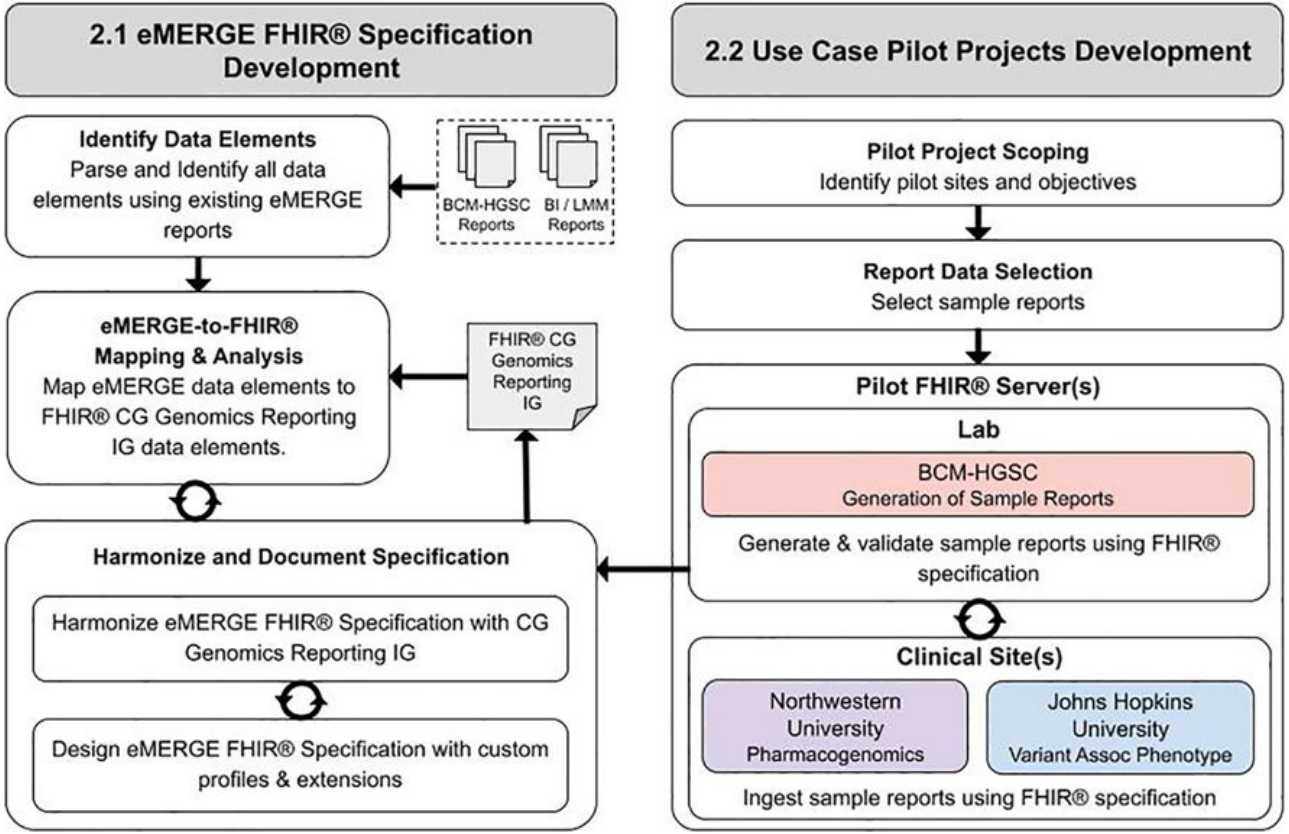
This Implementation Guide supports two reporting patterns for defining variants:

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

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

Clinical Genomics GR IG

2. Ease of Implementation



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

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

Getting Started? Go here

3.Multiple Pilot Efforts

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

- *Gen*
- FHI

1.Completed Major

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

2.Completed Minor

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

2.Completed Minor (cont'd)

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

3.Pending

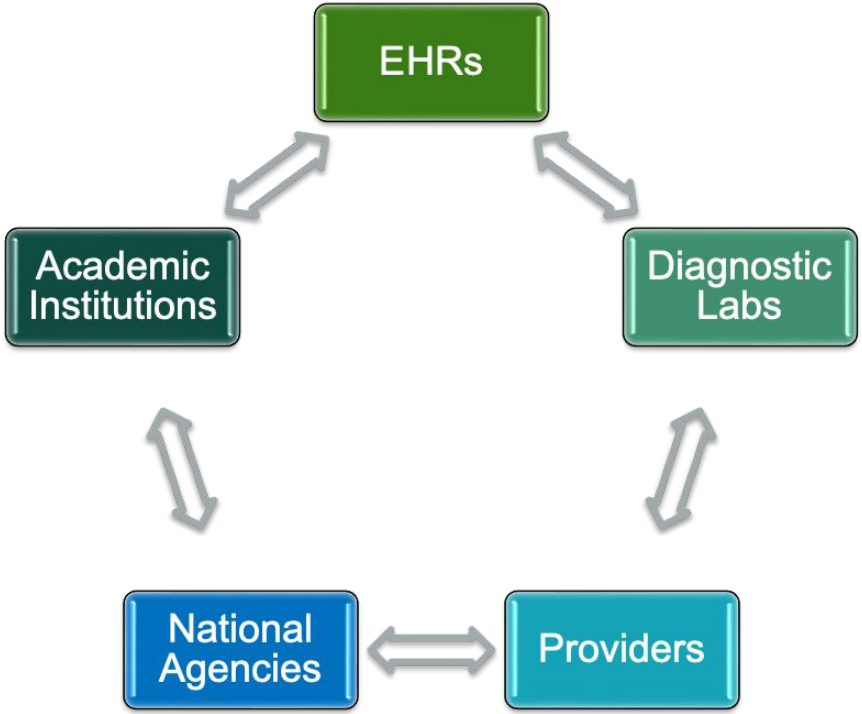
1. RecommendedAction Task reasonRef cardinality to 0..* (OO)
2. Add Age to US-Core Patient Profile (PatAdm)
3. Clinical vs Research Flag (Core)
4. Why is DR.code fixed to LOINC 81247-9? (CG)
5. RecommendedAction profile "code" should be extensible (CG)

4. Diversity of the tech landscape

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

5. Adoption and direction

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



Acknowledgements

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

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

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

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

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

Heidi Rehm (Broad Institute/MGH)

The Global Alliance for Genomics and Health Mission...

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

GA4GH achieves this by...

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

GA4GH Organization Structure

Global Alliance for Genomics & Health

STRATEGIC ADVISORY/FUNDING

SAB

Funders Forum

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

ORGANIZATIONAL GOVERNANCE

Board of Directors

- Finance, Audit, & Risk Committee
- Executive Committee
- Nominating Committee

CEO

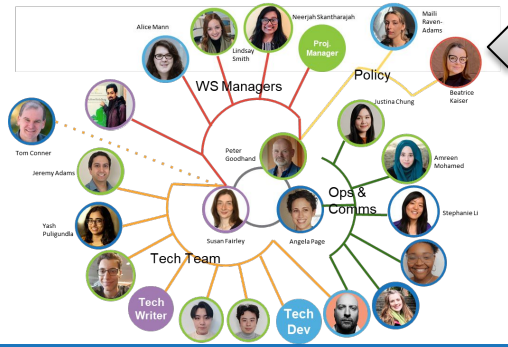
Secretariat

- Ops Team
- Technical Standards Team
- Work Stream Managers
- Comms Team

TECHNICAL GOVERNANCE

Standards Steering Committee

21 full-time staff



Executive Leadership

Peter Goodhand
Chief Executive Officer

Ewan Birney
Chair

Kathryn North
Vice Chair

Heidi Rehm
Vice Chair

Angela Page
Director of Strategy and Engagement

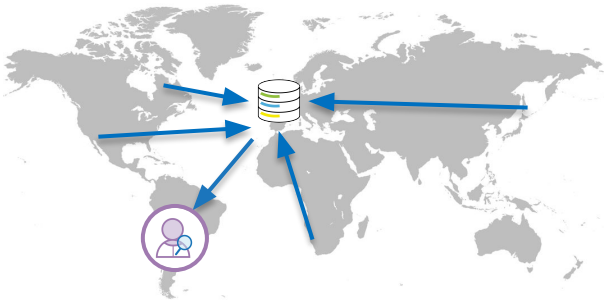
Susan Fairley
Chief Standards Officer

Assigned Experts & Active Contributors

Different Approaches to Data Sharing

Central Database

Genomic Knowledgebase

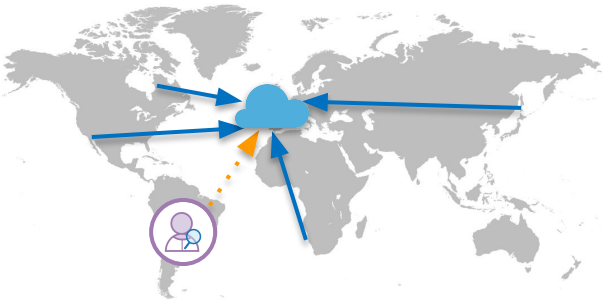


Aggregate data globally

Download, analyze locally

Secure Cloud

Large scale research datasets

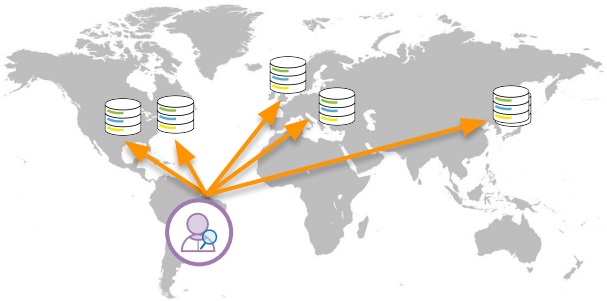


Aggregate data globally

Analyze centrally in secure cloud

Federation

Connecting national genomics initiatives



Host data locally

Visit data remotely and collate results

User — Data transmission — Secure access

How GA4GH Works

Technical Work Streams

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

Foundational Work Streams

- Regulatory & Ethics
- Data Security

Input from Driver Projects as exemplars of community

Partner Engagement

Output of standards that are taken up by the community

RARE DISEASES

Matchmaker Exchange

COMPLEX TRAITS

AUTISM SPEAKS

NHLBI TOPMed

GEM Japan

H3Africa

SPHN

ClinGen

Australian Genomics Health Alliance

All of Us Research Program

Genomics England

elixir

ENA

monarch initiative

epi share

Human Cell Atlas

CANCER

BRCA Exchange

ARGO

NIH National Cancer Institute

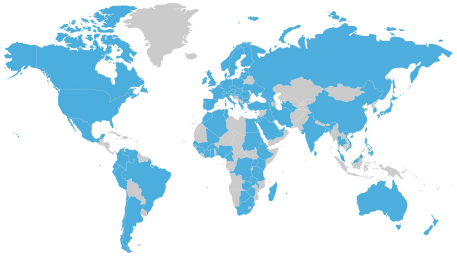
EU Can Can.

VICC Variant Interpretation for Cancer Consortium

BASIC BIOLOGY

24 Driver Projects

Real-World Driver Projects



Federated Analysis Systems Project (FASP)

Starter Kit

Technical Alignment Subcommittee (TASC)

Equity, Diversity and Inclusion Advisory Board

GA4GH 2020-2022 Strategic Roadmap

Global Alliance for Genomics & Health

- Approved
- Ongoing
- In Development

Clin/Pheno Data Capture
- Phenopackets
- Pedigree Representation

Cloud
- Workflow Execution Service API
- Tool Registry Service
- Data Repository Service API
- Task Execution Service
- Cloud Testbed Interoperability

Discovery
- Beacon API
- Service Info
- Service Registry
- Data Connect API

Data Security
- Authentication & Authorization Infrastructure
- Data Security and Infrastructure Policy
- Risk Assessment
- Breach Response Protocol
- Cloud Security & Privacy

Data Use & Researcher Identity
- Data Use Ontology
- GA4GH Passports
- Machine Readable Consent Guidance

Genomic Knowledge Standards
- Variation Representation
- Variation Annotation

Large Scale Genomics
- htsget API
- refget API
- Read file formats
- Variation File Formats
- Crypt4GH
- rnaget API
- BED File Format

Regulatory & Ethics
- Framework
- Policy Initiatives
  - Ethics Review Recognition, Accountability, Consent, Privacy & Security
  - Technical Standards & IP, Return of Results
- Responsible Data Sharing Resource Documents
- GDPR Forum

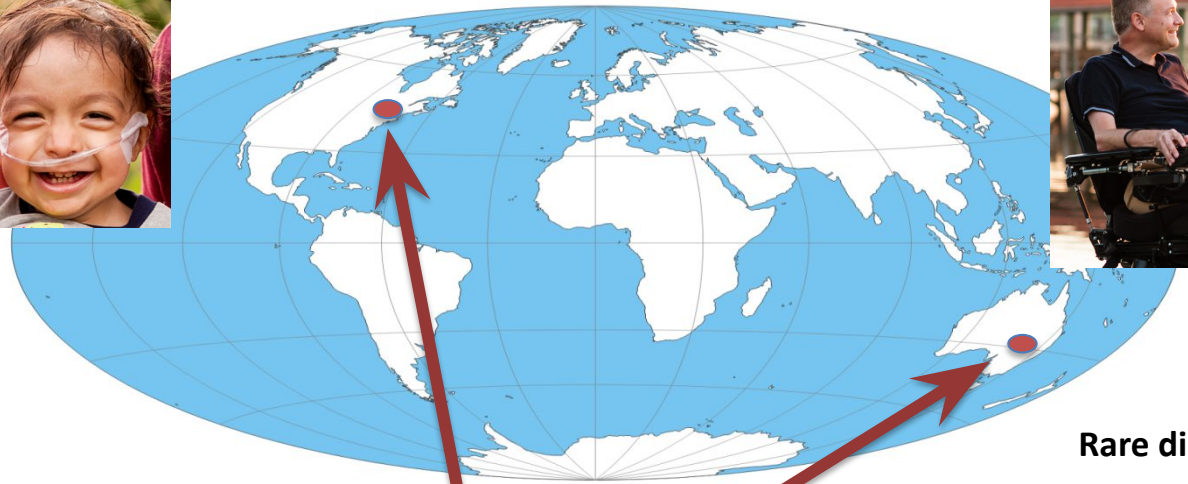
Challenges in rare disease gene discovery

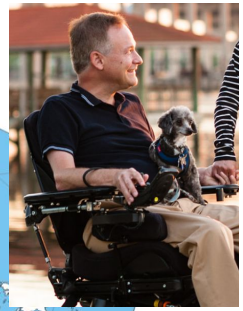
- 75% of rare disease cases remain unsolved
- 4,631 genes implicated in at least one disease but evidence for >10,000 more genes yet to be discovered for Mendelian disease (Bamshad, et al. AJHG 105, 448–455, 2019)
- The remaining genetic diseases are very, very rare – difficult for any one investigator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

Principles of Gene Matching

Individual with rare disease

Clinical geneticist





Individual with rare disease

Rare disease researcher

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

Genotypic Data
Gene D

→

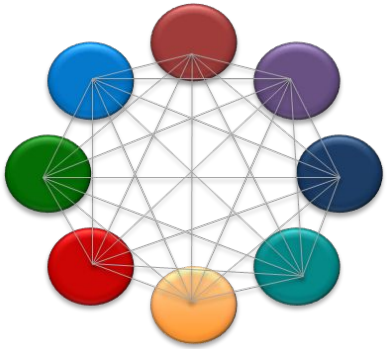
Genomic Matchmaker

←

Genotypic Data
Gene D

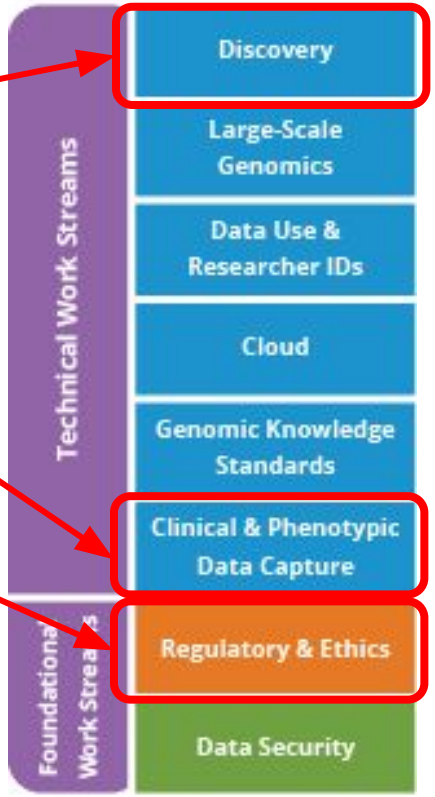
Phenotypic Data
Feature 1
Feature 3
Feature 4
Feature 5
Feature 6

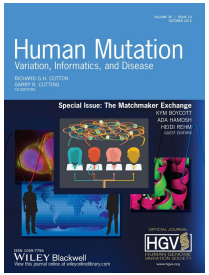
Developing the MME Federated Network using GA4GH Standards



Use of GA4GH standards:

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



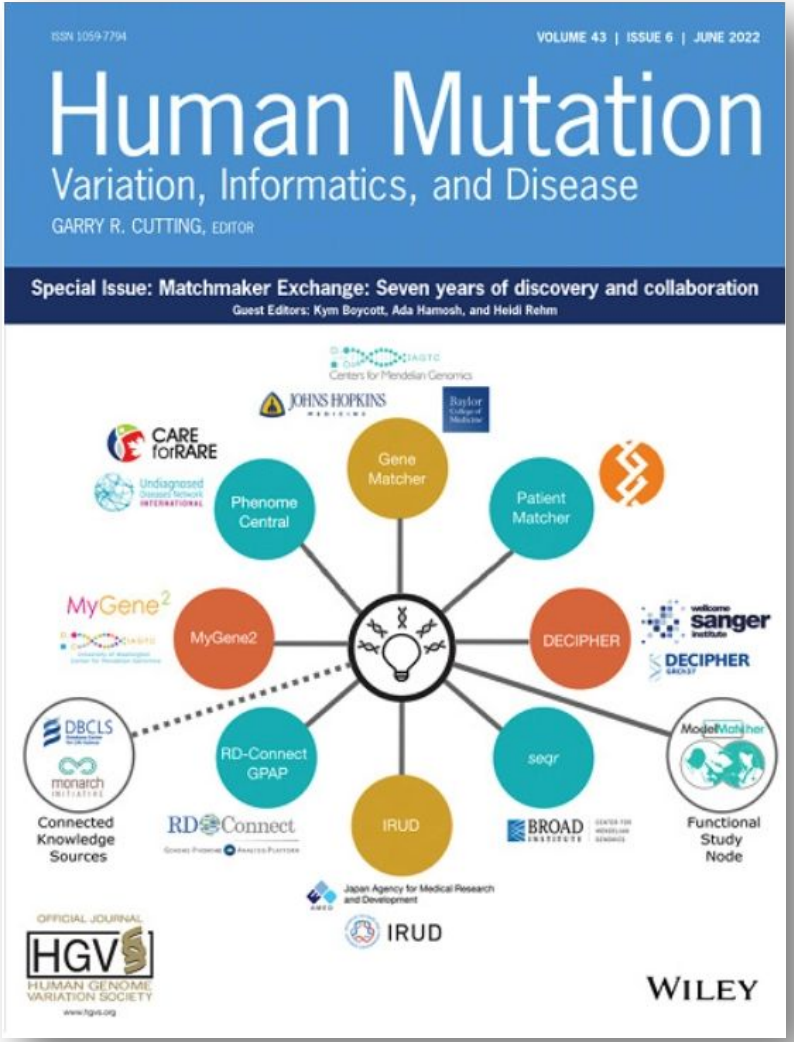


Philippakis et al. **The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery.** Hum Mutat. 2015;36(10):915-21.

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

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

www.matchmakerexchange.org



EDITORIAL INTRODUCTION

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

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

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

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

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

GeneDx
Illumina
Ambry

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

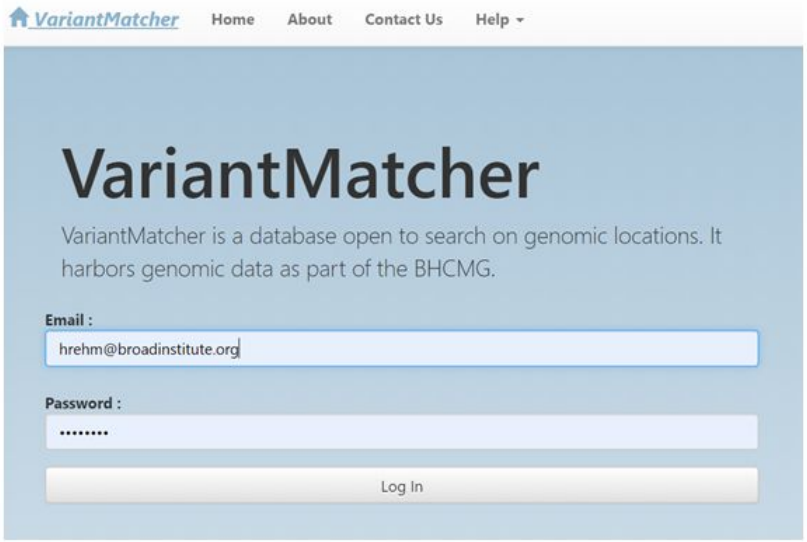
(a) Two-sided matchmaking

Matchmaker Exchange

(b) One-sided matchmaking

(c) Zero-sided matchmaking

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



VariantMatcher (VM) created by:

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

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

---

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

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

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

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

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

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

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

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

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

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

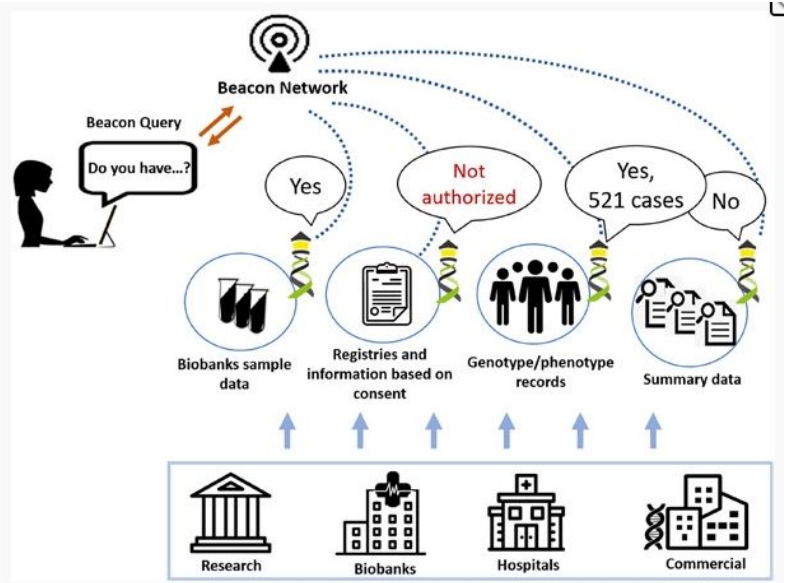
Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

INFORMATICS | Open Access

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

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

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



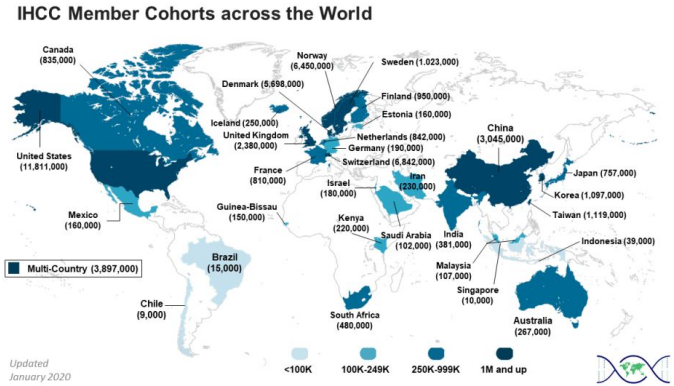
Human Mutation
Variation, Informatics, and Disease
OFFICIAL JOURNAL HGV HUMAN GENOME VARIATION SOCIETY

DATABASES | Open Access

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

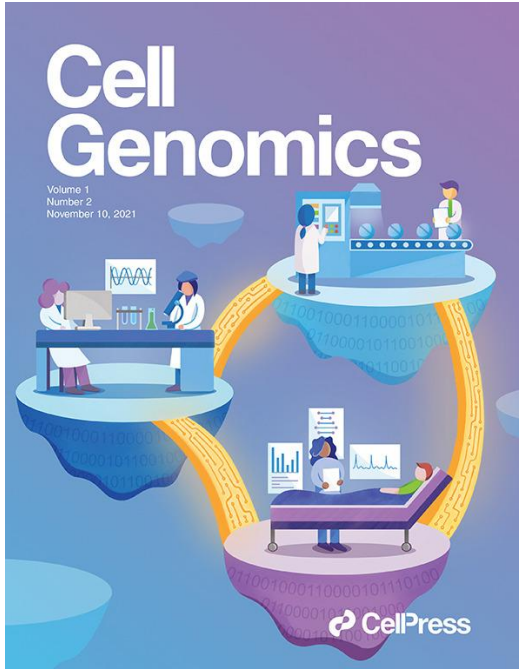
Eliete da S. Rodrigues, Sean Griffith, Renan Martin, Corina Antonescu, Jennifer E. Posey, Zeynep Coban-Akdemir, Shalini N. Jhangiani, Kimberly F. Doheny, James R. Lupski, David Valle ... See all authors

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



MyGene2, Geno2MP, VariantMatcher, Franklin

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



Cell Genomics

CellPress
OPEN ACCESS

Perspective

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

Heidi L. Rehm,[1,2,47] Angela J.H. Page,[1,3,*] Lindsay Smith,[3,4] Jeremy B. Adams,[3,4] Gil Alterovitz,[5,47] Lawrence J. Babb,[1] Maxmillian P. Barkley,[6] Michael Baudis,[7,8] Michael J.S. Beauvais,[3,9] Tim Beck,[10] Jacques S. Beckmann,[11] Sergi Beltran,[12,13,14] David Bernick,[1] Alexander Bernier,[9] James K. Bonfield,[15] Tiffany F. Boughtwood,[16,17] Guillaume Bourque,[9,18] Sarion R. Bowers,[15] Anthony J. Brookes,[10] Michael Brudno,[18,19,20,21,38] Matthew H. Brush,[22] David Bujold,[9,18,38] Tony Burdett,[23] Orion J. Buske,[24] Moran N. Cabili,[1] Daniel L. Cameron,[25,26] Robert J. Carroll,[27] Esmeralda Casas-Silva,[123] Debyani Chakravarty,[29] Bimal P. Chaudhari,[30,31] Shu Hui Chen,[32] J. Michael Cherry,[33] Justina Chung,[3,4] Melissa Cline,[34] Hayley L. Clissold,[15] Robert M. Cook-Deegan,[35] Mélanie Courtot,[23] Fiona Cunningham,[23] Miro Cupak,[6] Robert M. Davies,[15] Danielle Denisko,[19] Megan J. Doerr,[36] Lena I. Dolman,[19] Edward S. Dove,[38] L. Jonathan Dursi,[20,39] Stephanie O.M. Dyke,[9] James A. Eddy,[37] Karen Eilbeck,[40] Kyle P. Ellrott,[22] Susan Fairley,[3,23] Khalid A. Fakhro,[41,42] Helen V. Firth,[15,43] Michael S. Fitzsimons,[44] Marc Fiume,[6] Paul Flicek,[23] Ian M. Fore,[29] Mallory A. Freeberg,[23] Robert R. Freimuth,[45] Lauren A. Fromont,[52] Jonathan Fuerth,[6] Clara L. Gaff,[16,17] Weiniu Gan,[33] Elena M. Ghanaim,[46] David Glazer,[47] Robert C. Green,[1,48,49] Malachi Griffith,[50] Obi L. Griffith,[50] Robert L. Grossman,[44] Tudor Groza,[51] Jaime M. Guidry Auvil,[29] Roderic Guigó,[13,52] Dipayan Gupta,[23] Melissa A. Haendel,[53] Ada Hamosh,[54] David P. Hansen,[16,83] Reece K. Hart,[1,100,124] Dean Mitchell Hartley,[55] David Haussler,[35] Rachele M. Hendricks-Sturrup,[56] Calvin W.L. Ho,[57] Ashley E. Hobb,[6] Michael M. Hoffman,[19,20,21] Oliver M. Hofmann,[26] Petr Holub,[58,59] Jacob Shujui Hsu,[60] Jean-Pierre Hubaux,[61] Sarah E. Hunt,[23] Ammar Husami,[62] Julius O. Jacobsen,[63] Saumya S. Jamuar,[64,65] Elizabeth L. Janes,[3,66] Francis Jeanson,[126] Aina Jené,[52] Amber L. Johns,[67,68] Yann Joly,[9] Steven J.M. Jones,[69] Alexander Kanitz,[8,70] Kazuto Kato,[71] Thomas M. Keane,[23,72] Kristina Kekesi-Lafrance,[3,9] Jerome Kelleher,[73] Giselle Kerry,[23] Seik-Soon Khor,[74,75] Bartha M. Knoppers,[9] Melissa A. Konopko,[76] Kenjiro Kosaki,[77] Martin Kuba,[59] Jonathan Lawson,[1] Rasko Leinonen,[23] Stephanie Li,[1,3] Michael F. Lin,[78] Mikael Linden,[79,80] Xianglin Liu,[86] Isuru Udara Liyanage,[23] Javier Lopez,[101] Anneke M. Lucassen,[81] Michael Lukowski,[44] Alice L. Mann,[3,15] John Marshall,[68] Michele Mattioni,[82] Alejandro Metke-Jimenez,[83] Anna Middleton,[84,85] Richard J. Milne,[84,85] Fruzsina Molnar-Gabor,[86] Nicola Mulder,[87] Monica C. Munoz-Torres,[53] Rishi Nag,[23] Hidewaki Nakagawa,[88,89] Jamal Nasir,[90] Arcadi Navarro,[52,91,92,93] Tristan H. Nelson,[94] Ania Niewielska,[23] Amy Nisselle,[17,26,95] Jeffrey Niu,[20] Tommi H. Nyrönen,[79,80] Brian D. O'Connor,[1] Sabine Oesterle,[8] Soichi Ogishima,[96] Laura A.D. Paglione,[97,98] Emilio Palumbo,[13,52] Helen E. Parkinson,[23] Anthony A. Philippakis,[1] Angel D. Pizarro,[99] Andreas Prlic,[100] Jordi Rambla,[13,52] Augusto Rendon,[101] Renee A. Rider,[46] Peter N. Robinson,[102,103] Kurt W. Rodarmer,[104] Laura Lyman Rodriguez,[105] Alan F. Rubin,[25,26] Manuel Rueda,[52] Gregory A. Rushton,[1] Rosalyn S. Ryan,[106] Gary I. Saunders,[76] Helen Schuilenburg,[23] Torsten Schwede,[8,70] Serena Scollen,[76] Alexander Senf,[107] Nathan C. Sheffield,[108] Neerjah Skantharajah,[3,4] Albert V. Smith,[109] Heidi J. Sofia,[46] Dylan Spalding,[79,80] Amanda B. Spurdle,[110] Zornitza Stark,[16,17,26] Lincoln D. Stein,[4,19] Makoto Suematsu,[77] Patrick Tan,[64,111,112] Jonathan A. Tedds,[76] Alastair A. Thomson,[33] Adrian Thorogood,[9,113] Timothy L. Tickle,[1] Katsushi Tokunaga,[75,114] Juha Törnroos,[79,80] David Torrents,[92,116] Sean Upchurch,[115] Alfonso Valencia,[92,116] Roman Valls Guimera,[26] Jessica Vamathevan,[23] Susheel Varma,[23,117] Danya F. Vears,[17,26,95,118] Coby Viner,[19,20] Craig Voisin,[119] Alex H. Wagner,[31,32] Susan E. Wallace,[10] Brian P. Walsh,[22] Vivian Ota Wang,[29] Marc S. Williams,[94] Eva C. Winkler,[120] Barbara J. Wold,[115] Grant M. Wood,[1] J. Patrick Woolley,[73] Chisato Yamasaki,[71] Andrew D. Yates,[23] Christina K. Yung,[4,121] Lyndon J. Zass,[87] Ksenia Zaytseva,[9,122] Junjun Zhang,[4] Peter Goodhand,[4,3] Kathryn North,[17,26] and Ewan Birney[23,123]

Get Involved! Visit GA4GH.ORG

Join a Work Stream!

Contact secretariat@ga4gh.org

Become an Organizational Member

ga4gh.org/members

Subscribe to GA4GH Updates

ga4gh.org/subscribe

Interoperability Opportunities & Challenges with the Cloud and STRIDES

Nick Weber (NIH STRIDES)

Interoperability Opportunities & Challenges with STRIDES & Cloud

NCPI Spring Workshop

Nick Weber

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

NIH STRIDES Initiative

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

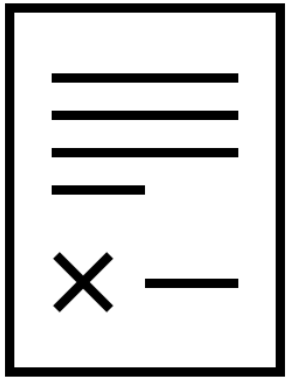
- State-of-the-art data storage and computational capabilities
- Training and education for researchers
- Innovative technologies such as artificial intelligence and machine learning
- Professional engineering and technical support

Partnerships with aws Google Cloud Microsoft Azure

Two Core Components of STRIDES

1) Other Transaction Agreement

Enables NIH-funded institutions to leverage STRIDES benefits



2) NIH Enterprise Cloud Platforms & Services

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

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

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

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

Sample of STRIDES-Supported Research Programs

All of Us RESEARCH PROGRAM

FRAMINGHAM HEART STUDY — Three Generations of Dedication — 1948 1971 2002 — A Project of the National Heart, Lung, and Blood Institute of Boston University

National Ophthalmic Disease Genotyping Network eyeGENE® A program of the National Institutes of Health

GTEx Portal

PRIMED consortium

NHLBI TOPMed

Metabolomics Workbench

COMMON FUND DATA ECOSYSTEM

HuBMAP Human BioMolecular Atlas Program

AMP PD

HEAL NIH · Helping to End Addiction Long-term

UniProt

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

SPARC

AnVIL

MoTrPAC The Molecular Transducers of Physical Activity Consortium

HMP

DK net

eRA

The Down Syndrome Consortium

Adolescent Brain Cognitive Development

Glycoscience

IDG ILLUMINATING the DRUGGABLE GENOME

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

NIMH Data Archive NDA | ABCD | CCF | NIAAA$_{DA}$ | OAI

NIA IMPACT COLLABORATORY TRANSFORMING DEMENTIA CARE

NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS Access the Data #NCIGDC

NIH Center for Information Technology

Galaxy PROJECT

SenNet

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®.

NEW: NIH Cloud Lab Offering

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

Primary Cloud Lab Use Cases

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

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

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

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

NIH Cloud Lab (continued)

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

Full Access to the Cloud Console

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

Bioinformatic Tutorials to Speed Uptake

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

Broad Access Across the NIH Community

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

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

Interoperability Challenges & Considerations

- New Data Management & Sharing Policy
- Modularity / portability / reusability
- Cross-cloud billing integration
- Cost enforcement
- Cost estimation

- Institution-level data mesh "nodes"?
- Pilot programs for standardization around products like Kubernetes, Docker, etc.?
- RAS as an underpinning for billing auth?
- NIH Cloud Lab examples / source code?
- NIH Cloud Lab & community contributions?

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

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

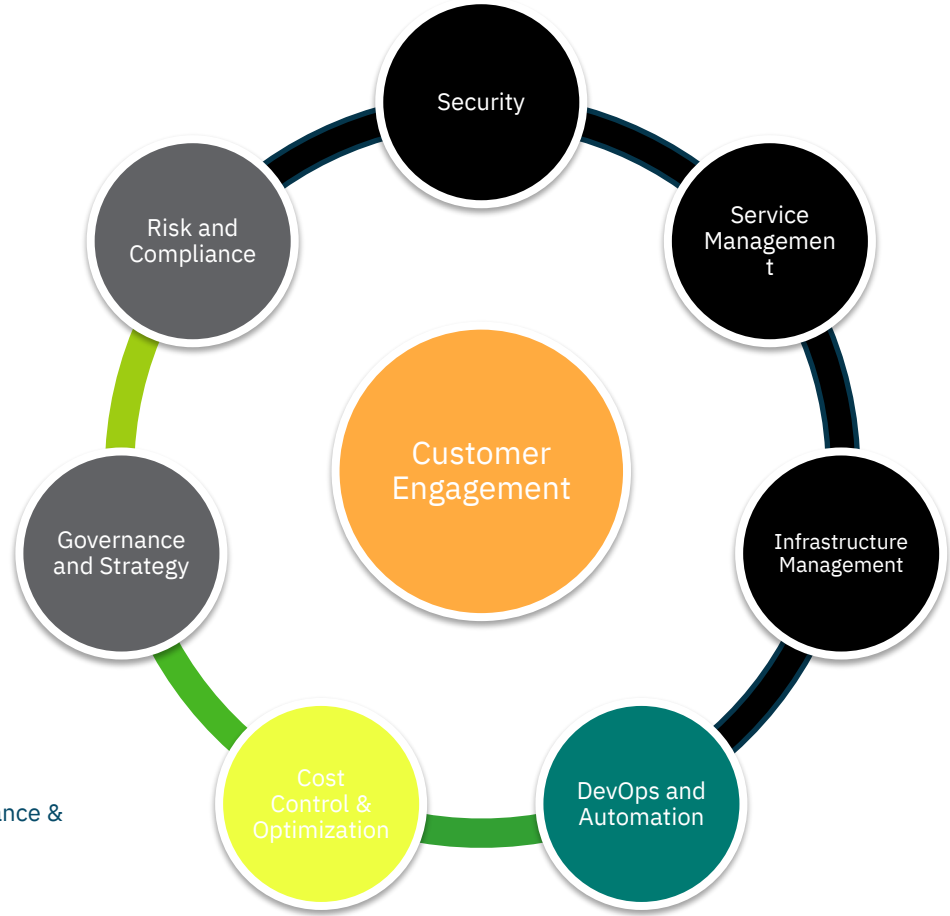
Interoperability in general requires mastery of the fundamentals (see: RAS); cloud infrastructure interoperability is no different

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

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

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

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



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

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

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

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

Concurrent Breakout Session

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

2:35 PM - 3:50 PM EDT

Topic 1: Bringing researchers to cloud computing

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

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

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

Provenance is a higher priority than perfect reproducibility

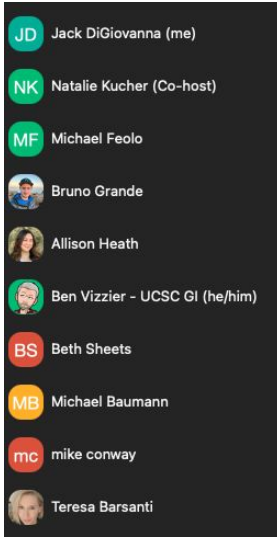
First step would be more information about data used

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

Two types of data releases important for different goals

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

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



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

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

Topic 4: Diversifying genomic data science

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

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
- 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)