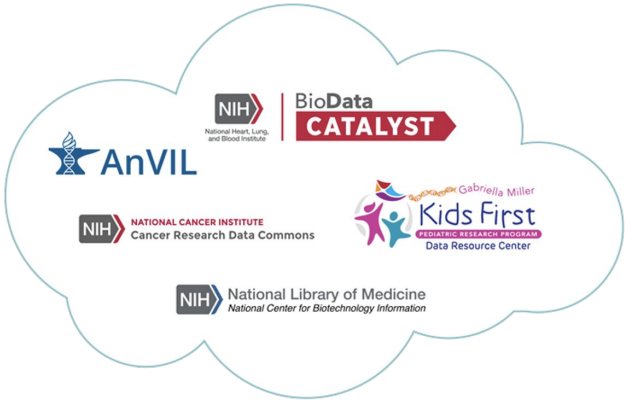
June 22-23, 2022

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

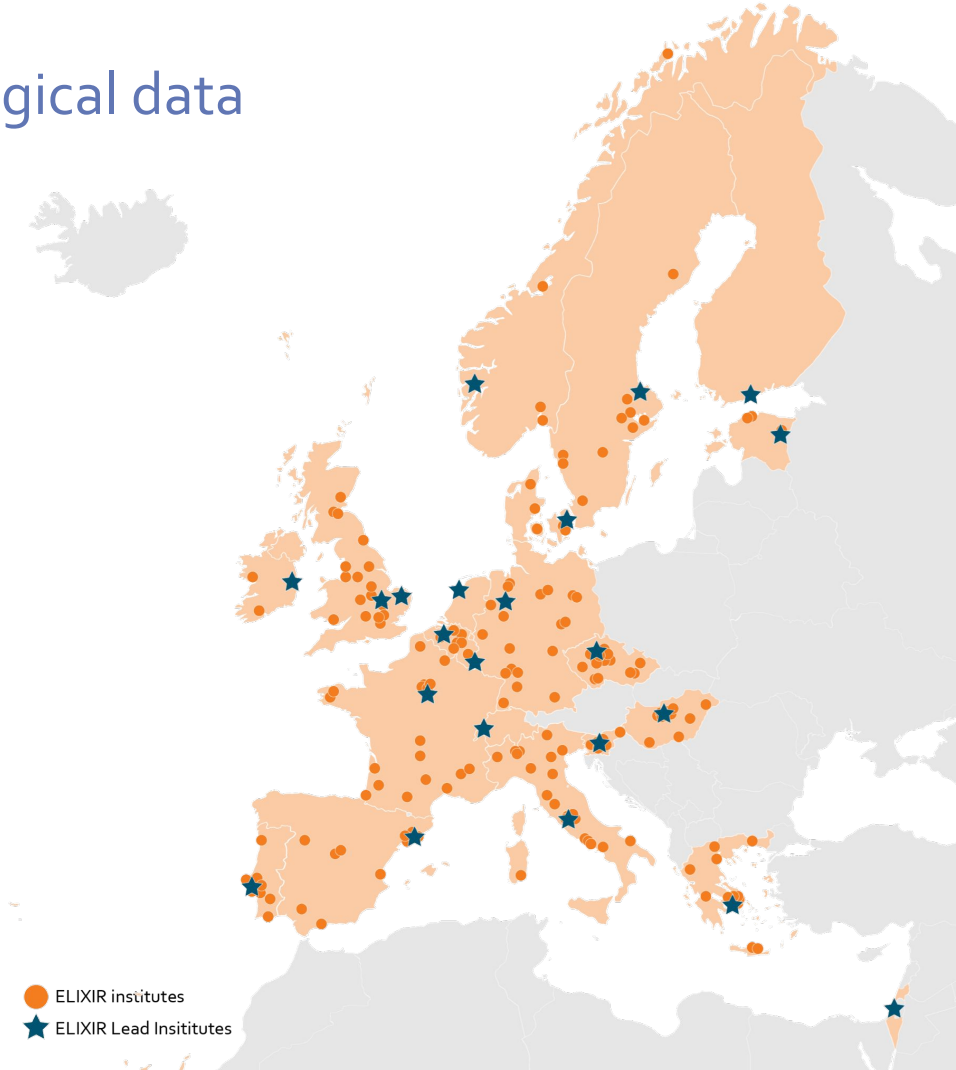
Jonathan Tedds (Compute, Tools Platform & EOSC Coordinator)

www.elixir-europe.org

A sustainable infrastructure for biological data

ELIXIR Members

ELIXIR Observers

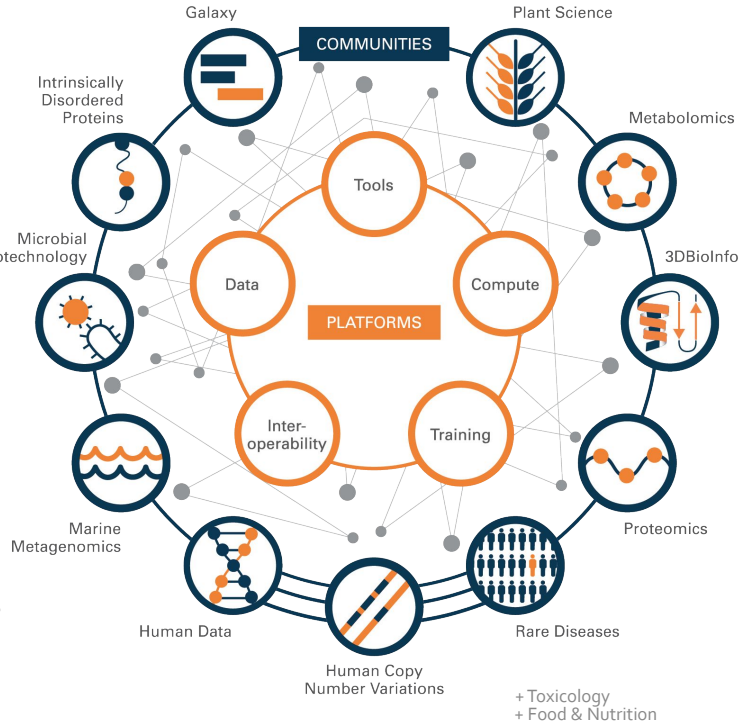


ELIXIR Services for all domains of life sciences

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

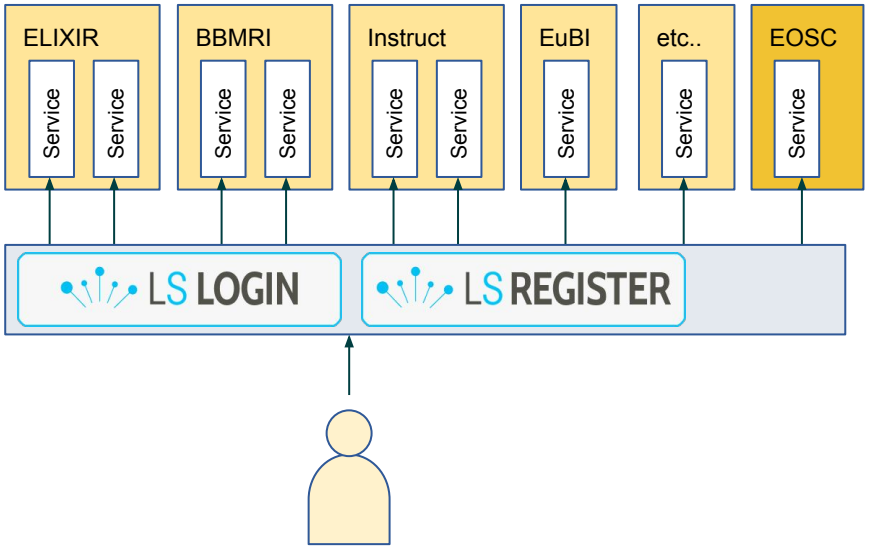
The vast majority of **ELIXIR services** are available free of charge and accessible globally by anyone interested



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

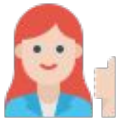
Services & Solutions

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

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

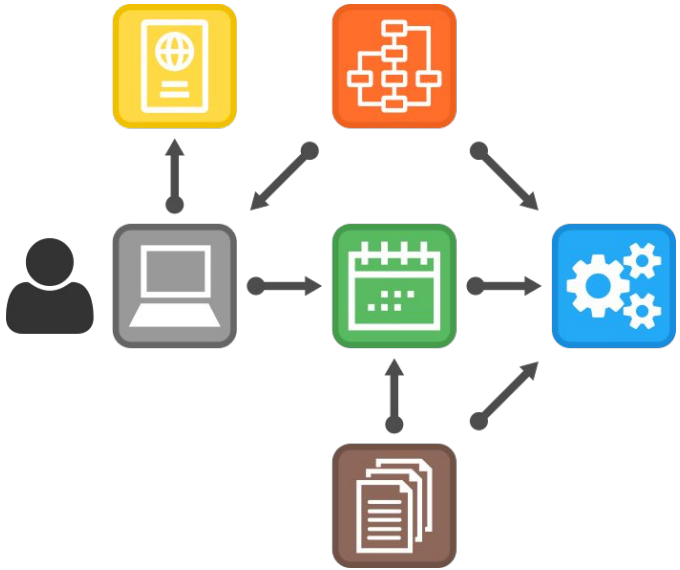
Access to data sets

WES: Workflow Execution Service API

Interpret workflows & schedule task execution

TES: Task Execution Service API

Execute tasks

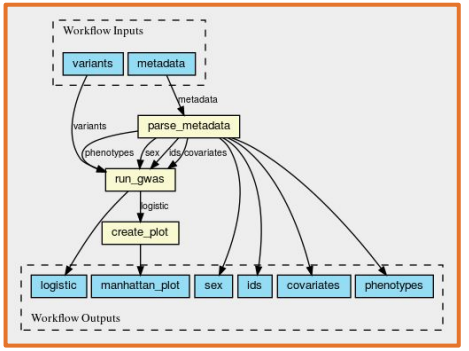


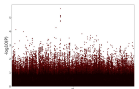
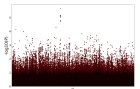
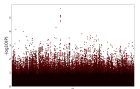
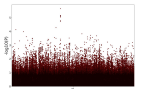
Moonshot demonstrator (8th GA4GH Plenary)

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

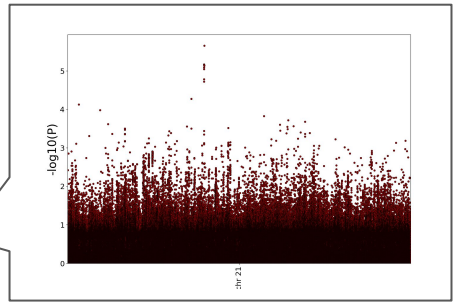
VCF, 1000 Genomes project

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**
- **BiGi - Bielefeld-Gießen Resource Center for Microbial Bioinformatics**
- **BioInfra.Prot - Bioinformatics for Proteomics**
- **CIBI - Center for Integrative Bioinformatics**
- **RBC - RNA Bioinformatics Center**
- **GCBN - German Crop BioGreenformatics**
- **BioData - Center for Biological Data**
- **de.NBI-SysBio - de.NBI Systems Biology Service Center**
- **Associated Partners**

➡ 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

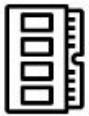
de.NBI Cloud Infrastructure

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

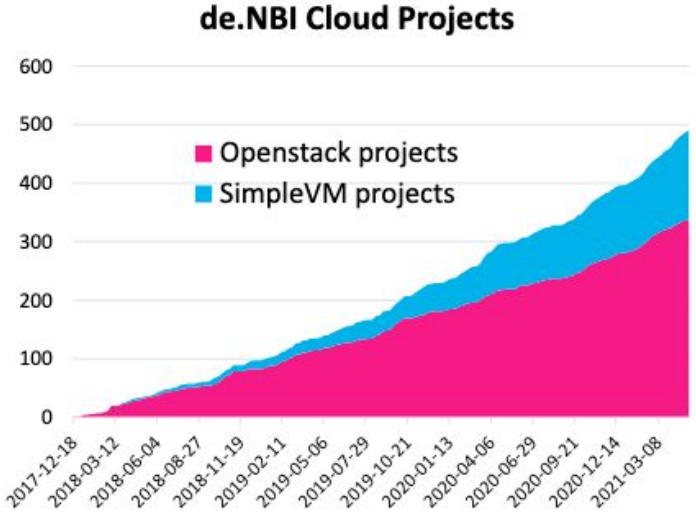
Storage capacity

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

80 PB

330 TB

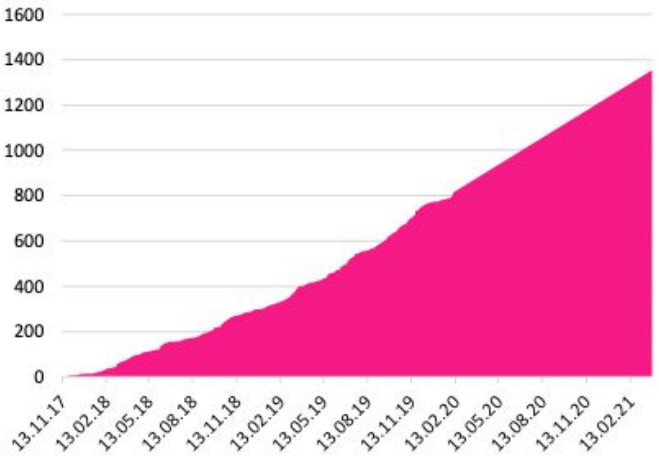
Project Numbers



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

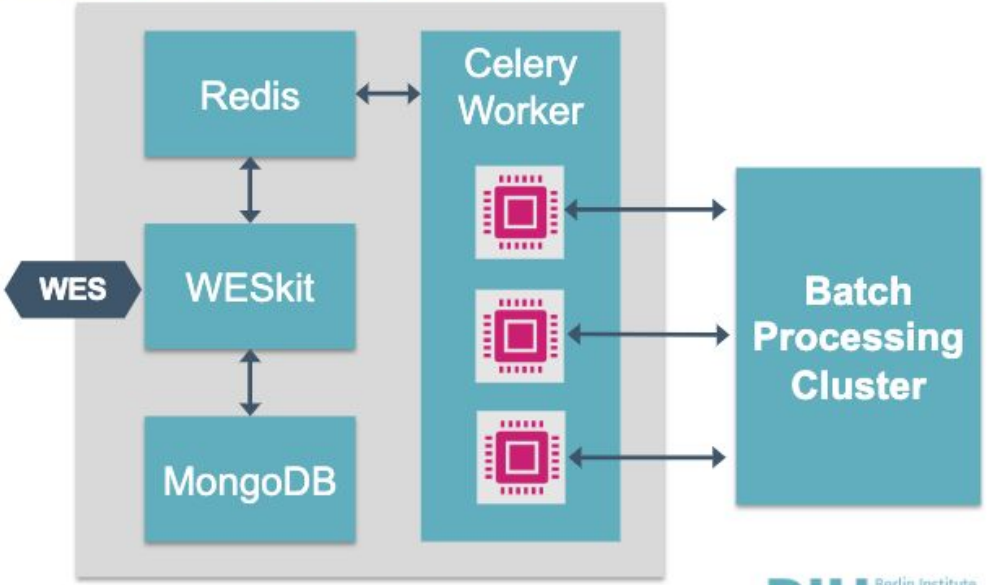
WESKIT

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

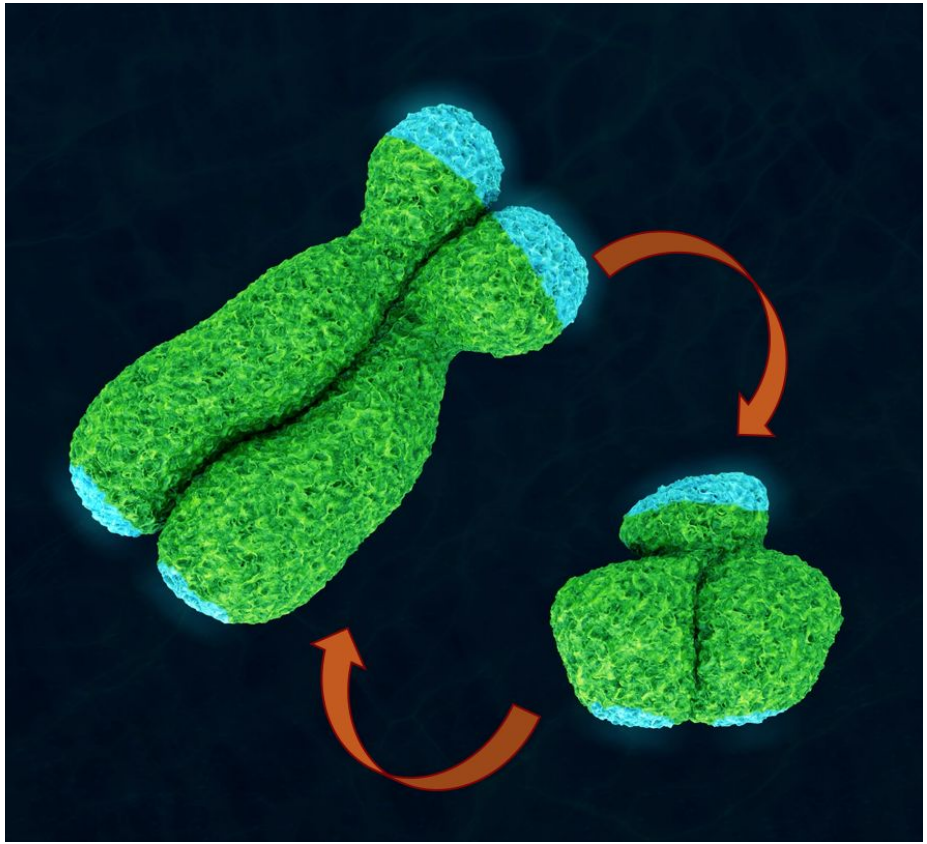
Melissa Wilson

@sexchrlab

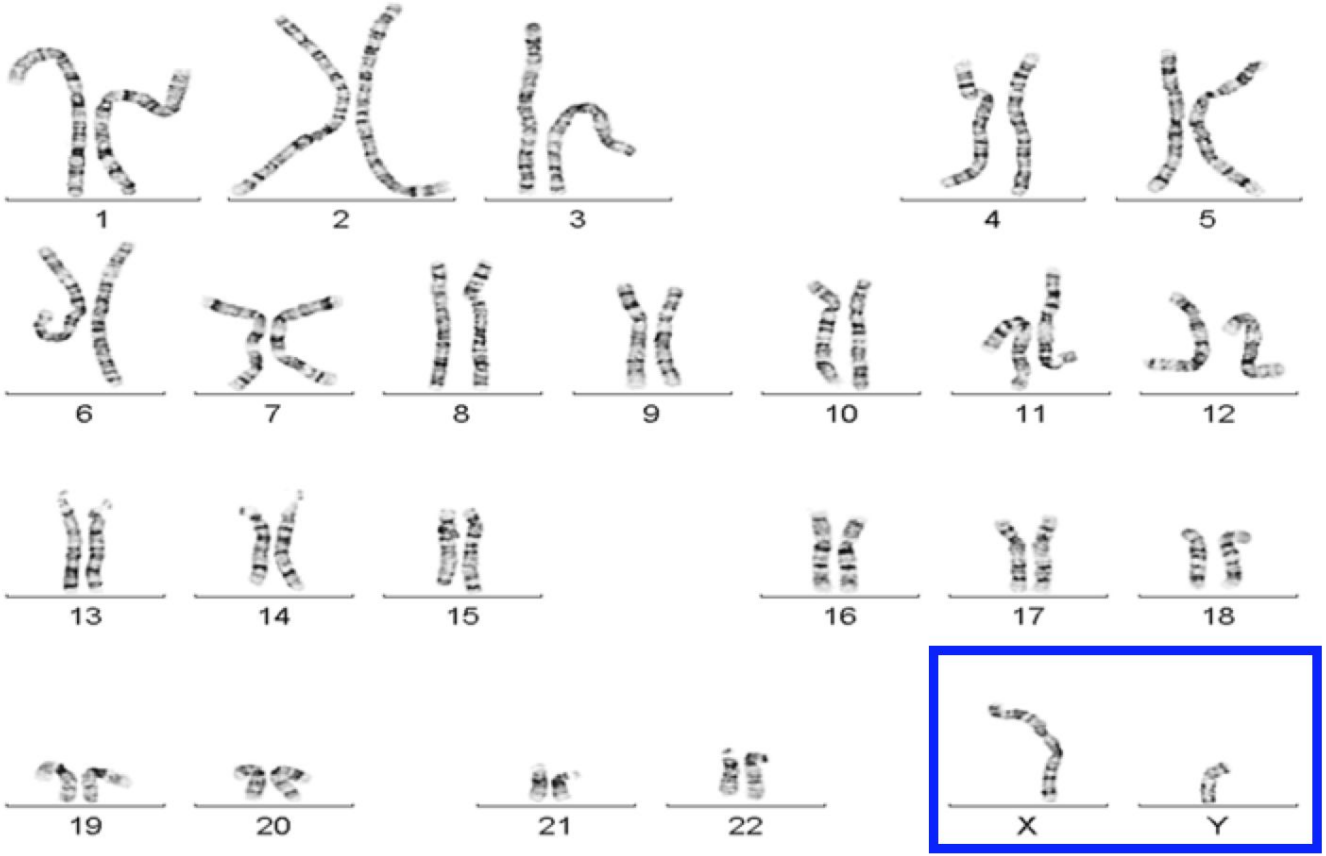
ANALYZE ALL THE GENOMES

Sex chromosomes share sequence similarity

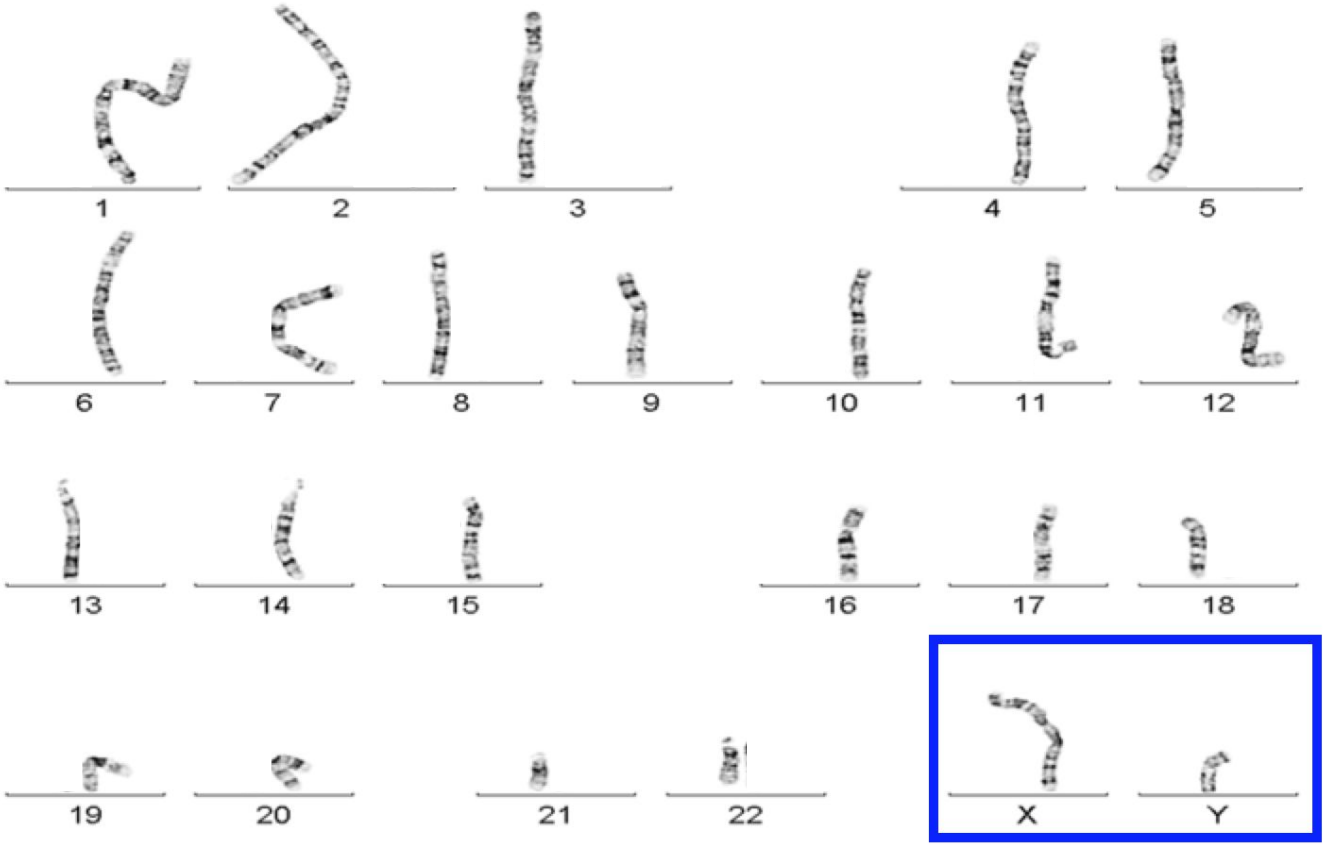
- The X and Y chromosomes share sequence similarity due to shared evolutionary ancestry that affects alignments and quantification of NGS data
- PARs share 100% homology



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

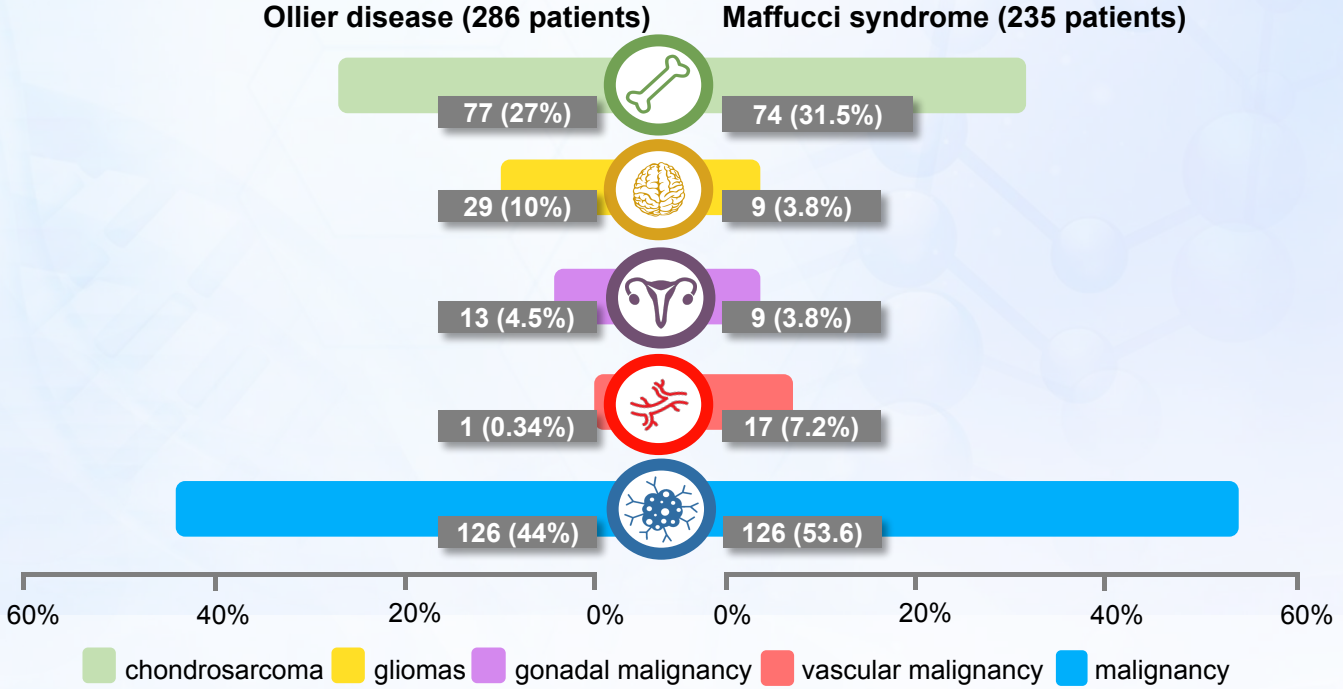
Nara Sobreira (Johns Hopkins University)

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

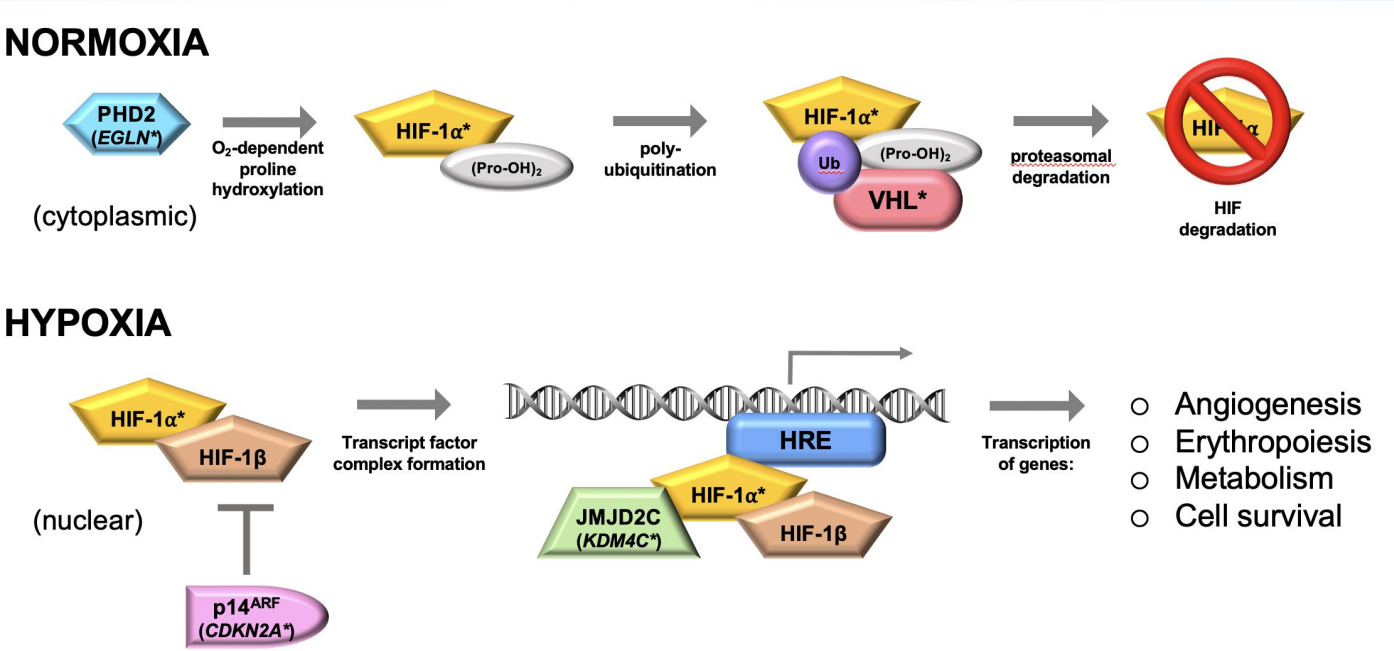
Renan Martin
Nara Sobreira
Johns Hopkins University School of Medicine

Scientific question

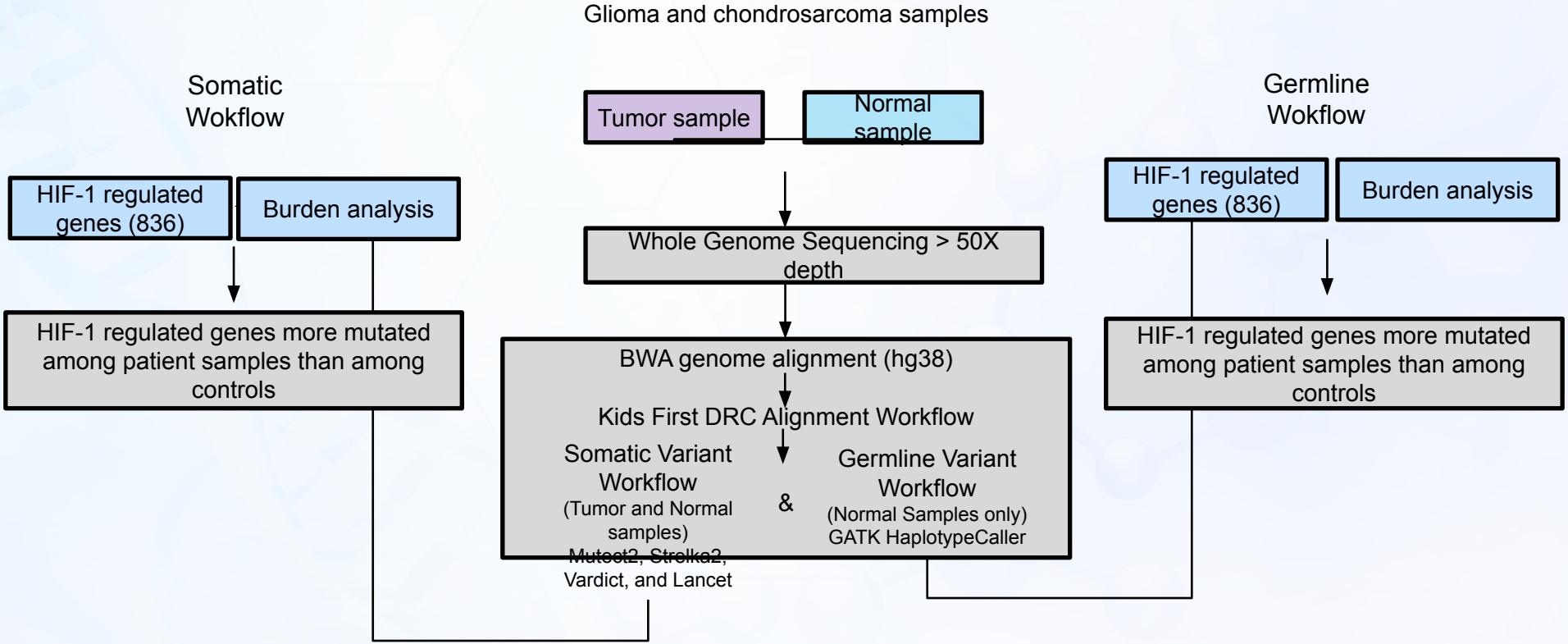
- Are pathogenic variants in genes related to HIF-1 pathway mutated in patients with Ollier disease and Maffucci syndrome and in patients with isolated forms of gliomas and chondrosarcomas?



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

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 Anno...

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?

AAA

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

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**

Search Studies

KF-DSD, Neuroblastoma...

Domain

Select All | None

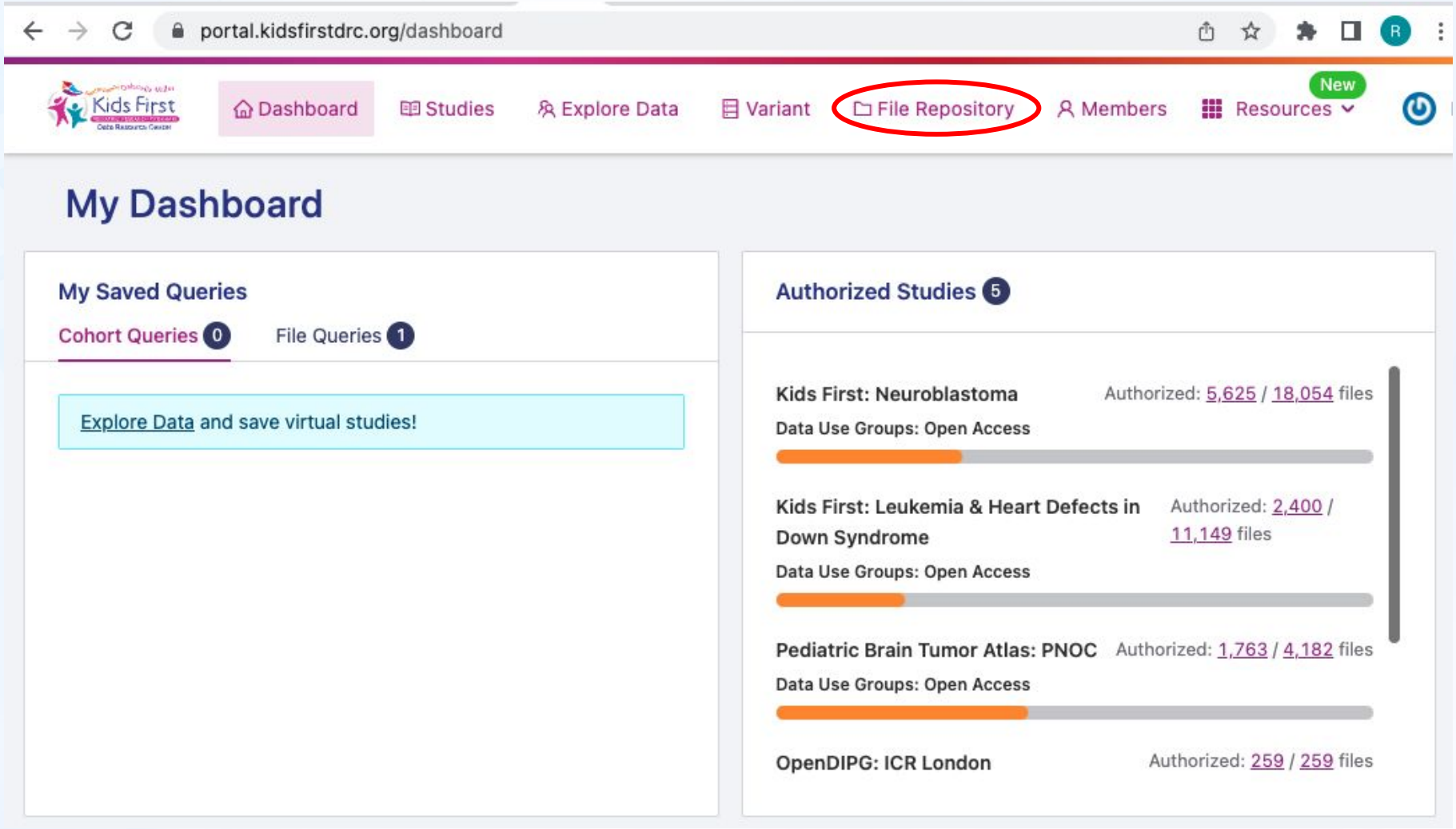
- Birth Defect 16
- Cancer 10

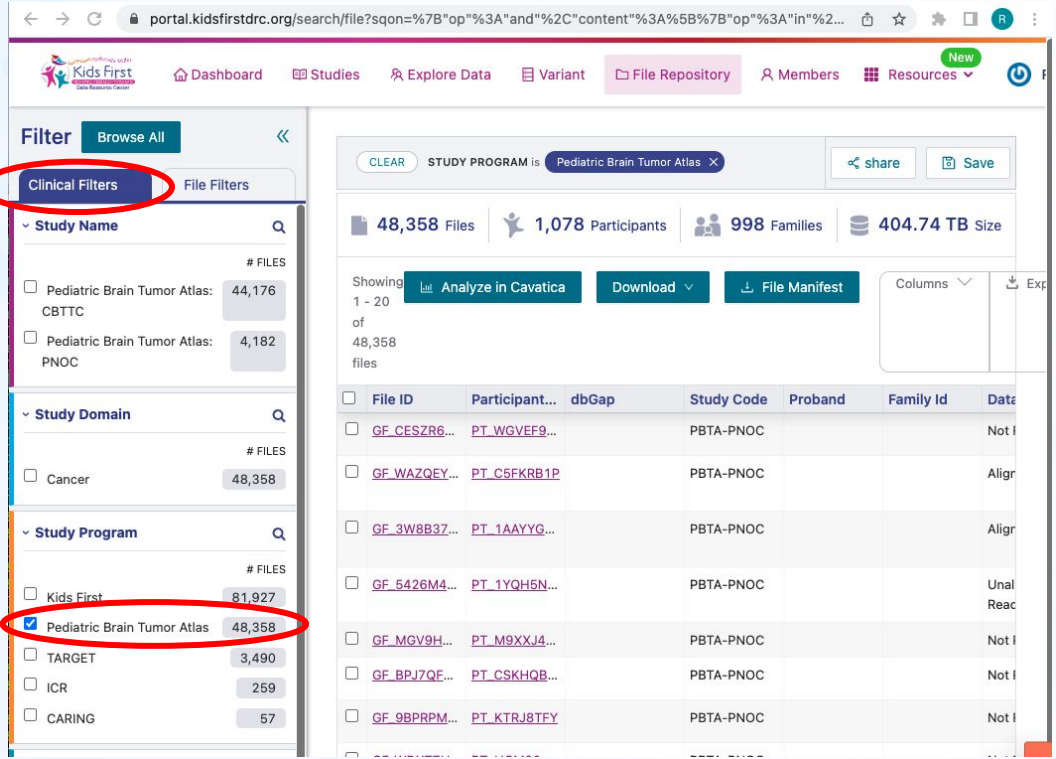
Program

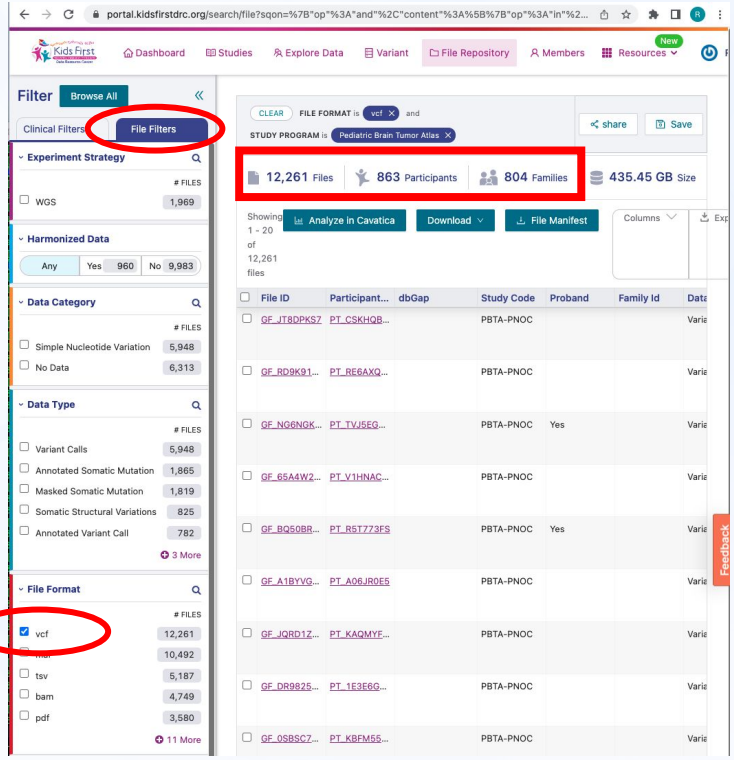
Select All | None

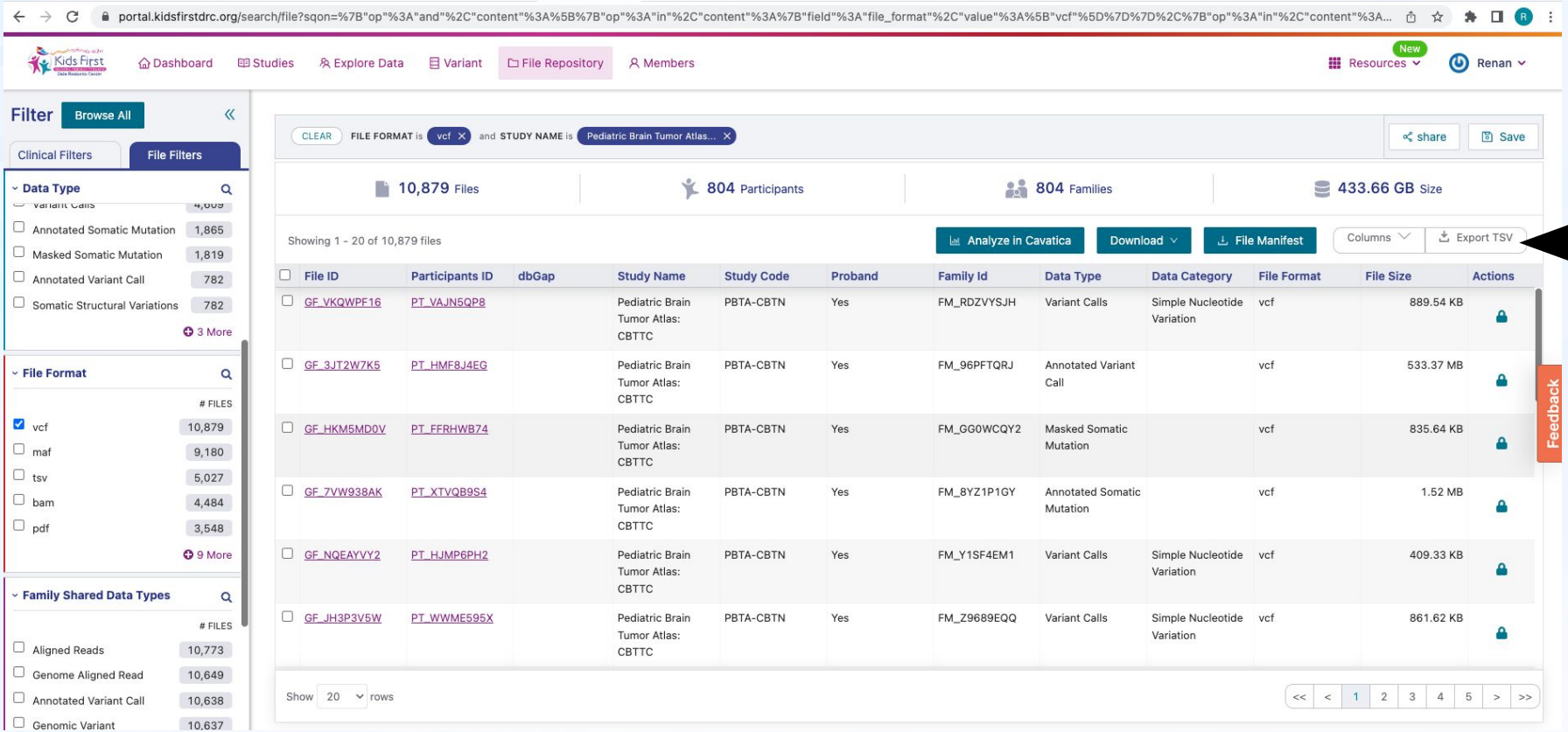
- Kids First 24
- Pediatric Brain Tumor Atlas 2
- TARGET 2
- CARING 1
- ICR 1

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

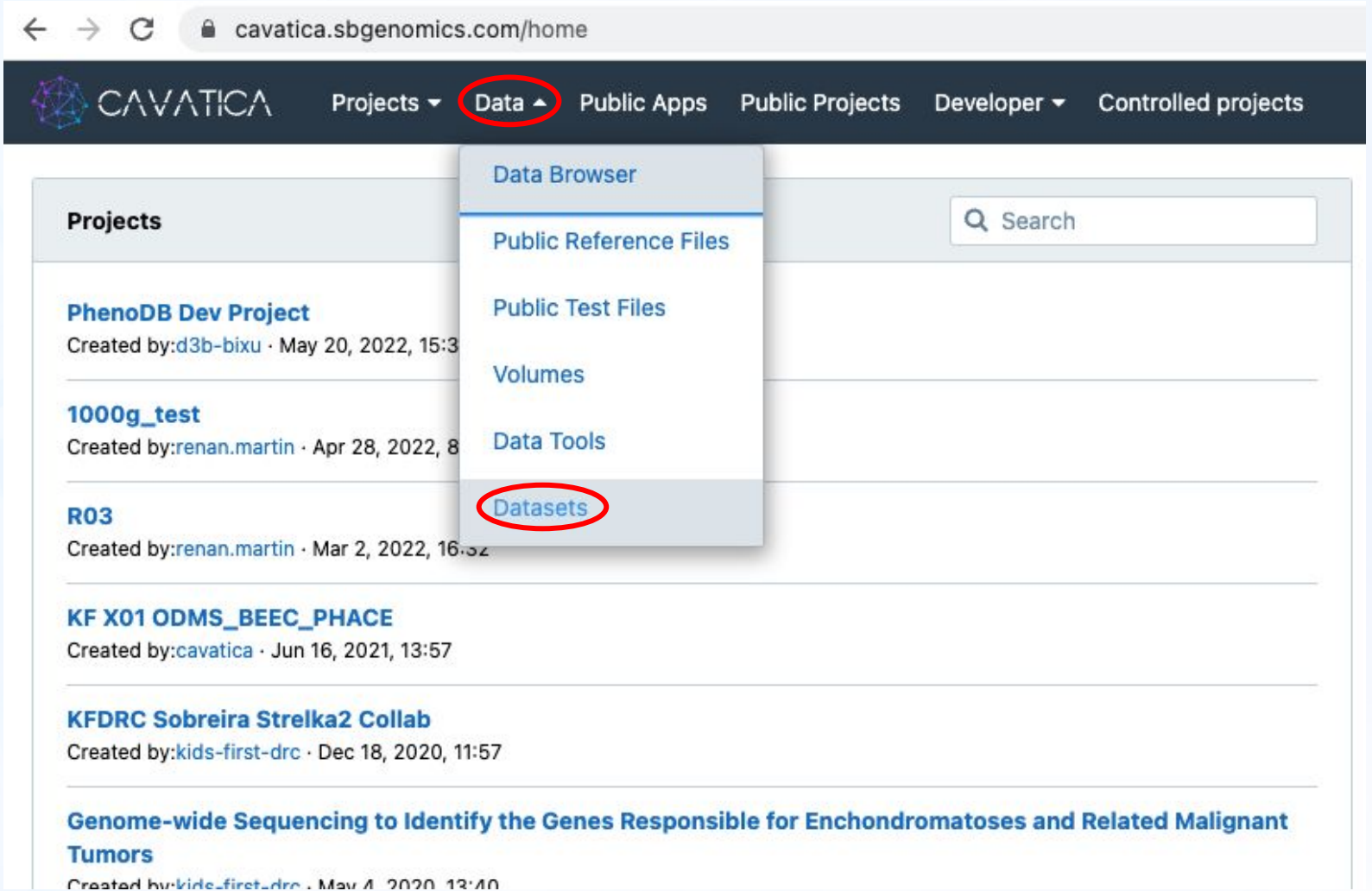

portal.kidsfirstdrc.org/dashboard
Kids First
Dashboard
Studies
Explore Data
Variant
File Repository
Members
New
Resources
My Dashboard
My Saved Queries
Cohort Queries 0
File Queries 1
Explore Data and save virtual studies!
Authorized Studies 5
Kids First: Neuroblastoma
Authorized: 5,625 / 18,054 files
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
Data
GF_CESZR6...
PT_WGVEF9...
PBTA-PNOC
GF_WAZQEY...
PT_C5FKRB1P
GF_3W8B37...
PT_1AAYYG...
GF_5426M4...
PT_1YQH5N...
GF_MGV9H...
PT_M9XXJ4...
GF_BPJ7QF...
PT_CSKHQB...
GF_9BPRPM...
PT_KTRJ8TFY


portal.kidsfirstdrc.org/search/file?sqon=%7B"op"%3A"and"%2C"content"%3A%5B%7B"op"%3A"in"%2...
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...
GF_NG6NGK...
PT_TVJ5EQ...
Yes
GF_65A4W2...
PT_V1HNAC...
GF_BQ50BR...
PT_R5T773FS
GF_A1BYVG...
PT_A06JR0E5
GF_JQRD1Z...
PT_KAQMYF...
GF_DR9825...
PT_1E3E6G...
GF_0SBSC7...
PT_KBFM55...
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, 16:32
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

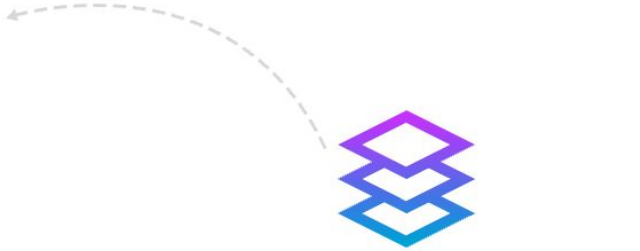
CAVATICA | Projects | Data | Public Apps | Public Projects | Developer | Controlled projects | renan.martin

Datasets

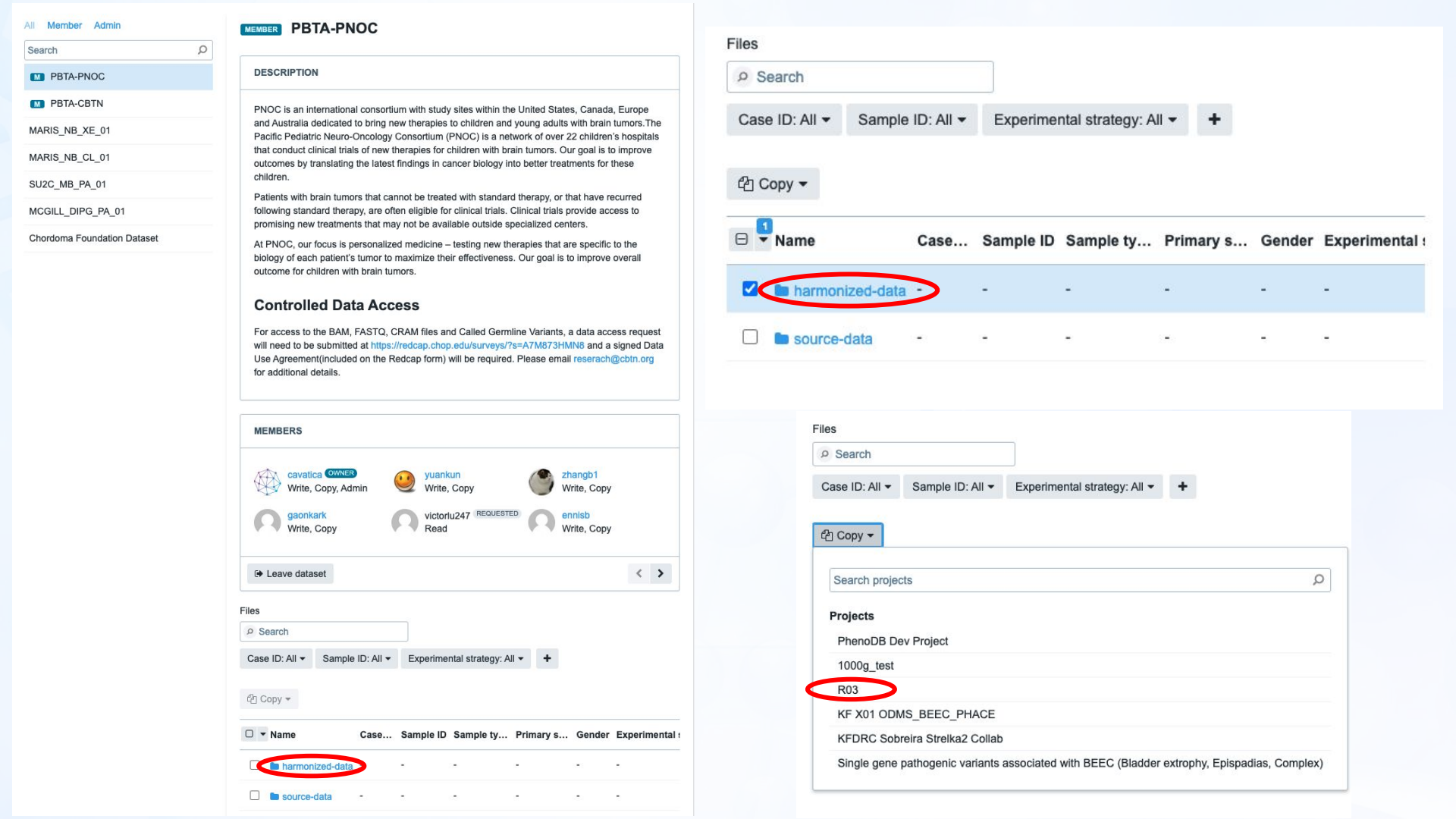
All Member Admin

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
PBTA-PNOC
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
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

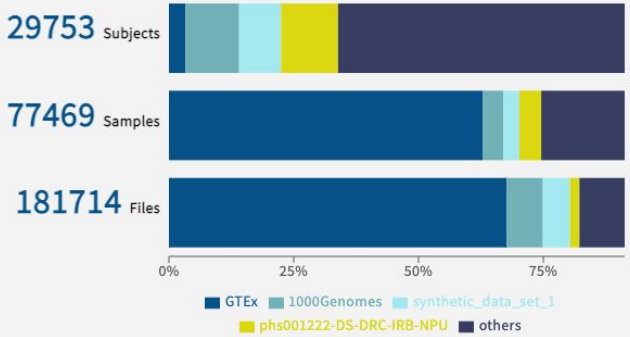
The AnVIL

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.

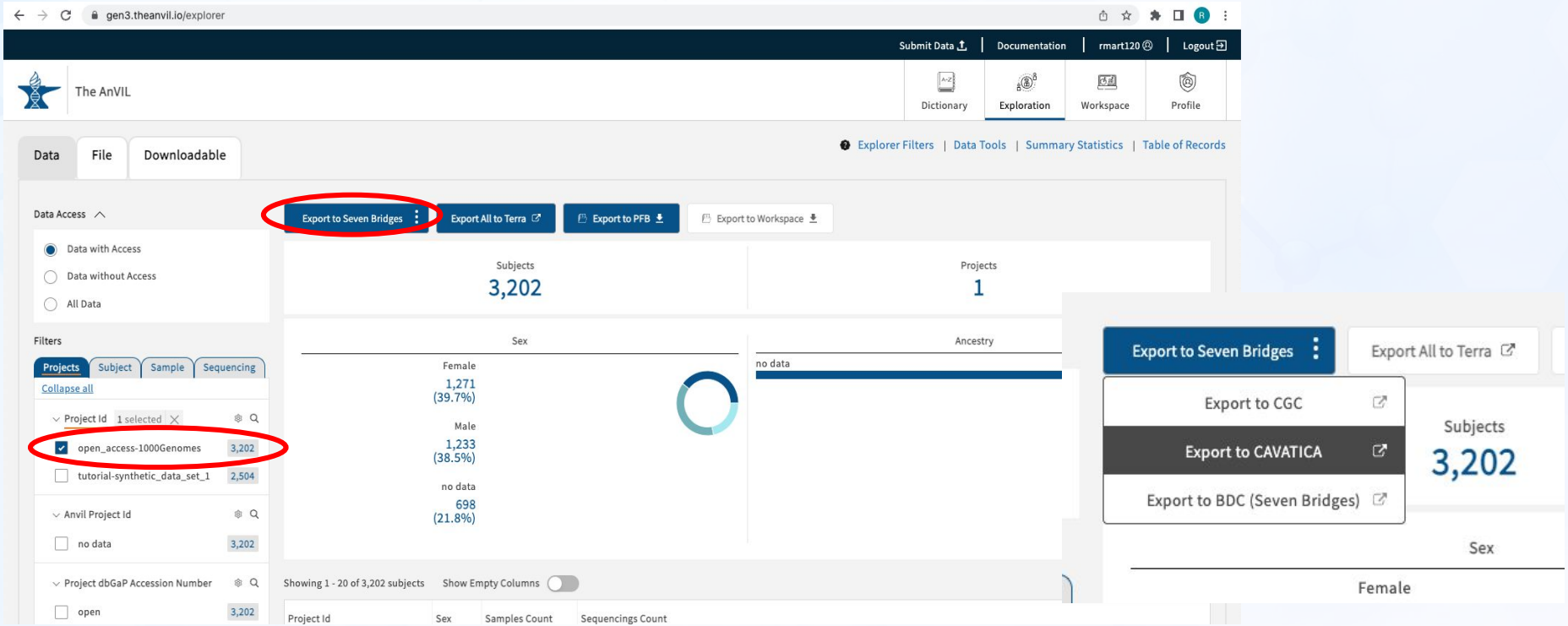
Submit Data

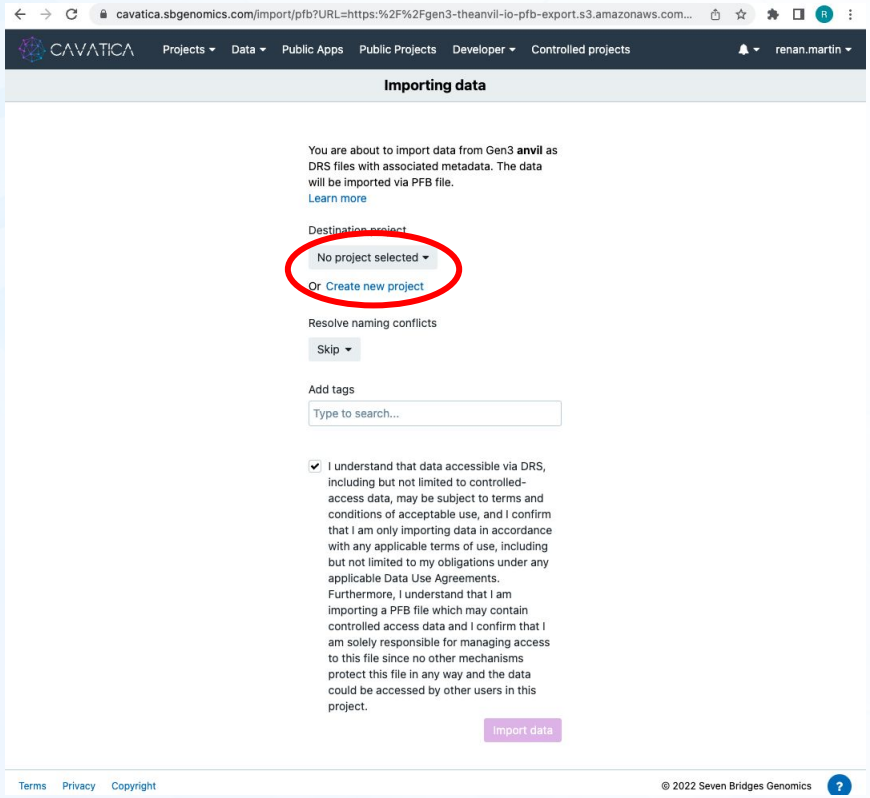


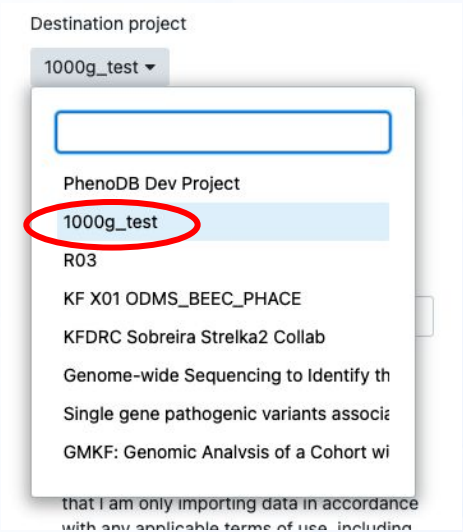
Define Data Field

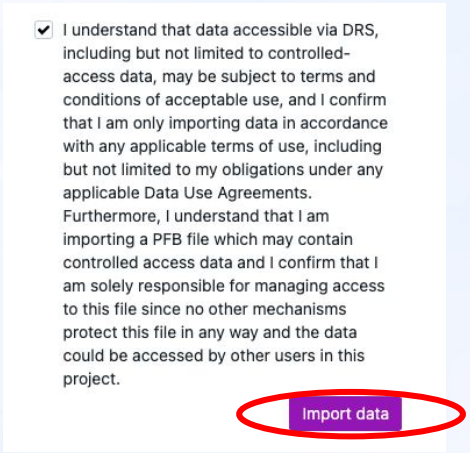
Explore Data

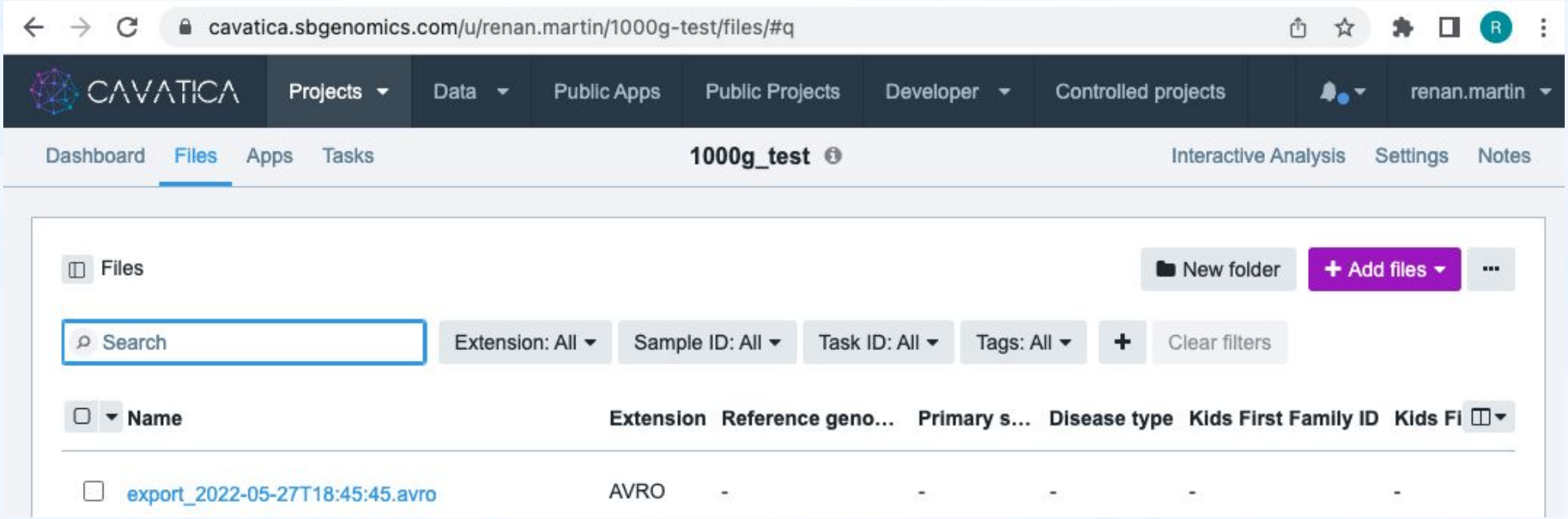
Analyze Data


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
Export to Seven Bridges
Export All to Terra
Export to PFB
Export to Workspace
Data with Access
Data without Access
All Data
Subjects
3,202
Projects
1
Filters
Sex
Ancestry
Female
1,271
(39.7%)
Male
1,233
(38.5%)
no data
698
(21.8%)
no data
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
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

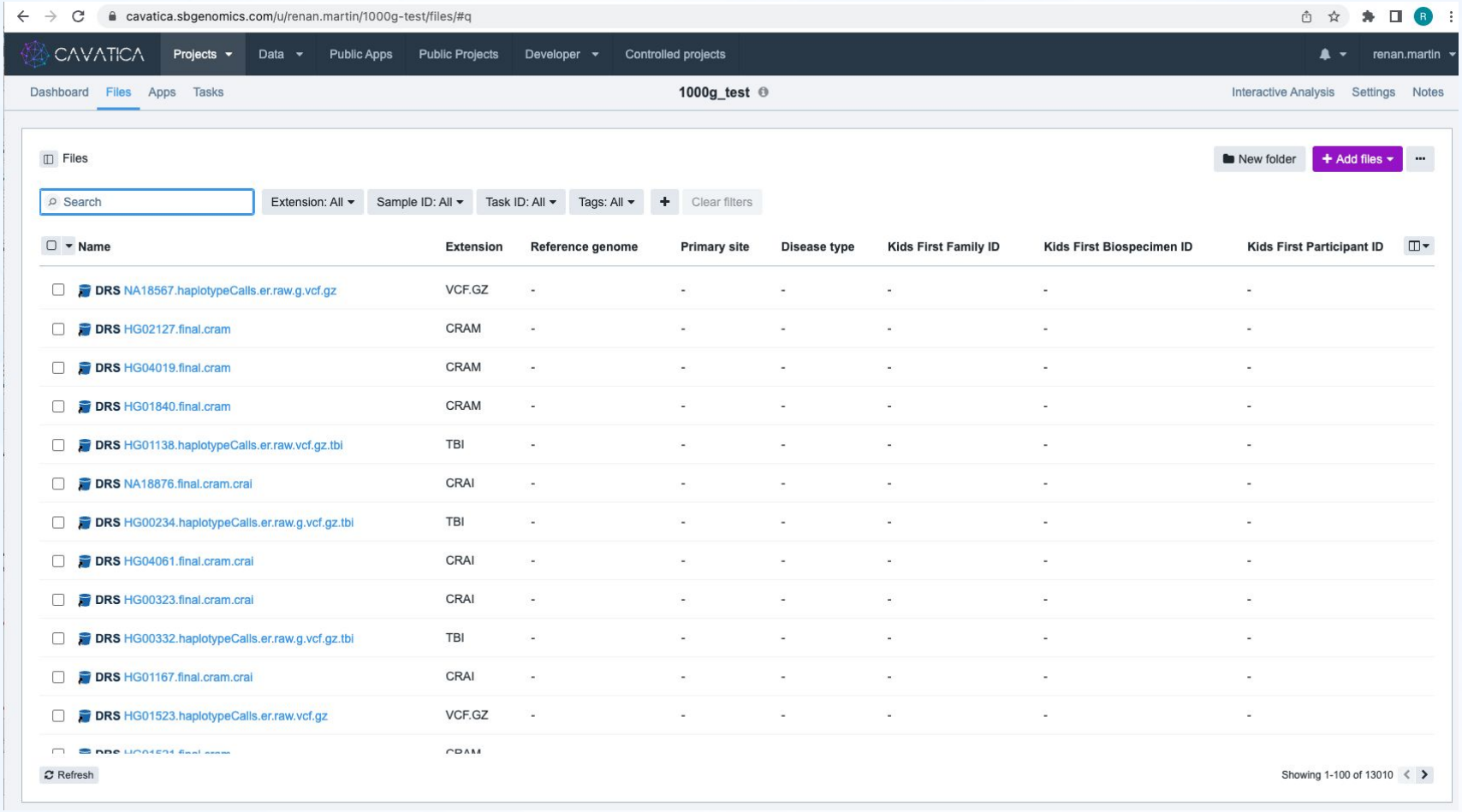

cavatica.sbgenomics.com/import/pfb?URL=https:%2F%2Fgen3-theanvil-io-pfb-export.s3.amazonaws.com...
CAVATICA
Projects
Data
Public Apps
Public Projects
Developer
Controlled projects
renan.martin
Importing data
You are about to import data from Gen3 anvil as DRS files with associated metadata. The data 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
© 2022 Seven Bridges Genomics


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


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



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

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

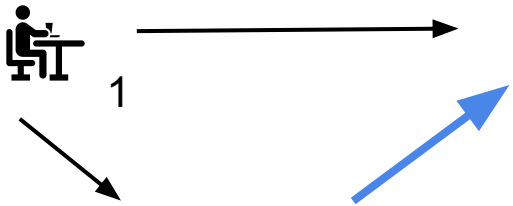
Key Concepts

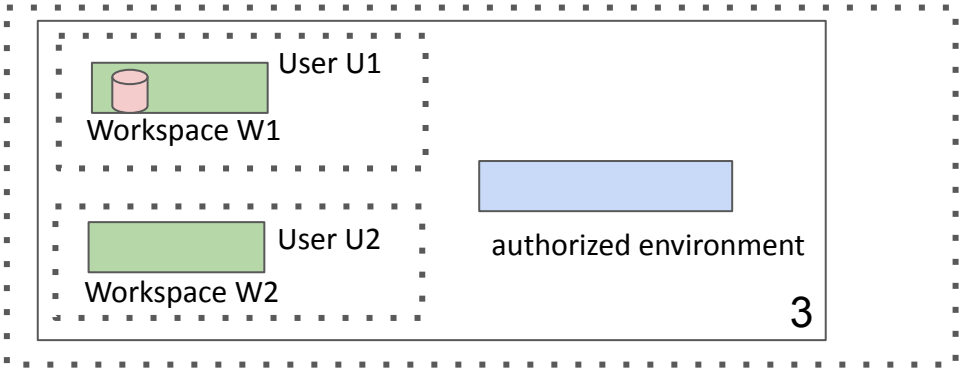
Project Sponsor - Entity responsible for data and platform governance.

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

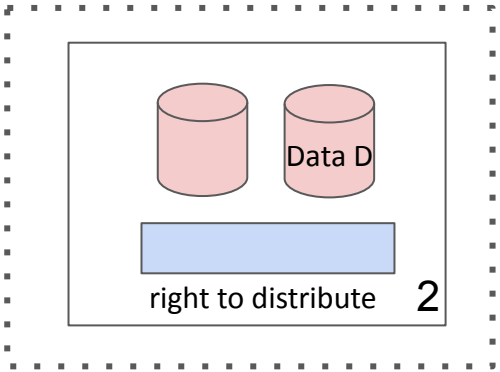
Authorized environment - the project sponsor determines whether the target cloud platform has appropriate security, compliance and governance to support the analysis of the data on the cloud platform by authorized researches

Overview





Cloud Platform B boundary



Cloud Platform A boundary

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.

Cloud platform portal

Cloud platform boundary

Workspace for user

Security and compliance boundary

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

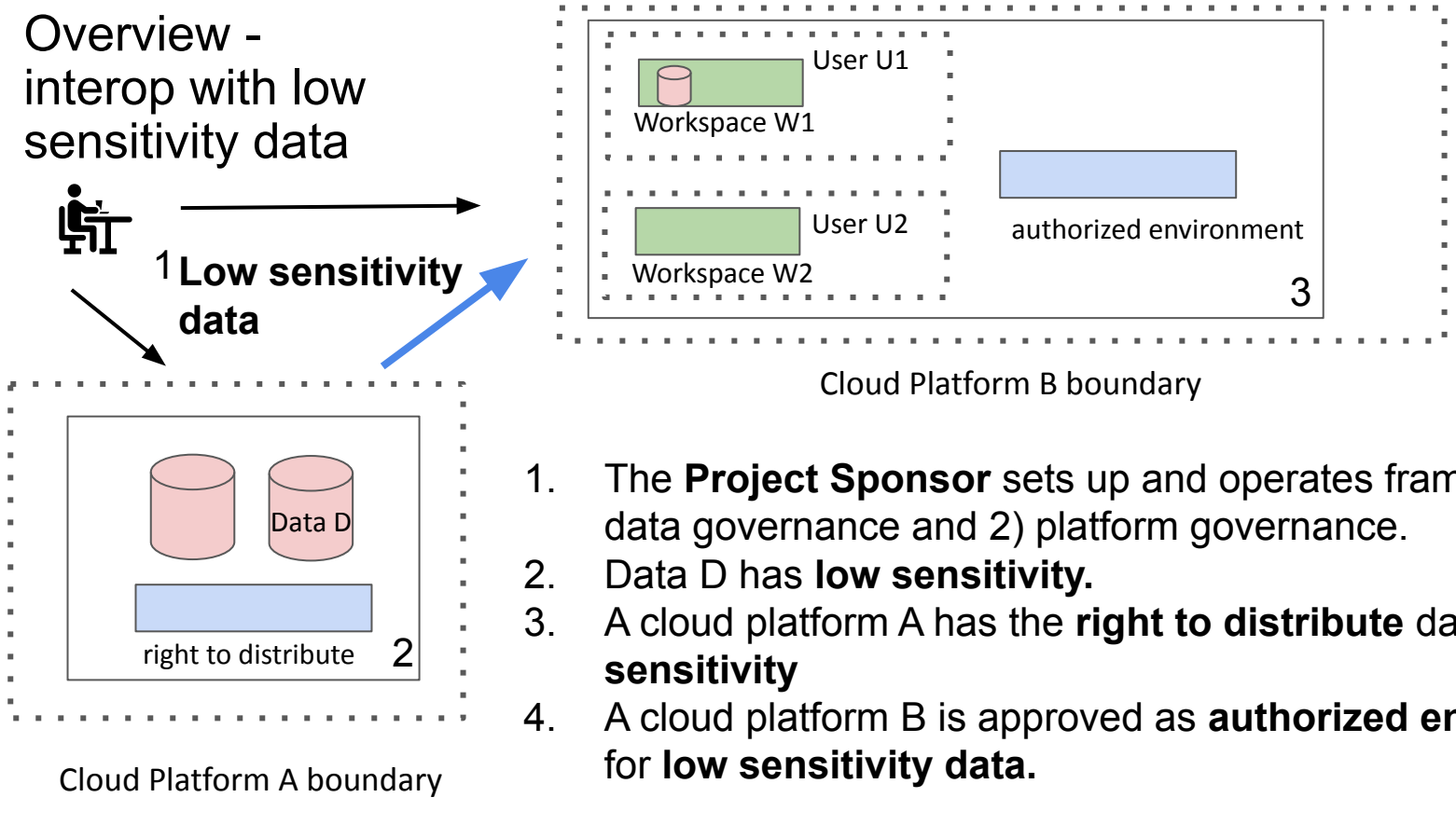
- Seek approval for the current NCPI Platforms as authorized environments for data from one of the other NCPI Platforms.
- Seek approval for selected other platforms that follow NIST 800-53 Moderate as authorized environments for one or more NCPI platforms.

Phase 2 - Interop for Low Sensitivity Data

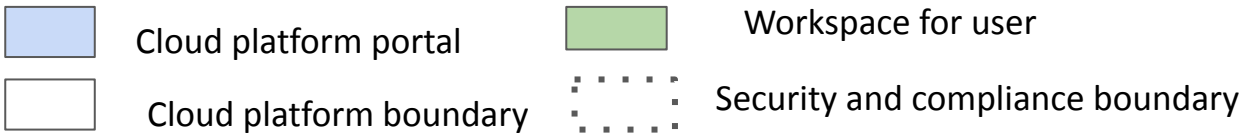
Basic Idea

- 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.

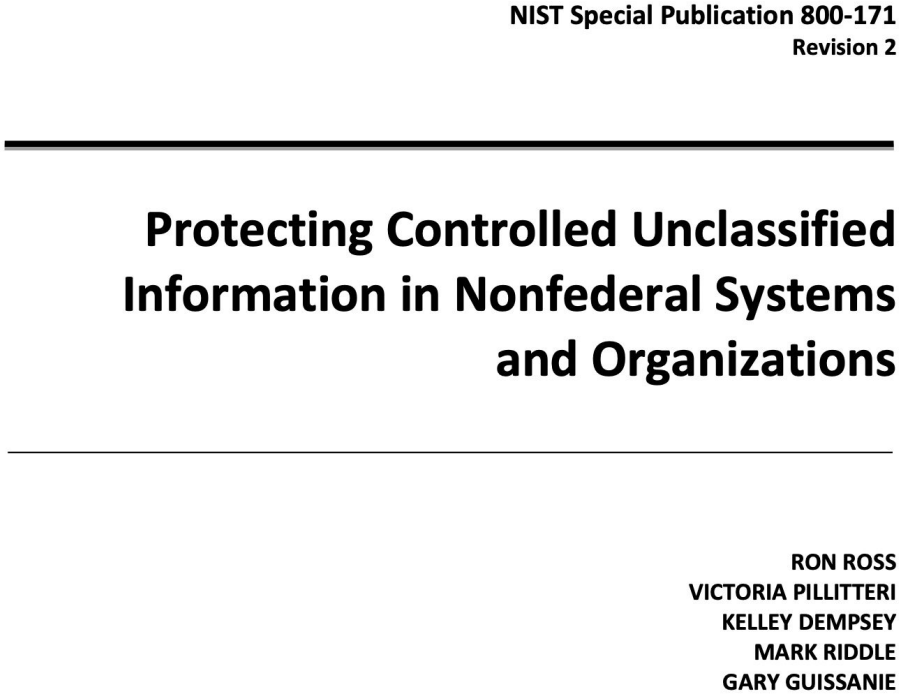
Overview - interop with low sensitivity data



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.**

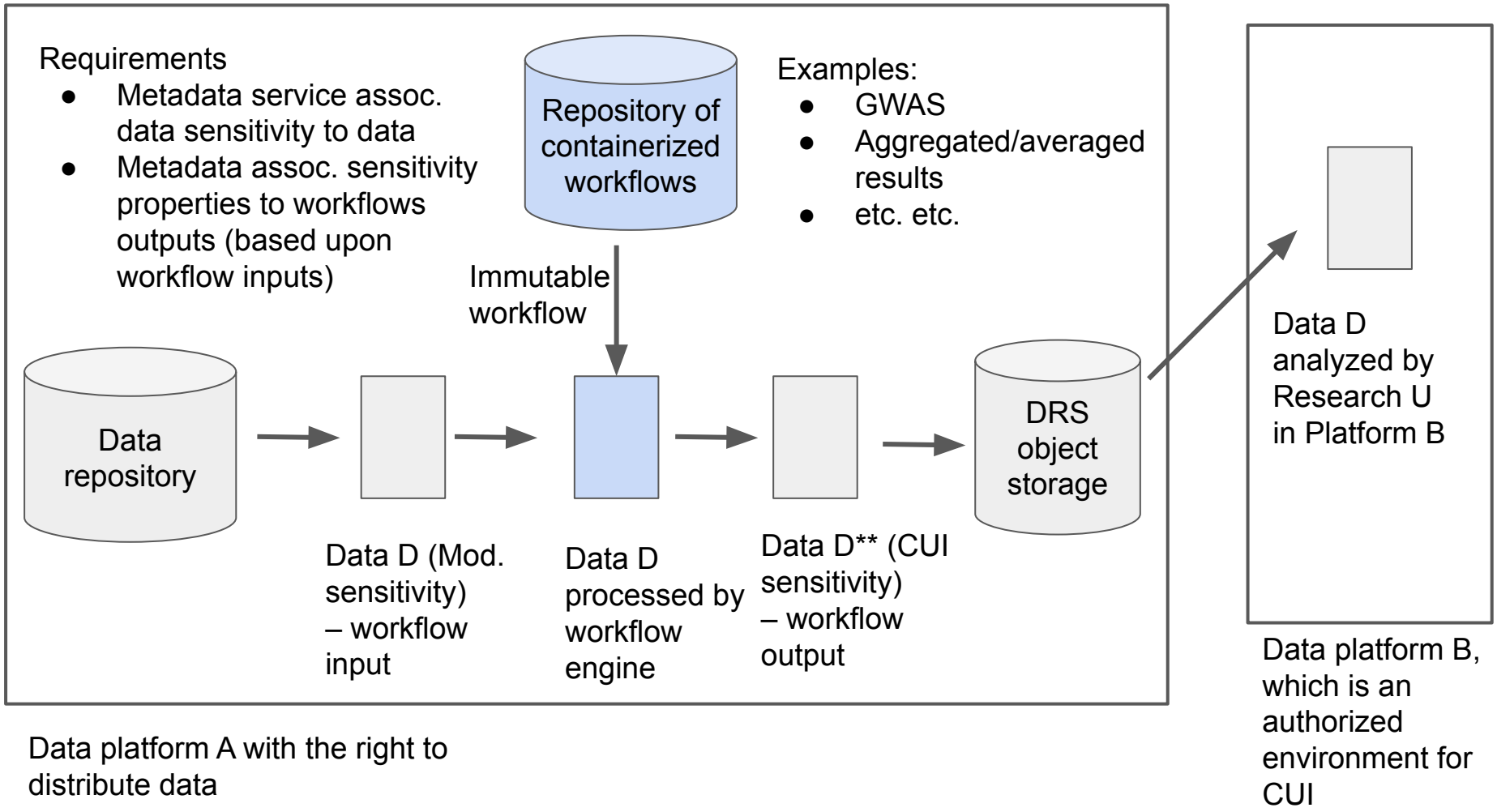


Controlled Unclassified Information (CUI)



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

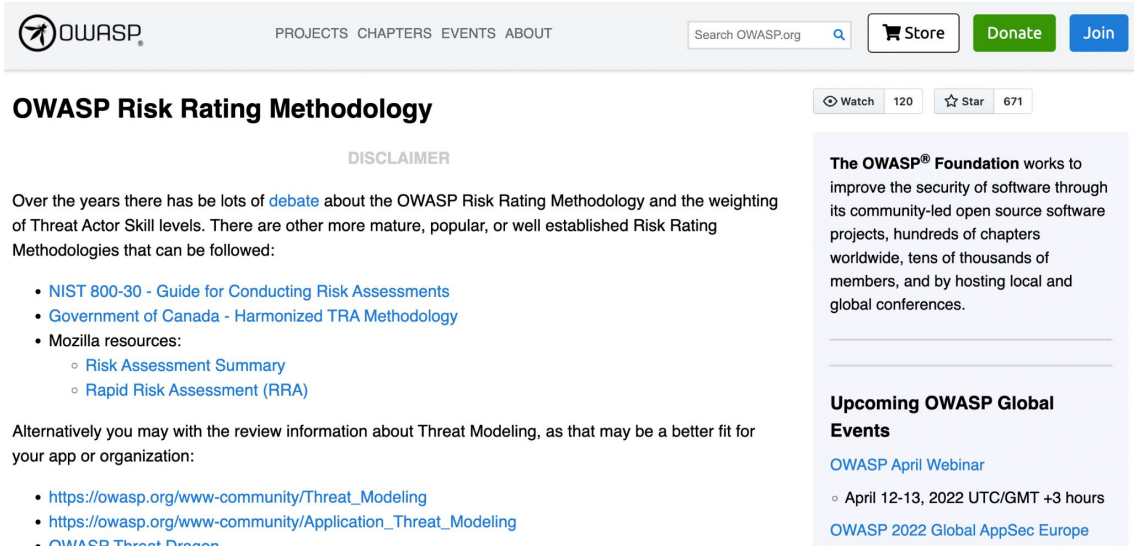
A very simple use case of low sensitivity data being generated by applying approved workflows to genomic data.


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



- 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?

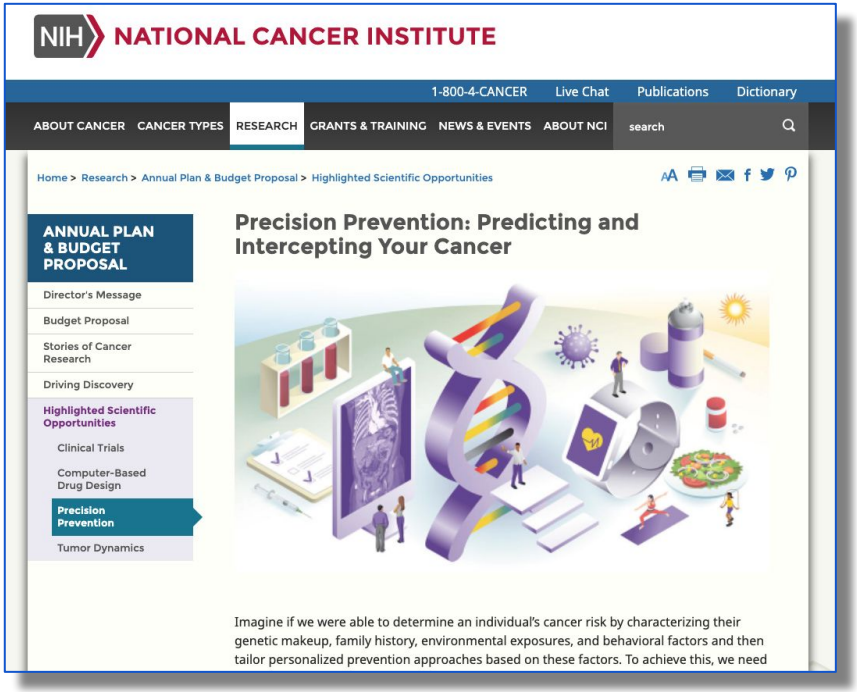


Image credit:
https://www.cancer.gov/research/annual-plan/scientific-topics/precision-prevention

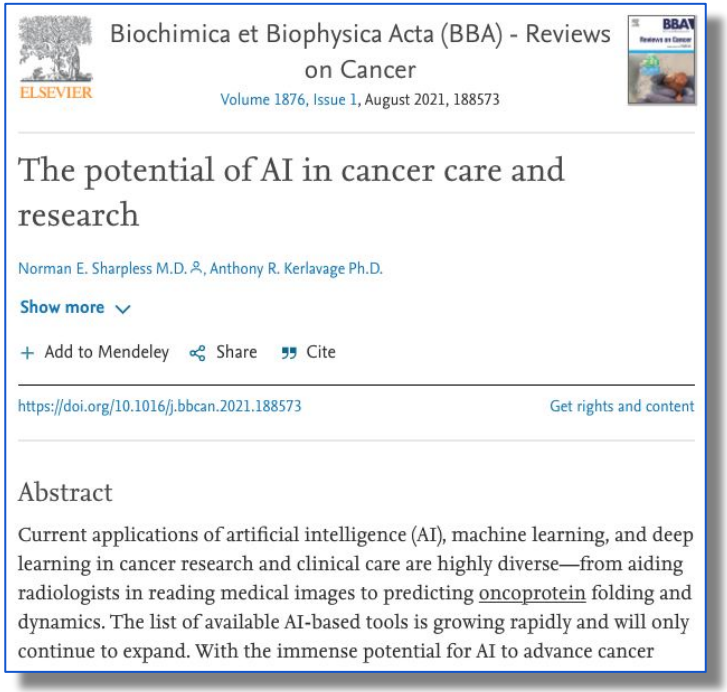


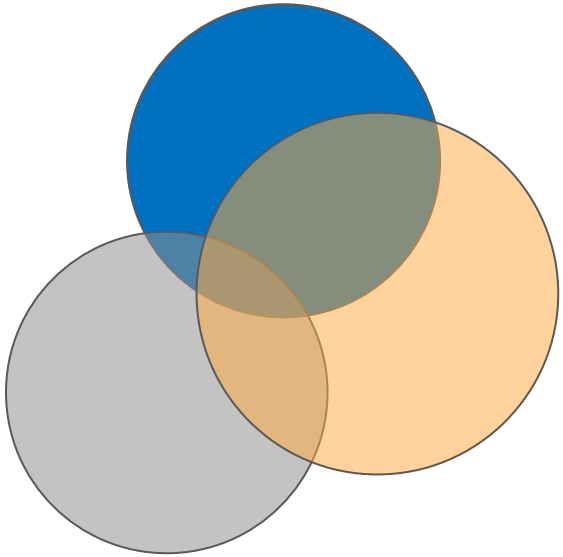
Image credit:
https://www.sciencedirect.com/science/article/abs/pii/S0304419X21000706

System Interoperability mission to empower diverse researchers

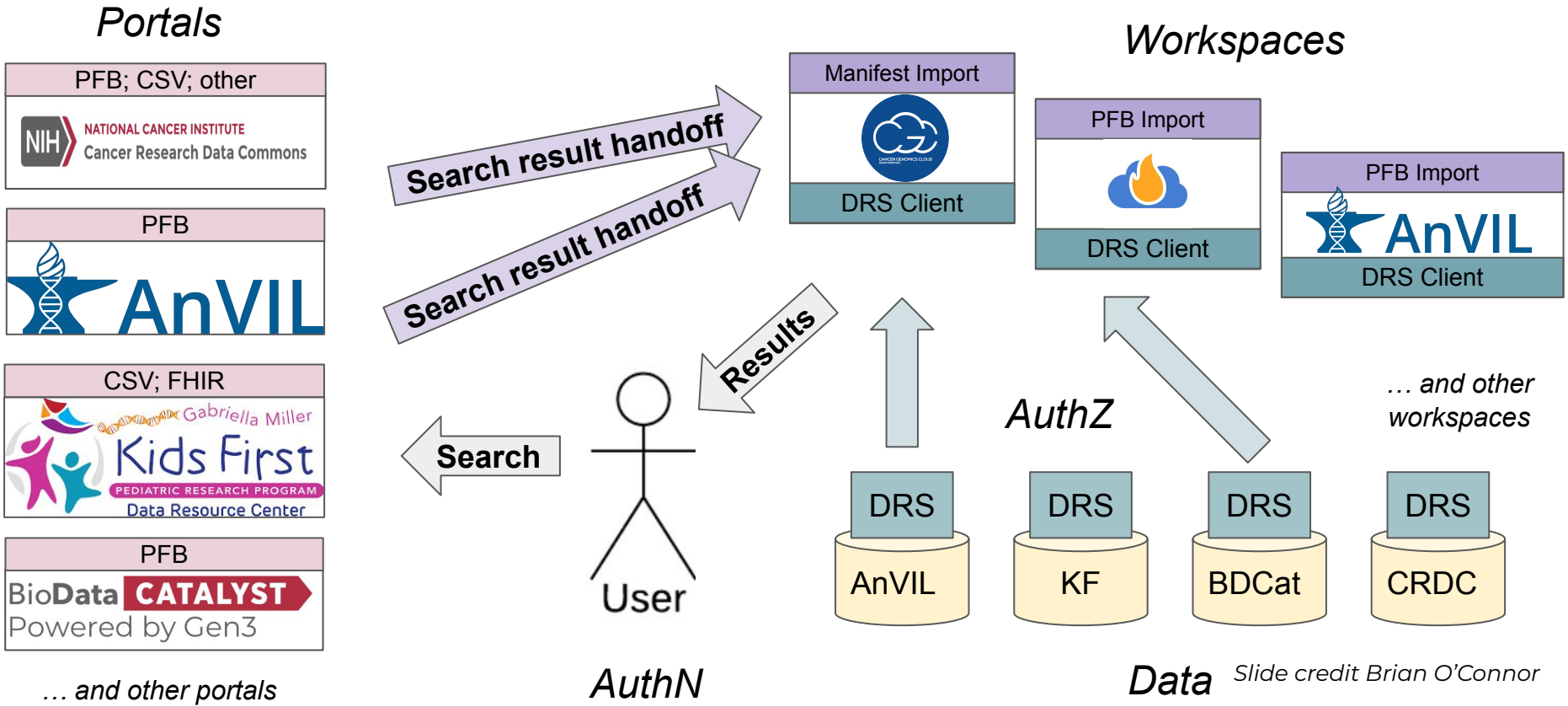
Empower **diverse researchers** to complete **scientific projects** across ICs by spearheading **technical improvements** across cloud "stacks”

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



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

Wilson McKerrow, David Fenyö, et al

Cloud costs funded via Collaborative Project

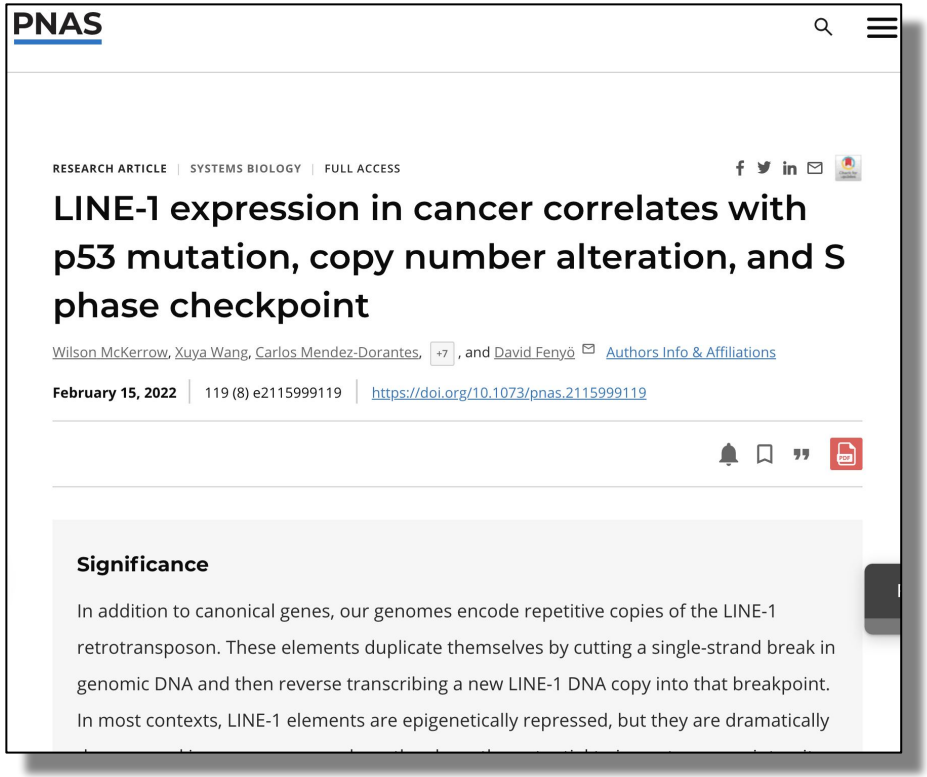
CRDC

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

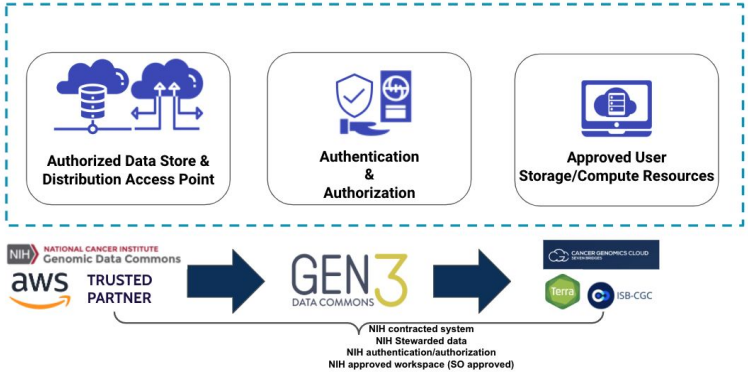
February 15, 2022 | 119 (8) e2115999119 | https://doi.org/10.1073/pnas.2115999119

Significance

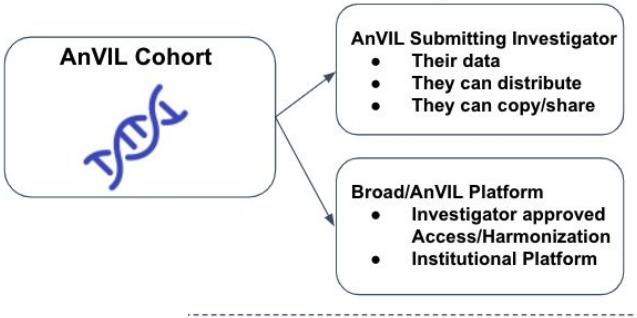
In addition to canonical genes, our genomes encode repetitive copies of the LINE-1 retrotransposon. These elements duplicate themselves by cutting a single-strand break in genomic DNA and then reverse transcribing a new LINE-1 DNA copy into that breakpoint. In most contexts, LINE-1 elements are epigenetically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

NCPI is trailblazing interoperability policy as well

CRDC



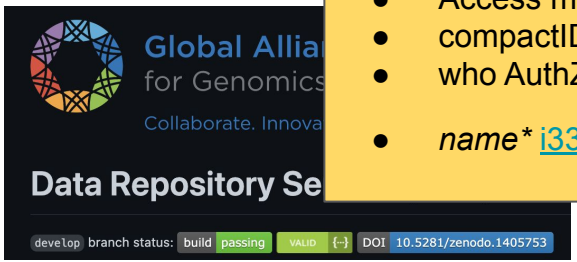
AnVIL



Together we've made it easier for the next researcher

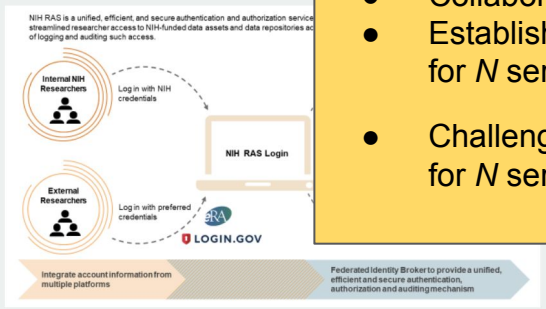
Agreed on a finite set of technical methods

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

Minimal data attributes

Manifests (PFB or CSV)

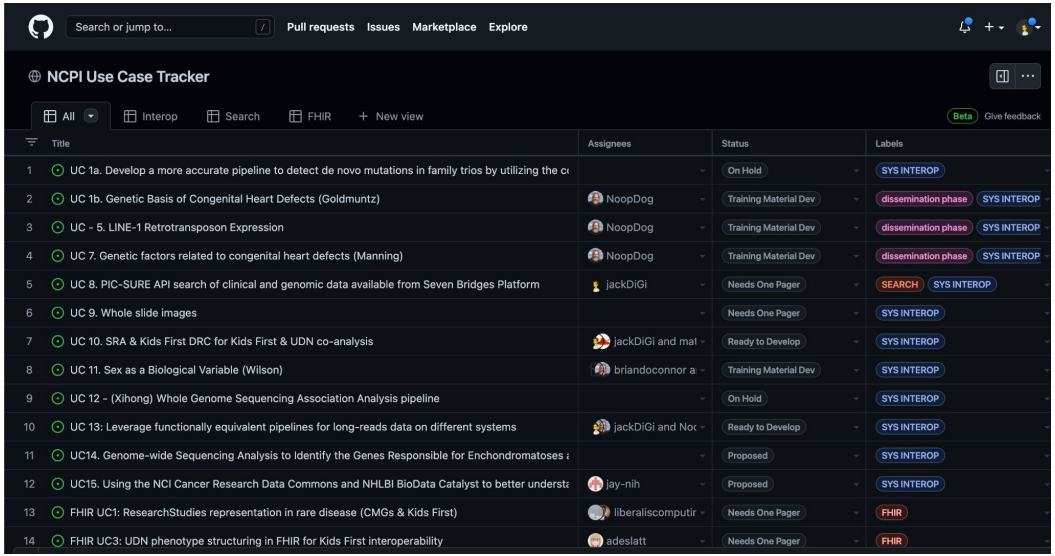
Attribute	Definition
drs_uri	DRS URI as defined by GA4GH DRS spec for pointers to file objects.
study_registration	External source from which the identifier included in study_id originates (answer can be dbGaP for example)
study_id	Unique identifier that can be used to retrieve more information for a 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

Happy to see how things have progressed

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

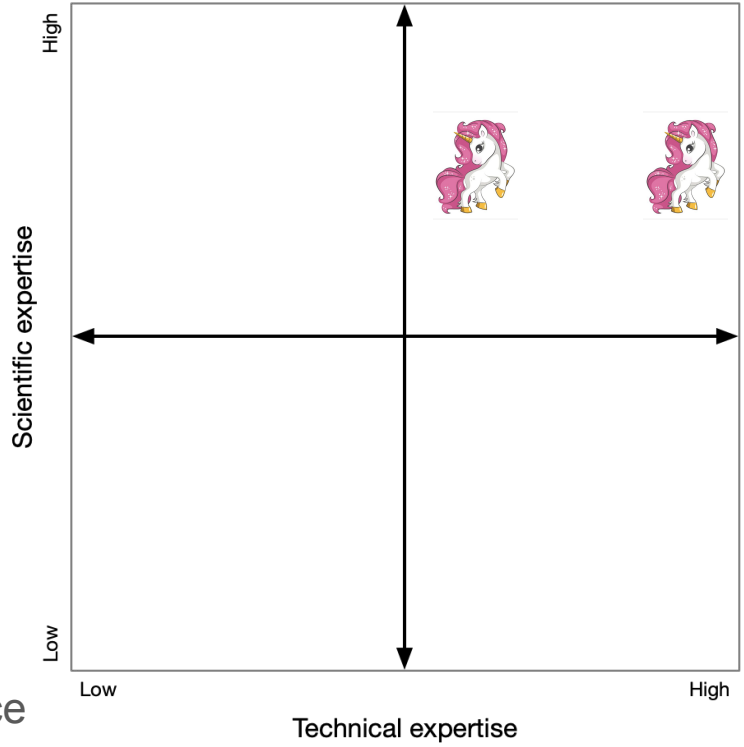
Funding & roadmap management was also *very challenging*

Use cases are now **publishing** manuscrip[illegible]

More interop is happening

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

Tech and policy are hardening to **reduce barr[illegible]s** [illegible]science



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