



NIH Workshop on Cloud-Based Platforms Interoperability
Remote Event - April 16, 2020

NCPI Workshop

NIH Workshop on Cloud Platforms Interoperability

Remote Event - Apr 16, 2020



NIH Workshop on Cloud-Based Platforms Interoperability
Remote Event - April 16, 2020

NCPI Workshop

NIH Workshop on Cloud Platforms Interoperability

Remote Event - Apr 16, 2020



Today's Agenda

Meeting Goals

- Updates on Progress, Roadmaps and Challenges
- Identify Action Plans (next steps) for ALL Identified Challenges

Agenda

- | | |
|---|-----------------------|
| • Welcome / Opening Remarks | 11:00 - 11:15pm (EDT) |
| • Working Group Updates | 11:15 - 12:15pm (EDT) |
| • Interoperability Use Case Updates | 12:15 - 1:00pm (EDT) |
| • <i>(Break)</i> | 1:00 - 2:00pm (EDT) |
| • General Interoperability Solution (long-term) | 2:00 - 2:50pm (EDT) |
| • Closing | 2:50 - 3:00pm (EDT) |



Opening Remarks

Host: Anthony Philippakis



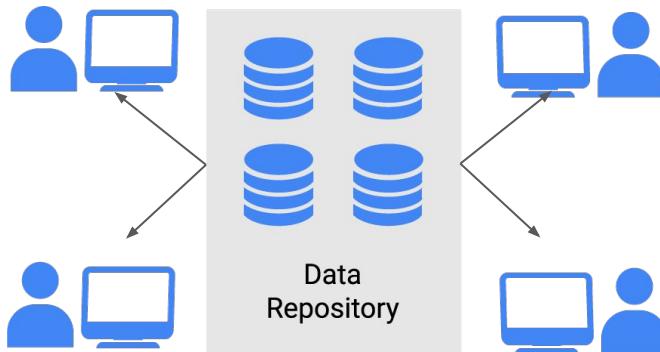
In Memoriam



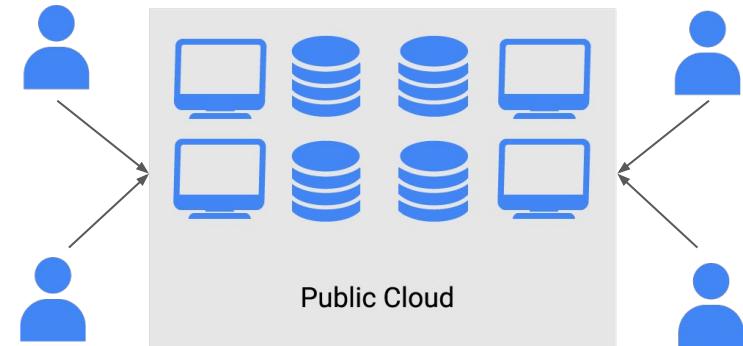
James Taylor
Software Visionary

Inverting the Model of Data Sharing

Old Way: Bring data to the researchers



New Way: Bring researchers to the data



Problems

Data sharing = data copying
Security
Accessibility
Inelastic

Solutions

Less Expensive
Audit Trails
Greater Accessibility
Elasticity



Inverting the Model of Data Sharing

There is a natural tension between the following goals

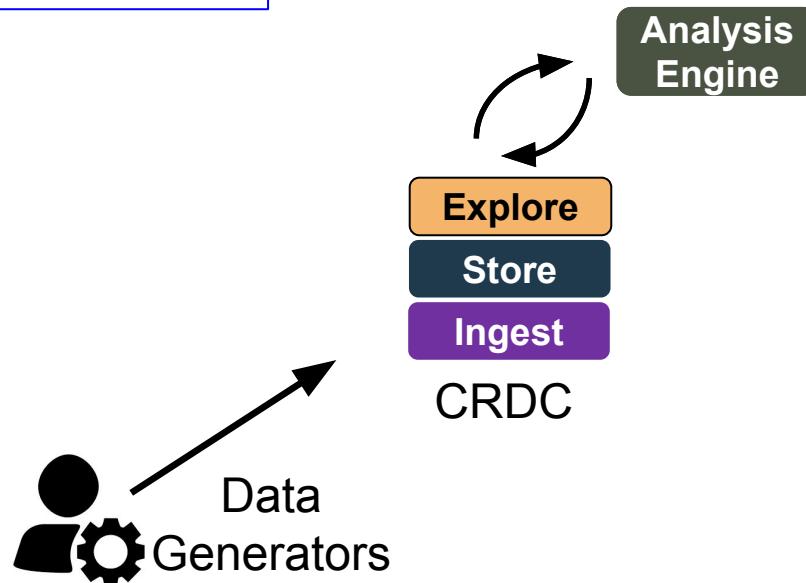
- We want an ecosystem of data holders and tool builders
 - *Neither desirable nor realistic for one monolithic entity to hold all of the world's genomic and clinical data*
- We want to easily combine diverse datasets
 - *We need to maximize statistical power, and jointly analyze orthogonal data types*

Answer: Build our systems in a federated and interoperable way

Making this Vision Real

NCI Data Flow:

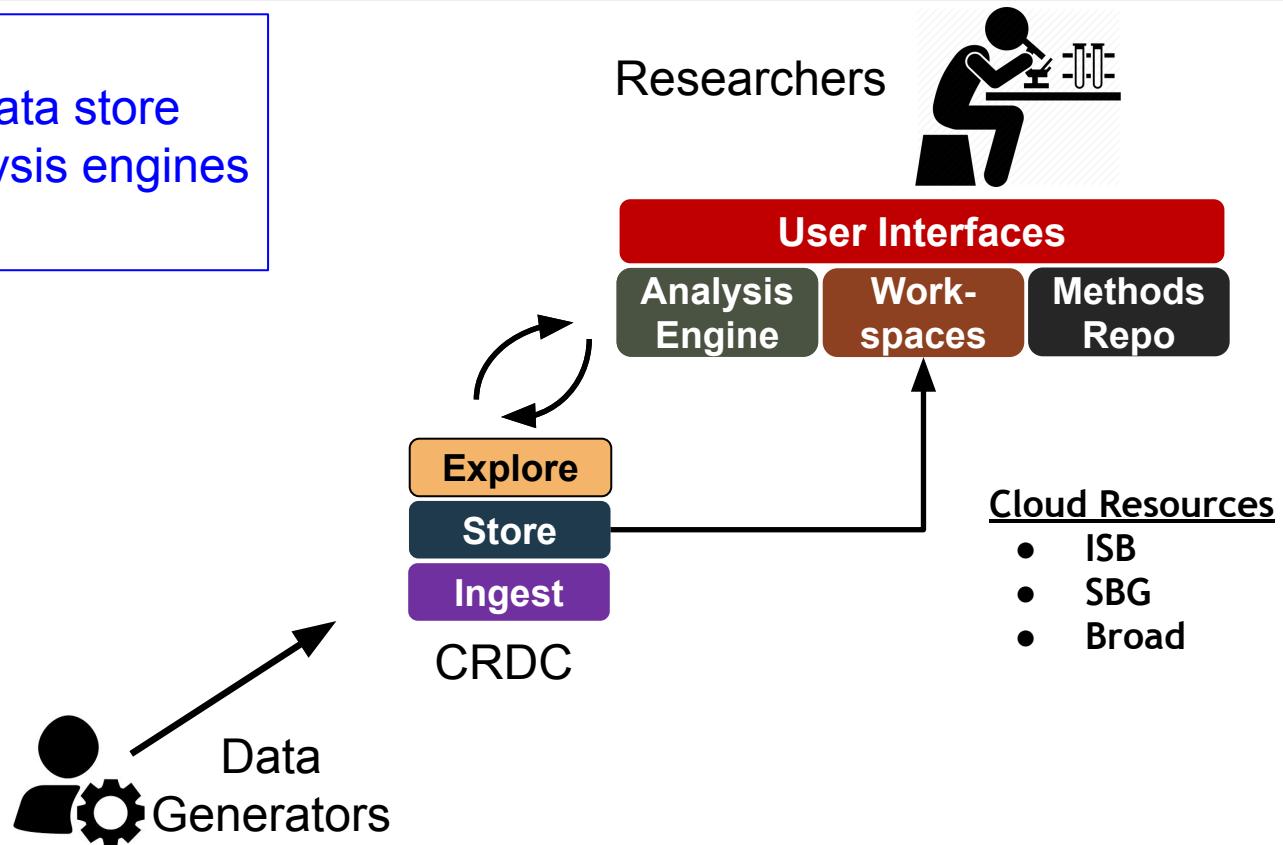
- Generators upload to data store
- Process data with analysis engines
- Indexing and search



Making this Vision Real

NCI Data Flow:

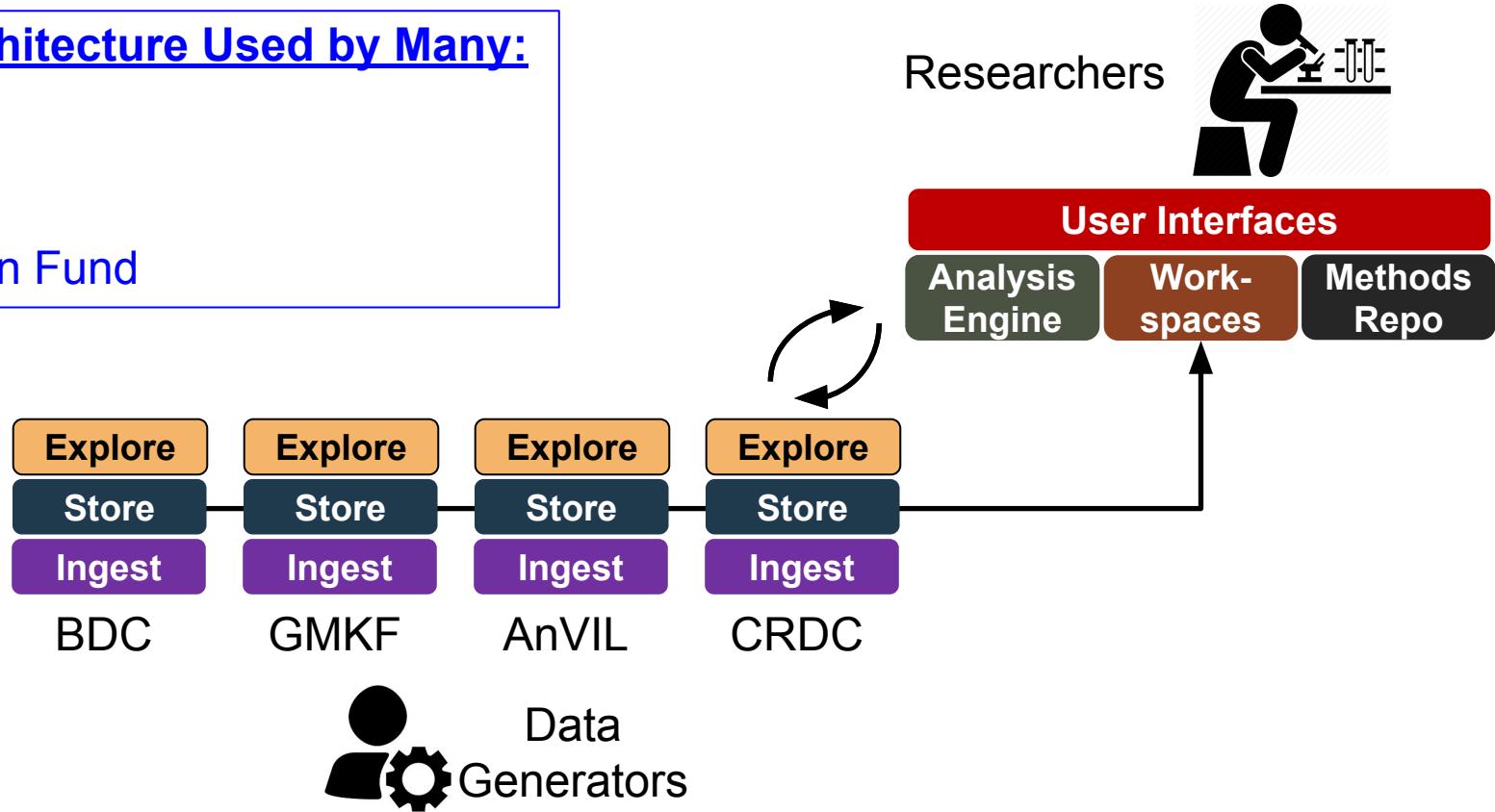
- Generators upload to data store
- Process data with analysis engines
- Indexing and search



Making this Vision Real

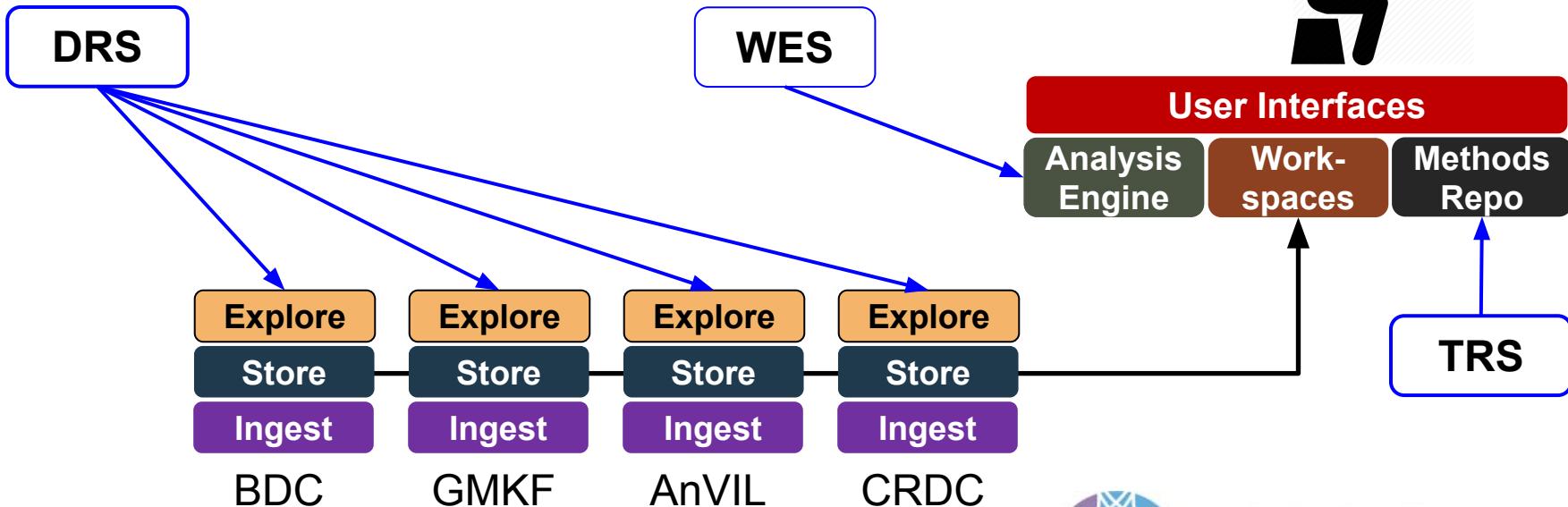
Same Architecture Used by Many:

- NCI
- NHLBI
- NHGRI
- Common Fund



Making this Vision Real

Standards



Passports
DUO



Data
Generators



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



What we have accomplished

- Clarity on architectural vision (BIG!)
 - Agree on basic parts list (data repo, workspace, etc)
 - Real plan for federation and interoperability
 - Can demo pulling data or tools from various repositories and pulling them into the same workspace.

What we have yet to do

- For the most part, interoperability is still only a prototype.
- No ability to search across repositories (we need “data aggregators” that can see across data repositories)
- AuthN and AuthZ are still complex
 - RAS in its infancy. Data use ontologies unevenly utilized.



What we have accomplished

- Succeeded in getting to a model where NIH ICs are aggregating their data in one place and processing it in a consistent way (Big!)

What we have yet to do

- Uneven collection of phenotypic data
- Use of common and consistent data models (especially for phenotypes and metadata) -- new FHIR WG to help address
- Consistent processing of data across efforts -- without this, interoperability is challenging.
- Governance around models for data and metadata standards

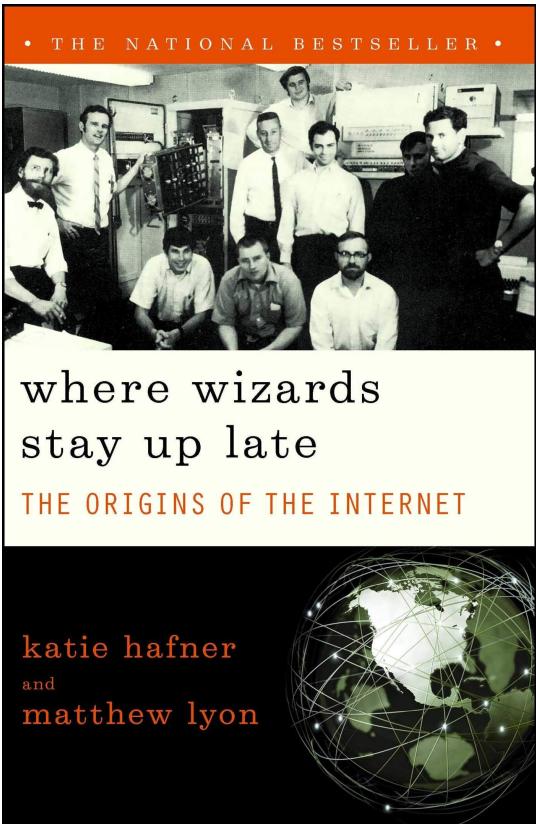
What we have accomplished

- Have basic standards in place
 - DRS for Data, TRS for Tools, WES for analytics.

What we have yet to do

- DRS, TRS, WES are still evolving and unevenly utilized
- While we have an increasingly clear vision for how to handle data use and researcher identities (DUO, Passports), it is still early days.
- We see an international agreement on converging on these standards in the genomics field, but still early.

Closing Thought



We should not underestimate the importance of this effort...

- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
 - Others have done it before (SWIFT, the internet, the web).
- If we fail, we will degenerate into a collection of monolithic data silos
 - Others have done this before too (medical records in US hospitals)...



Working Group Updates

Community / Governance WG

Bob Grossman

Outreach / Training

Mo Heydarian

FHIR WG

Robert Carroll / Allison Heath

RAS

Rebecca Rosen

Systems Interoperation

Brian O'Connor / Jack DiGiovanna



Working Group Update - Community / Governance

Chairs: Bob Grossman / Stan Ahalt

Progress Since October Workshop

- We developed a Principles **FAQ** to help create a common vocabulary and common POV
- Interop Principles Ver A (Jan 13, 2020)
- Interop Principles Ver B (Mar 20, 2020)
- Interop Principles Ver C (Apr 8, 2020)

Roadmap (3-6 months)

- Finalize Interop Principles and roll out
- Develop preliminary Interop “metrics” to evaluate how well a particular platform follows the Interop principles

Challenges

- Agreeing on a general definition of a “**trust relationship**” between two data platforms (**done**)
- Clearly separately **technical guidelines** (use DRS identifiers) from **operating principles** (provide users access to your data in the least restriction manner) (**done**)
- Developing appropriate **metrics** to evaluate adherence to the Interop Principles



Working Group Update - Outreach / Training

Chair: Ashok Krishnamurthy

Progress Since October Workshop

- NCPI knowledge base - what are the platforms? How are they similar/different?
- “Train your colleague” virtual training session
 - [Recordings and presentation materials](#) on NCPI Stacks and use cases

Roadmap (3-6 months)

- Host NCPI knowledge base on AnVIL Portal, include NCPI background, WG information, training materials (NCPI community contributions encouraged)
- Cloud cost resources - real examples of analysis on each stack

Challenges

- In-person training/collaborating - capture virtual sessions/materials to build training resources for asynchronous learning (lots of opportunities: MaGIC Jamboree, BCC2020, BioC2020, ISMB, ASHG)



FHIR Working Group Update - Background



Chairs: Robert Carroll / Allison Heath

Problems Facing the FHIR research community

- Good progress on EHR interoperability using FHIR but it is still a serious challenge
- Research world, including NCPI, has similar interoperability challenges
- FHIR offers a lot of promise, but effective tooling, standard model dev and processes not in place

Goals To help prepare us for solving the problem(s)

- Take a hands-on approach to learning and prototyping with FHIR through the WG kickoff project ("Project Forge") for current NCPI participants
- Learn how to effectively collaborate on FHIR modeling
- Gain a shared understanding of the problems FHIR solves and its current weaknesses

Objectives

- Take a practical approach to learning and prototyping with FHIR
- Provide feedback to the HL7 FHIR and NIH communities
- Create a roadmap for clinical data interoperability among datasets and platforms
- Engage the appropriate stakeholders for the longer term



FHIR Working Group Update - Roadmap



Chairs: Robert Carroll / Allison Heath

Progress Since October Workshop

- Group has officially been formed and met
- KFDRC team has built a preliminary FHIR modeling toolchain
- Organization of pilot to support interoperability

Roadmap ("[Project Forge](#)" - 3 months)

- KFDRC will organize initial infrastructure and dev process
- Groups will identify key datasets and familiarize themselves with those and FHIR modeling
- Collaborative and iterative process to create a baseline interoperable FHIR for research model
- Along the way - discuss naming conventions, model release process (i.e. versioning, distribution)
- Load 2 key datasets into the FHIR test server, review data with the FHIR Data Dashboard

Challenges

- Collaborative FHIR modeling with a fully remote, multi-organizational team
- Security framework and data sharing among IC-sponsored platforms
- Long term FHIR server licensing, deployment, and operation

FHIR Working Group - Potential User Story

Chairs: Robert Carroll / Allison Heath



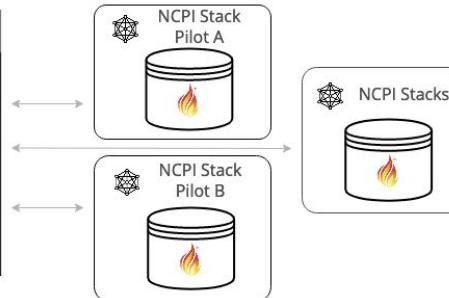
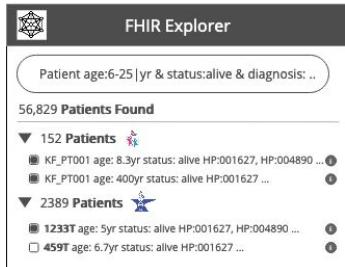
Down Syndrome (DS) Researcher

I want to **understand structural birth defect variability** among the DS population in order to **identify genetic modifiers** that impact quality of life from birth to late age



Create Patient Cohort:

Need: >= 2000 patients
Age: 6 to 25 years
Vital status: alive
Diagnosis: down syndrome, autism, ...
HPO: abnormal heart morphology | pulmonary hypertension | ...



Exchange Data:

Use FHIR Explorer to exchange data between:

- 1) FHIR-enabled workspaces (local or cloud-based)
- 2) FHIR servers in stacks



Explore Data:

Explore data and schemas in Jupyter/RStudio in the workspace of choice utilizing Python/R libraries





Working Group Update – RAS (NIH Researcher Auth Service)



Chairs: Rebecca Rosen, Susan Gregurick

Context for RAS and Progress so far

- **VISION:** Develop a unified, efficient, and secure authentication and authorization service that enables streamlined access by researchers to NIH-funded data resources across multiple systems; provides standardized methods of logging and auditing such access; and is compliant with NIST and GA4GH standards
- **PROGRESS:** V1.0 OIDC AuthN/Z endpoints deployed in testing environment for Phase 1 partners: KFDRC/BDCatalyst and CRDC/AnVIL; implemented basic auditing and logging of data and metrics; internal system monitoring and notifications [<https://auth.nih.gov/docs/RAS/serviceofferings.html>]

Roadmap (2020)

- Summer production deploy of SSO and AuthZ endpoints for Phase 1 interoperability use cases
- Use case / requirements gathering and planning workshop for Phase 2 integrations: NIMH Data Archive (NDA), dbGaP/NCBI, Common Fund Data Ecosystem, and All of Us
- End of year production deploy to support Phase 2 integration use cases and account linking

Challenges

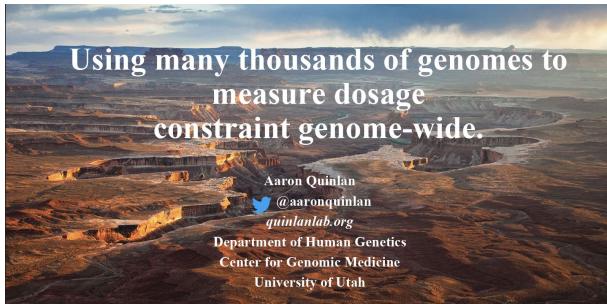
- Aligning development timelines with partner systems who have overlapping projects and priorities
- Facilitating long-running analytic pipelines



Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna

Our Working Group was created as an outcome of the NIH Workshop on
Cloud-Based Platforms Interoperability (RENCI Oct 3-4, 2019)



Sex as a Biological Variable: Use Cases and Challenges



Melissa A. Wilson
@sexchrlab
sexchrlab.org
Arizona State University

CAUSES AND IMPLICATIONS OF LINE-1 EXPRESSION VARIATION IN HEALTHY SOMATIC TISSUES

WILSON MCKERROW (POSTDOC)

FENYO LAB, INSTITUTE FOR SYSTEMS GENETICS, NYU LANGONE SCHOOL OF MEDICINE

DataSTAGE Science Use Case: Nutritional Omics of Lung Function Decline R01HL149352, 2019–2023



Dana B. Hancock, PhD
Director, Center for Omics Discovery and Epidemiology



Patricia A. Cassano, PhD
Director, Division of Nutritional Sciences
With acknowledgement to Bonnie Patchen



Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna



The group's Charter establishes the group's mission, members/teams, high-level scientific and technical goals, and timeline.

The group will spearhead **technical improvements** to cloud "stacks" created by the Common Fund, NCI , NHGRI, and NHLBI that enable improved interoperability. We will demonstrate progress in **realistic researcher use cases** every 6 months.

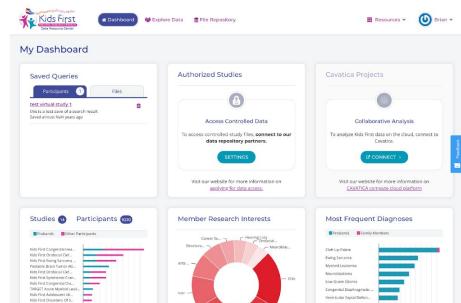
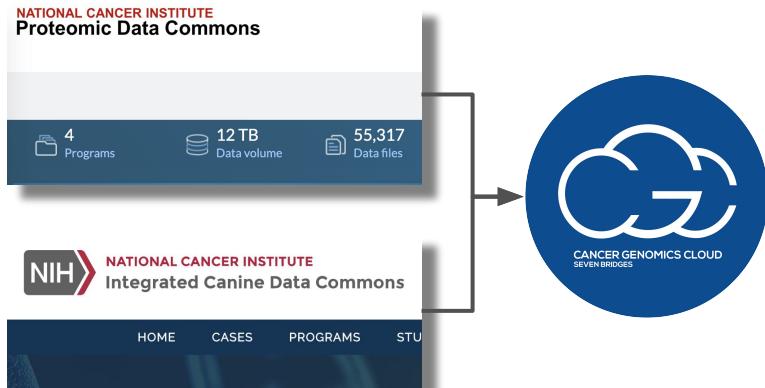
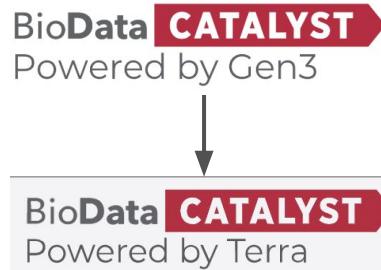
Please join if you are interested.



Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna

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



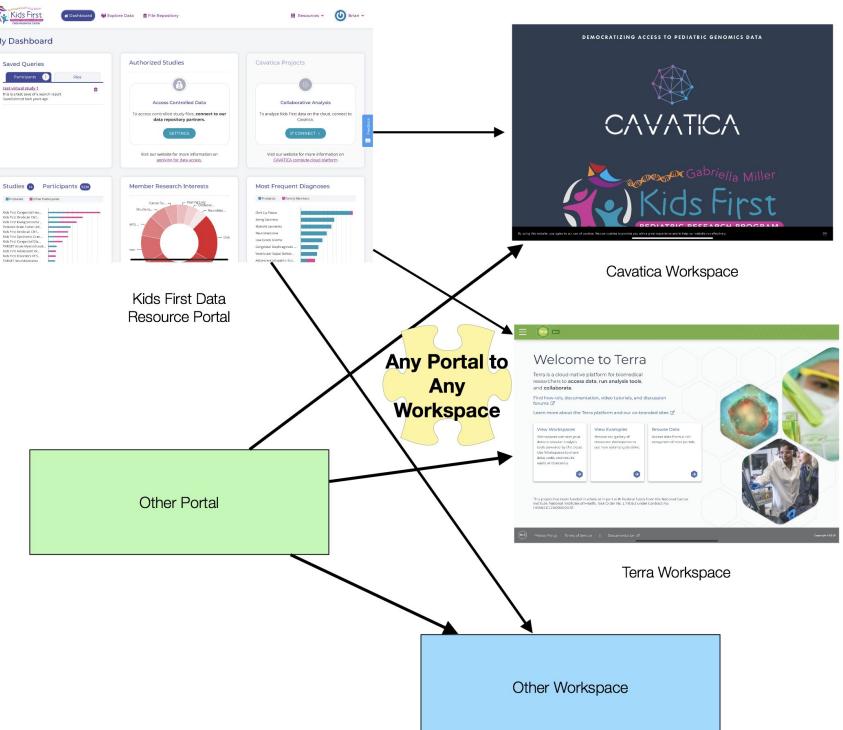
Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna

Connecting systems (*intra- & inter-IC*) would unblock key user stories

Goals (first 6 months):

- Analyze & develop interop between 4 distinct *Data Portals* and 3 Workspace environments (Terra, SB, and Gen3)
- Focus on achievable improvements to interoperation between these multiple groups. (Not analyzing *all* APIs initially)



See the Technical Plan



Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna

Immediately looked for scientific “driver projects”

Our WG quickly identified five interesting researcher use cases that required interoperability both within and between ICs

- CRDC + AnVIL (n=1);
- BioData Catalyst + Kids First (n=3)
- BioData Catalyst + Kids First + AnVIL (n=1)

[Charter](#) has details on the use cases



Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna



Initial technical effort focused on **lightweight mechanism** by which **Data Portals** can "hand off" search results to compute environments.

Allows researchers to leverage data on Kids First, AnVIL, BDCCat, and CRDC and compute in Terra, SB, and Gen3.

PFB and DRS were the initial candidates to send data references and metadata to compute environments

Working Group Update - Systems Interoperation

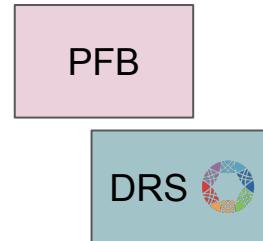
Chairs: Brian O'Connor / Jack DiGiovanna

Progress



1

Wrote a charter describing what we want to do based on 8+ use cases. IC signoff 1/17/2020



2

Wrote a technical plan, including dev guides for PFB and DRS. IC signoff 1/17/2020



3

Groups have been implementing & providing feedback to GA4GH, U. Chicago, and each other



Working Group Update - Systems Interoperation



Chairs: Brian O'Connor / Jack DiGiovanna

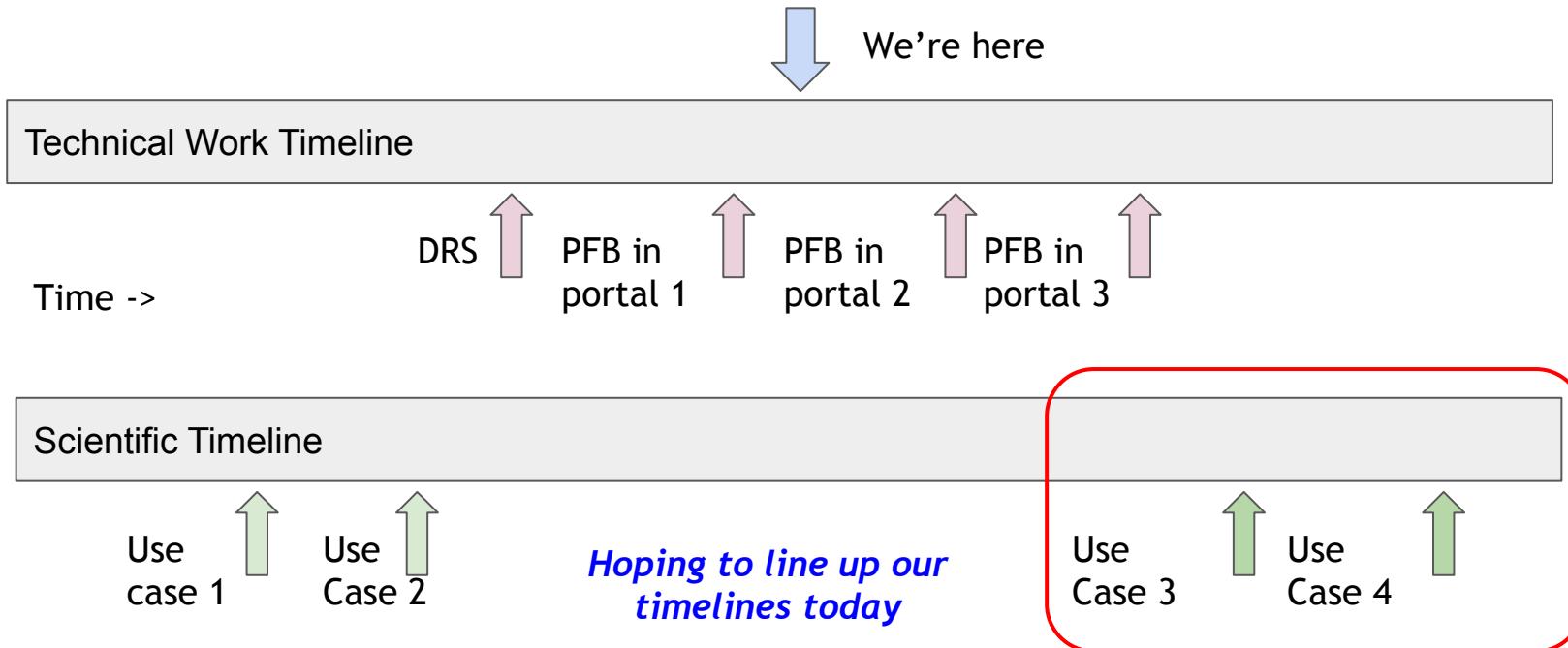
Challenges (science never sleeps)

- Agreement on the overall goals happened easily and groups understood a clear path to make improvements.
 - However, initial steps were difficult to coordinate
- *Standards needed updates* which introduced delays
 - Wanted to align persistent GUIDs/compact identifiers with GA4GH DRS
 - Different versions of DRS supported
 - PFB was not widely implemented, analyzed and feedback developed
- Aligning group timelines proved challenging
 - Who's funded for this work and when is it in their work plan?
 - To work on interop we need groups to align their timelines
- Aligning use cases to users stories would also help tremendously

Working Group Update - Systems Interoperation

Chairs: Brian O'Connor / Jack DiGiovanna

Looking for the Goldilocks use cases happening (or repeating) at just the right time





Use Case Updates

KF DRC / NHLBI BDC / NCI CRDC

Allison Heath

NCI CRDC / NHGRI AnVIL

Wilson McKerrow & Jack DiGiovanna

NHLBI BDC / KF DRC

Gina Kuffel & Garrett Rupp

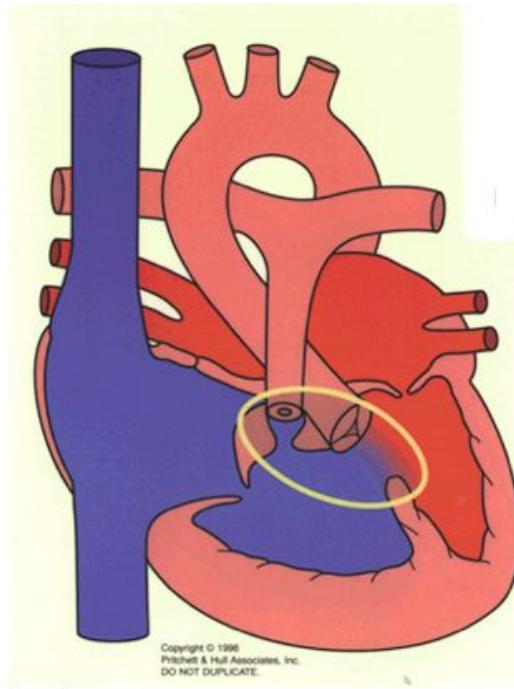
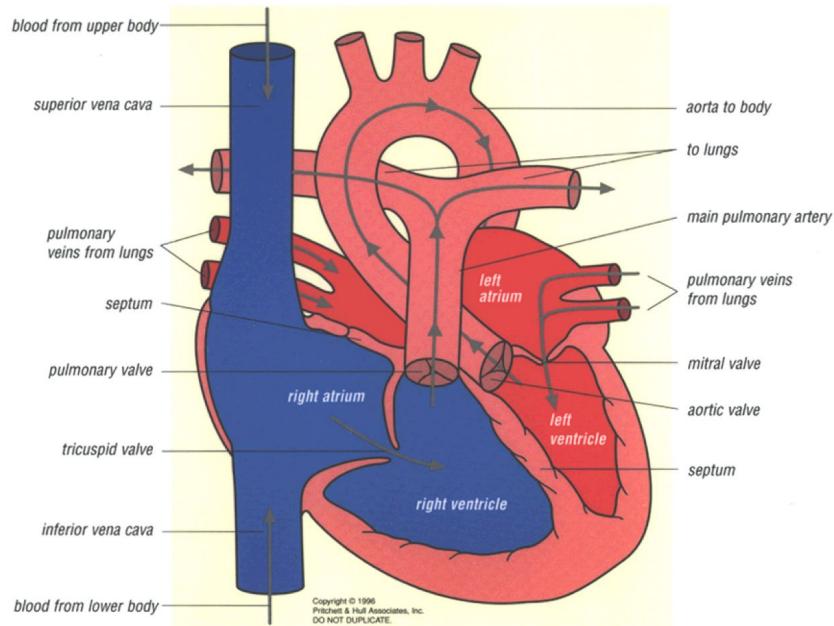
NHGRI AnVIL / KF DRC / NHLBI BDC

Alisa Manning

Use Case Updates: Congenital Heart Defect Analysis

Presenter: Allison Heath (CHOP)

Researchers: Betsy Goldmuntz, Deanne Taylor, et al.



Pritchett & Hull Associates, Inc. Atlanta, GA 800-241-4925 Providing Quality Health Education Materials Since 1973



Pritchett & Hull Associates, Inc. Atlanta, GA 800-241-4925 Providing Qu



Use Case Updates: Congenital Heart Defect Analysis

Presenter: Allison Heath (CHOP)

Researchers: Betsy Goldmuntz, Deanne Taylor, et al.

Background

- CHD is the most common birth defect, occurring in 1% of all live births and 10% of stillbirths
- Pediatric Cardiac Genomics Consortium began enrolling in 2010, over 29,000 participants, 3,300+ trios characterized with WES
- Analysis with WES using SFARI (Autism) data as controls, characterized about 40% of cases - 60% still remain unexplained

Goals

- Fill gaps that exist in understanding the etiology of CHDs
- Better understand cardiogenesis and assess risk of disease

Objectives

- Harmonize WES data to GRCh38
- Utilize Aortic Arch Anomalies (AAA) data generated at CHOP as controls
- Utilize curated diagnostic groups to analyze severe vs. mild CHD diagnoses
- Share data for analysis with collaborators at UTHSC and UPenn



Use Case Updates: Congenital Heart Defect Analysis

Presenter: Allison Heath (CHOP)

Researchers: Betsy Goldmuntz, Deanne Taylor, et al.



Progress

- Completed all alignment tasks on BioData Catalyst powered by Seven Bridges
 - Brought workflow from Cavatica to BDC
 - 12,506 PCGC exome samples
 - 161 AAA samples
- Output into S3 buckets managed by KFDRC
- Cohort joint genotyping on Cavatica
 - Split gVCF files in 89 regions (~2 million files)

Next Steps

- Joint genotyping on each region
 - Merge/gather VCFs and VQSR to final VCFs
- Shared project for collaborators
- Bring together phenotypic and genotypic data for analyses

Use Case Updates: Congenital Heart Defect Analysis

Presenter: Allison Heath (CHOP)

Researchers: Betsy Goldmuntz, Deanne Taylor, et al.

Alignment Statistics on BDC

	Min	Max	Mean	Median	Sum
Run hours	0.06555556	42.29639	2.22909	1.64389	27386.61
Cost	0.05	44.66	1.686	1.39	20714.38

Status	Total task	Total Cost	Total run hours
COMPLETED	12217	19982.94	25701.08
ABORTED	59	723.64	1669.101
FAILED	10	7.8	16.43

	Min	Max	Mean	Median	Sum
Run hours	0.5456	13.4739	2.1037	1.6403	25701.08
Cost	0.52	16.13	1.636	1.390	19982.94



Use Case Updates: Congenital Heart Defect Analysis

Presenter: Allison Heath (CHOP)

Researchers: Betsy Goldmuntz, Deanne Taylor, et al.



Preliminary gVCF Split Statistics in CAVATICA

Status	Total task	Total Cost	Total run hours
COMPLETED	12217	1456.11	3174.006
ABORTED	35	146	132.8789
FAILED	23	11.82	26.33083

Use Case Updates: Interop Driven by Common Need

Presenter: Allison Heath (CHOP)

What PCCG researchers want to do is not uncommon!

Similar cancer use cases help drive initial KFDRC/NCI CRDC integration with TARGET NBL and TARGET AML



Volume 105, Issue 3, 5 September 2019, Pages 658-668



Report

Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma

Laura E. Egolf^{1, 2, 3}, Zalman Vaksman^{2, 3, 4}, Gonzalo Lopez^{2, 3, 4}, Jo Lynne Rokita^{2, 3, 4}, Apexa Modi^{2, 3, 5}, Patricia V. Basta^{6, 7}, Hakon Hakonarson^{8, 9}, Andrew F. Olshan^{6, 7}, Sharon J. Diskin^{1, 2, 3, 4, 5, 10}

Show more

<https://doi.org/10.1016/j.ajhg.2019.07.020>

Under an Elsevier user license

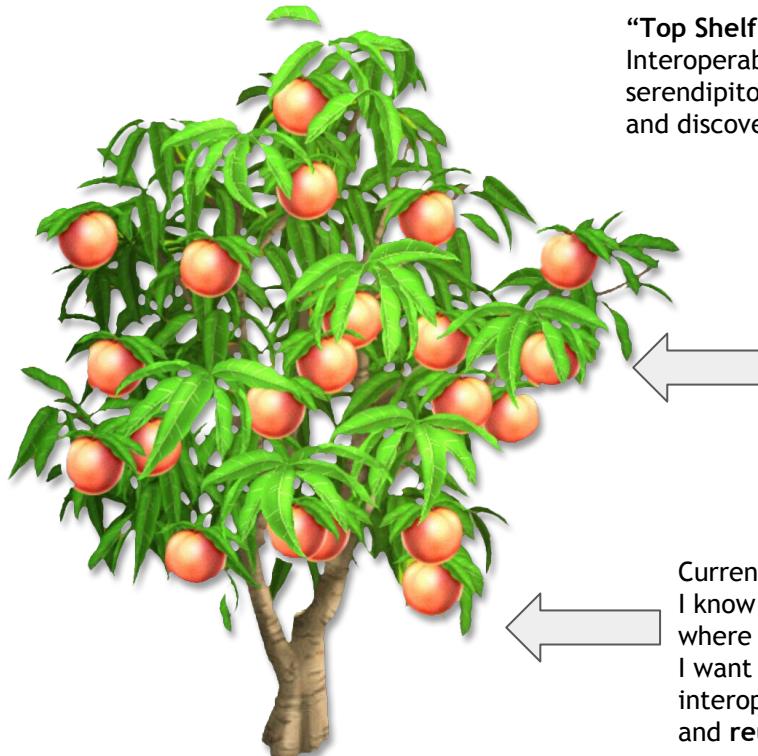
Get rights and content

[open archive](#)

- 5,585 Neuroblastoma cases SNP arrays
- 23,505 cancer-free controls SNP arrays
- 7 local cases WGS
- 5 TARGET NBL and Kids First NBL WGS
- 3 Kids First NBL WGS trios

Use Cases to User Stories (as Informed by Pediatrics)

Presenter: Allison Heath (CHOP)



“Top Shelf Fruit”:

Interoperable resources allow me to make serendipitous or unexpected connections and discoveries.



Change in the current scientific process. Cannot be a goal, but could be an emergent property.

Only happens with interoperability.

Goal: Accelerating the Current Scientific Process

A Step Up:

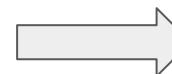
I know what analysis/data I want to do, but want to easily **find** other datasets that can augment it. E.g. case/controls, syndromic disorders, rare diseases

Without interoperability: Going to conferences, reading papers, talking to people etc.



Current use cases are “low hanging fruit”:

I know what datasets I want to analyze, where they’re located and what workspace I want to use. I need platforms to interoperate so data and resource access and **reuse** is easy in minimal time.



A Step Up:

I have used (or have created) tools/visualizations and I want to explore all data I have access to in them.

Without interoperability: pedcBioPortal

Much of our focus because we **have** to be able to have access and it needs to be **easier and faster**.

Without interoperability:
Local download and analysis

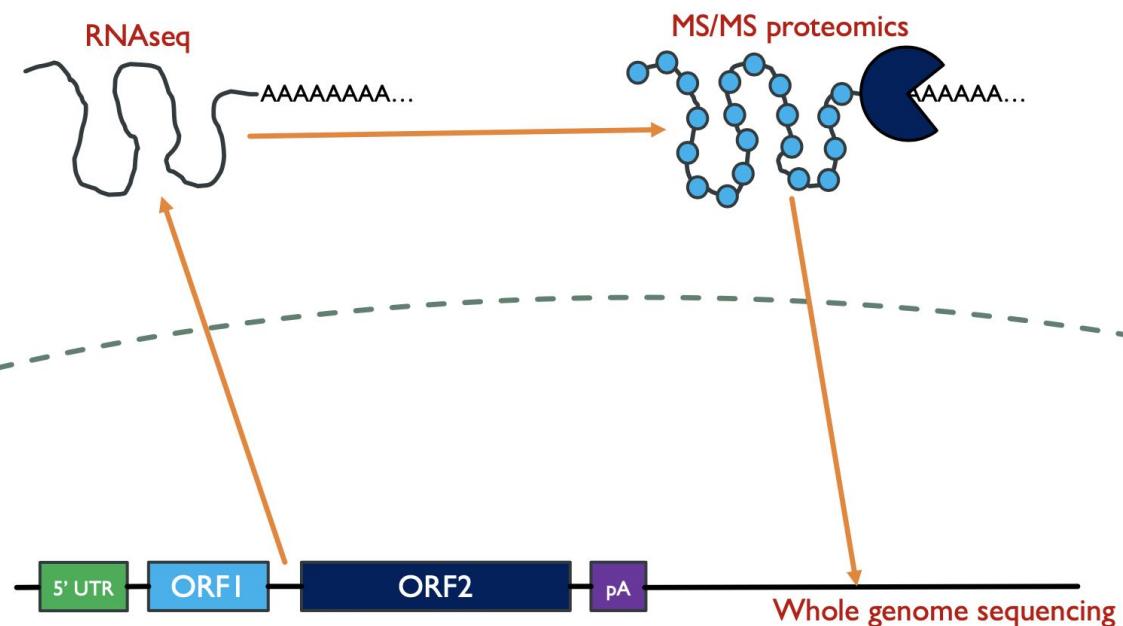
Just a few examples to potentially frame how to prioritize activities!

Use Case Update: NCI CRDC + NHGRI AnVIL

Presenters: J DiGiovanna & Wilson McKerrow

Researchers: W McKerrow, D Fenyö, et al.

SCIENTIFIC QUESTION: Causes & Implications of LINE-1 expression variation in healthy somatic tissues



Data

- CPTAC (Gen3) **[CRDC]**
- TCGA (Gen3) **[CRDC]**
- GTEx (Terra) **[AnVIL]**

Portals

- Proteomics Data Commons **[CRDC]**
- Data Browser **[CRDC]**
- GTEx Workspace **[AnVIL]**

Platforms

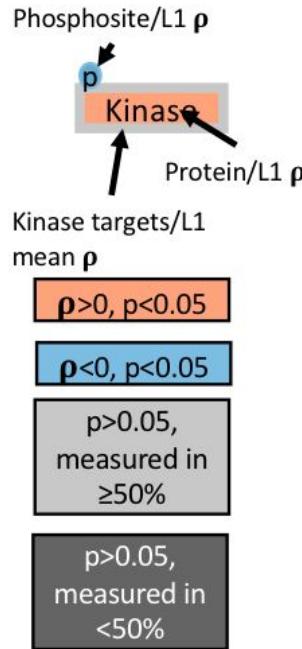
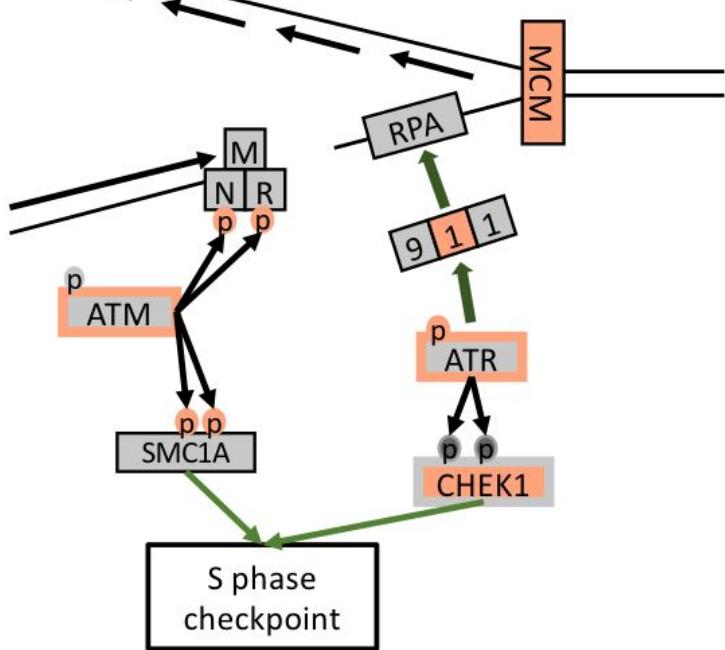
- Cancer Genomics Cloud **[CRDC]**
- Terra **[AnVIL]**

Use Case Update: NCI CRDC + NHGRI AnVIL

Presenter: Wilson McKerrow

Researchers: Wilson McKerrow, David Fenyö, et al.

CANCER PROJECT: LINE-1 expression correlated with DNA damage in S phase



Data

- Genomics [CRDC]
- Transcriptomics [CRDC]
- Proteomics [CRDC]
- Phosphoproteomics [CRDC]

Platforms

- Cancer Genomics Cloud [CRDC]

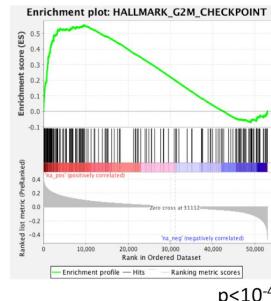
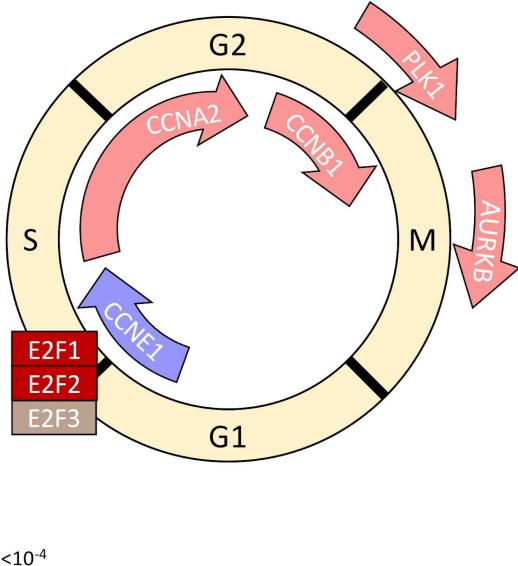
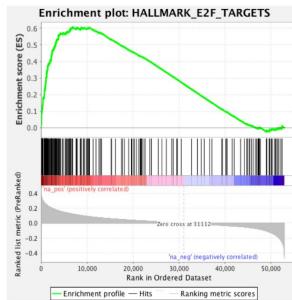
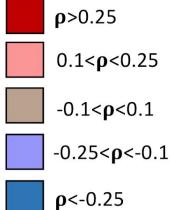
Use Case Update: NCI CRDC + NHGRI AnVIL

Presenter: Wilson McKerrow

Researchers: Wilson McKerrow, David Fenyö, et al.

NORMAL SKIN: LINE-1 expression correlated with changes in cell cycle regulation

Sun Exposed Skin
Partial correlation (accounting for RIN)



Data

- GTEx RNA-seq [AnVIL]

Platforms

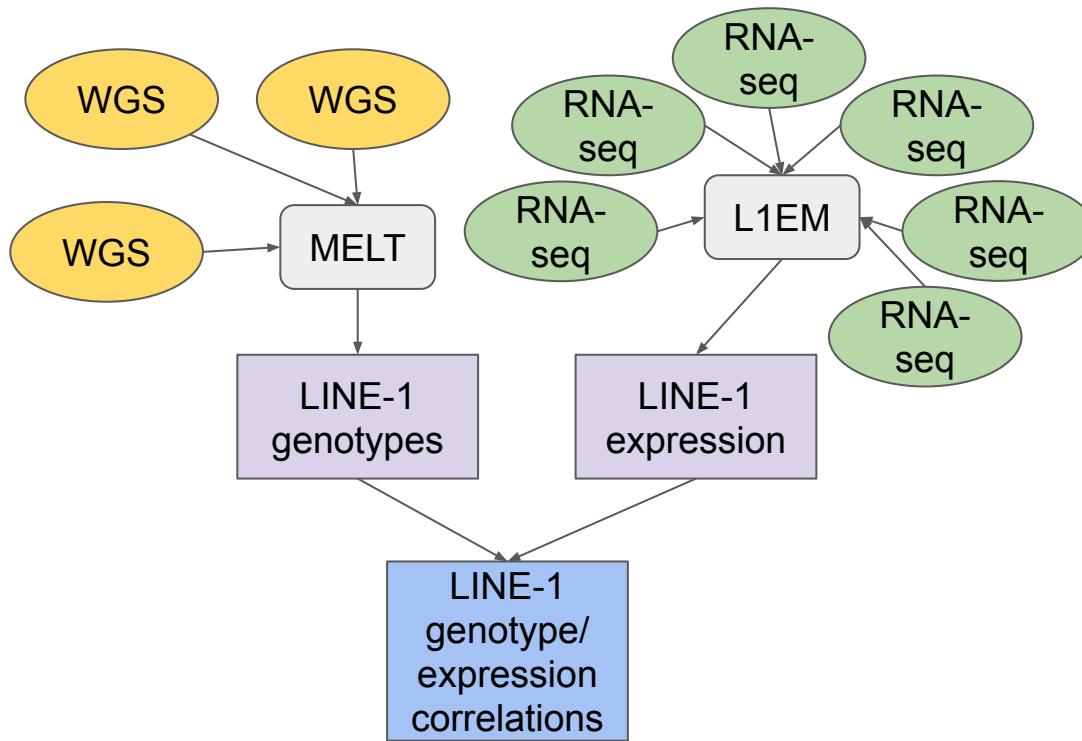
- Cancer Genomics Cloud [CRDC]
- Terra [AnVIL]

Use Case Update: NCI CRDC + NHGRI AnVIL

Presenter: Wilson McKerrow

Researchers: Wilson McKerrow, David Fenyö, et al.

NEXT STEP: Do polymorphic LINE-1 explain variation in LINE-1 RNA?



Data

- GTEx RNA-seq [\[AnVIL\]](#)
- GTEx WGS [\[AnVIL\]](#)
- TCGA RNA-seq [\[CRDC\]](#)
- TCGA WGS [\[CRDC\]](#)

Platforms

- Cancer Genomics Cloud [\[CRDC\]](#)
- Terra [\[AnVIL\]](#)

Use Case Update: NCI CRDC + NHGRI AnVIL

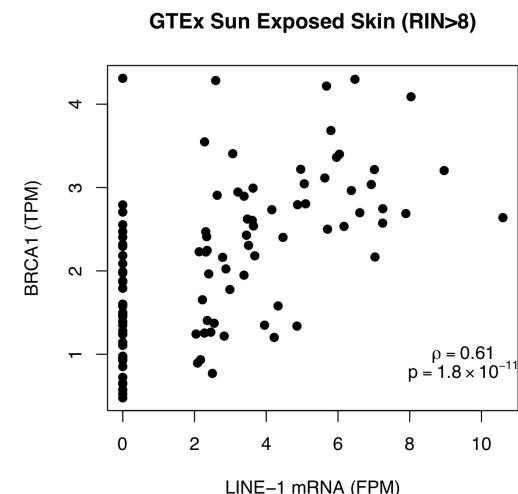
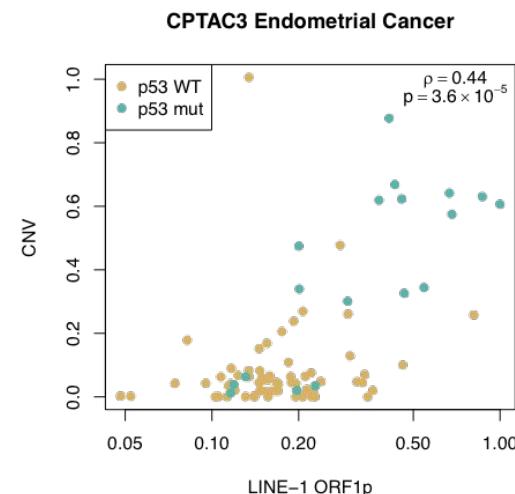
Presenter: Wilson McKerrow

Researchers: Wilson McKerrow, David Fenyö, et al.

PRELIMINARY RESULTS: High LINE-1

expression correlated with persistent DNA damage in tumors; Detectable LINE-1 expression correlated with checkpoint expression in some healthy tissues.

- McKerrow, Fenyö. "Cloud based tools for multiomic LINE-1 quantification". (In prep.)
- McKerrow, et al., Fenyö. "LINE-1 expression in cancer correlates with DNA damage response, copy number variation, and cell cycle progression". Submitting to *Cancer Cell* this month.
- McKerrow et al., Fenyö. "LINE-1 expression in healthy tissues...". (Computation ongoing)





Use Case Update: NCI CRDC + NHGRI AnVIL

Presenter: Jack DiGiovanna

Researchers: Wilson McKerrow, David Fenyo, et al.



ANALYSIS COST per GTEx RNA-seq \$0.25

- Storage = \$0.10
- Compute = \$0.15
- Egress = \$0

TOTAL RNA-seq COST (estimate)

- $\$0.25 \times 17382 = \4346

APPROXIMATE TIME (calendar)

- Getting data access: 30% (Dec-Jan)
- Wiring Platforms and Portals together: 40% (Dec-Feb)
- Pipeline/workflow running: 30% (March-April)



Use Case Update: NCI CRDC + NHGRI AnVIL

Presenter: Jack DiGiovanna

Researchers: Wilson McKerrow, David Fenyo, et al.

HOW WE CONNECTED:

[CRDC - CRDC] Portals and Platforms used existing manifest standards within CRDC coupled with fence IDs and metadata from indexd. *Reusable approach already leveraged in CRDC for other DCCs*

[AnVIL - CRDC] CRDC Platform used similar manifests, but the flow to find data on Terra (hosted by Broad) and move it to Seven Bridges was complex due to lack of common standards and security concerns around AuthZ.

GTEX is in an AnVIL bucket, was not connected through Fence, and users could not get signed URLs from Terra, so the following flow was established for the existing CGC user:

1. Create a Google account; create a Terra account; link dbGaP credentials
2. AuthZ to access GTEX via Terra
3. Copy links to GTEX cohort into Terra Workspace
4. Get signed URLs to pass to Project (GCP) on SB CGC
 - a. Could call Terra API to generate signed urls, but requires passing tokens
 - b. Instead download a manifest from Terra Workspace to local machine
 - c. Run a local script with system variables to more securely pass credentials, returns signed URLs
5. Use SB API to pass signed URLs to a task input; that task will produce the files (now copied) in an SB CGC Project
6. Upload a metadata manifest; set GTEX metadata appropriately
7. As Project owner/admin, you are responsible for the data inside the project (policy-level control)

Use Case Update: NCI CRDC + NHGRI AnVIL

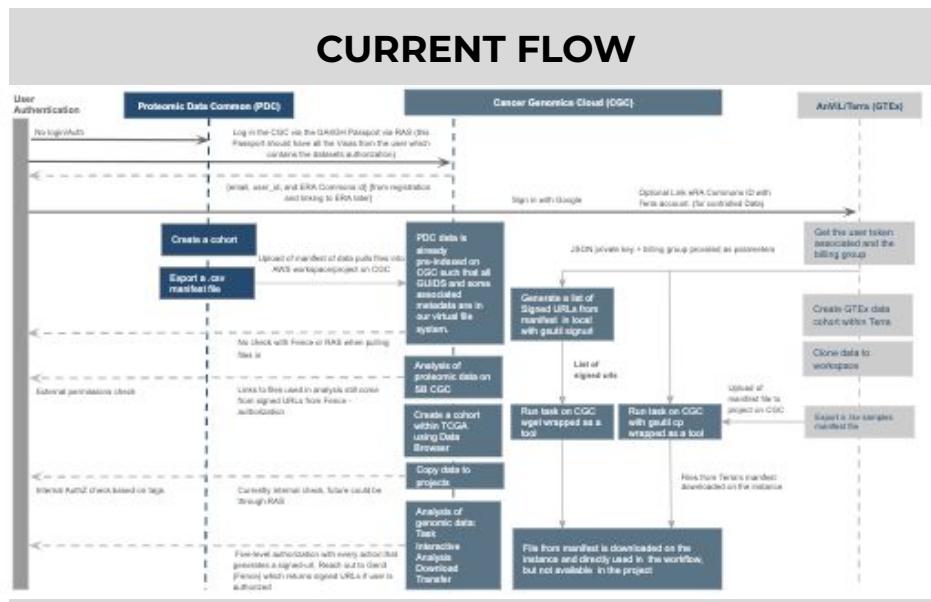
Presenter: Jack DiGiovanna

Researchers: Wilson McKerrow, David Fenyo, et al.

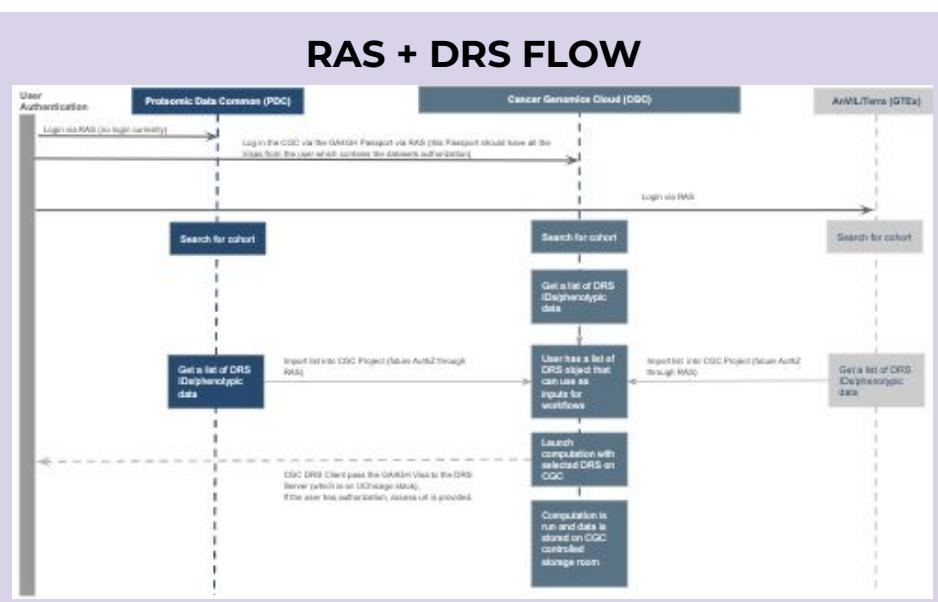
GAPS:

- DRS servers on some CRDC *Platforms*, but not *Portals*. DOS server on AnVIL Platform. DRS clients?
- Current version of PFB is not supported in CRDC Portals or Platforms; PFB not yet available in AnVIL.
- Policy-based auth and inefficient interop UX. *DRS + RAS handling AuthZ would be a major step forward*.

CURRENT FLOW



RAS + DRS FLOW





Use Case Update: NCI CRDC + NHGRI AnVIL



Presenter: Jack DiGiovanna

Researchers: Wilson McKerrow, David Fenyo, et al.

STANDARDS FUTURE WORK:

- DRS servers would have unified the code for accessing raw data and saved multiple steps (assuming a search-API or Portal to form the cohorts of ids).
- DRS client could pull over the files “just-in-time” rather than making a copy.
- An Avro-based standard manifest could improve passing cohorts within and between ICs. It could be especially useful for metadata, but utility over current manifests is unclear for GUIDs
- RAS for AuthN with Fence for AuthZ would enable technical-level data controls on all platforms

SCIENCE FUTURE WORK:

- Connect GTEx back to TCGA to validate cancer connection with healthy tissue
- LINE-1 doesn't seem active in blood and brain cancers in adults but perhaps functions differently in pediatric patients. Apply for access to Kids First datasets.
- Leverage LINE-1 work in healthy tissue to study its role in contexts other than cancer- exercise, aging, etc

Use Case Update: NHLBI BioData Catalyst + Kids First DRC

Presenter: Gina Kuffel

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez

SCIENTIFIC QUESTION: Explore the patterns of variation in genes associated with relevant inherited cardiac conditions

Data

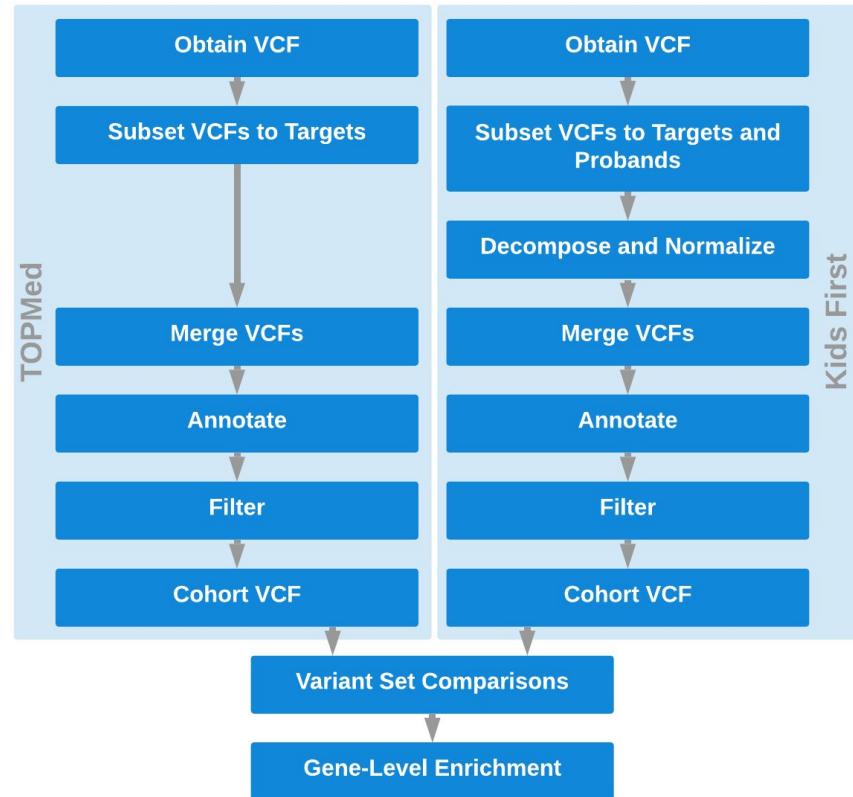
- TOPMed **[NHLBI BD**Cat**]** - 1021 germline VCFs with atrial fibrillation
- PCGC **[NHLBI Kids First]** - 2,148 gVCFs across 716 families

Portals

- BioData Catalyst Commons
- Kids First Data Portal
- Cavatica **[KFDRC]**

Platforms

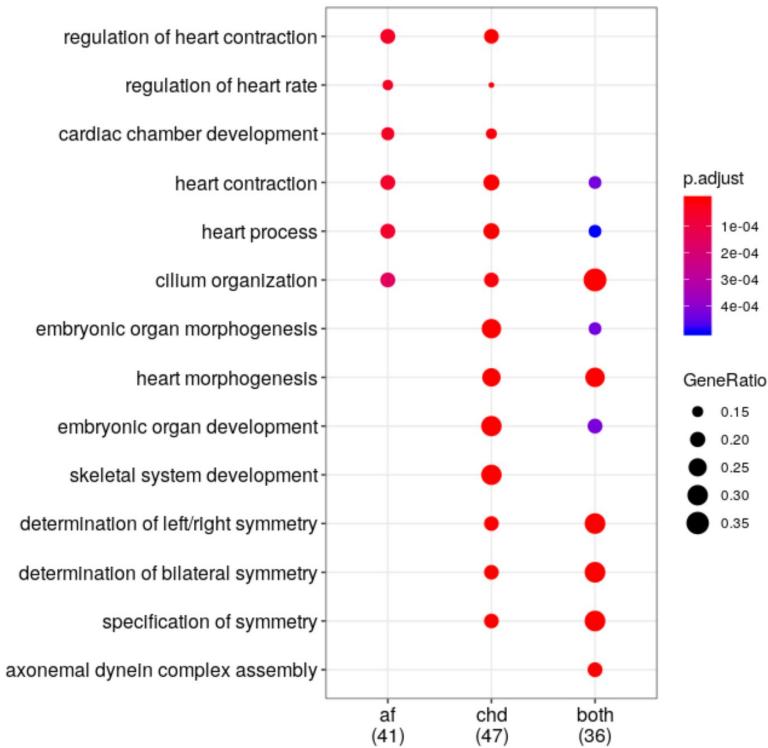
- Cavatica **[KFDRC]**



Use Case Update: NHLBI BioData Catalyst + Kids First DRC

Presenter: Gina Kuffel

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez



PRELIMINARY RESULTS: A gene enrichment analysis revealed clear and interesting patterns. The presented approach seems viable to generate feasible candidate genes. Further investigation could inform a comprehensive panel for AF development.

APPROXIMATE TIME (Period of several months)

- Getting data access: IRB for TOPMed dataset, dbGaP access
- Wiring Platforms and Portals together: Initial approach was manual
- Pipeline/workflow running: many iterations over several months



Use Case Update: NHLBI BioData Catalyst + Kids First DRC



Presenter: Gina Kuffel

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez

HOW WE CONNECTED:

[BioData Catalyst] Gen3 Exploration into a Workspace and then into a VM

[KFDRC] Used Cavatica API to get files into a VM for the data processing

No way to get data into a single Gen3 workspace from different data commons

1. Create a Google account; create a Cavatica account; link dbGaP credentials
2. AuthZ to access data via Cavatica, and Gen3
3. Select virtual cohort using Gen3 Data Exploration page in BDCatalyst and export to Gen3 Workspace
4. Run pipeline in Jupyter notebook and store result
5. Select cohort in KFDRC and download manifest
6. Spin up a virtual machine
7. Get signed URLs to load data into vm
 - a. Provide KFDRC manifest to Cavatica API to download the files which returns Gen3 pre-signed URLs
8. Extract proband metadata
9. Run pipeline in VM and store result
10. Merge results in VM



Use Case Update: NHLBI BioData Catalyst + Kids First DRC



Presenter: Gina Kuffel

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez

HOW WE CONNECTED:

[BioData Catalyst] Used Gen3 API to get files into a VM for data processing

[KFDRC] Used Cavatica API to get files into a VM for the data processing

Enable ability to mount data files from multiple Data Commons to the same Gen3 Workspace.

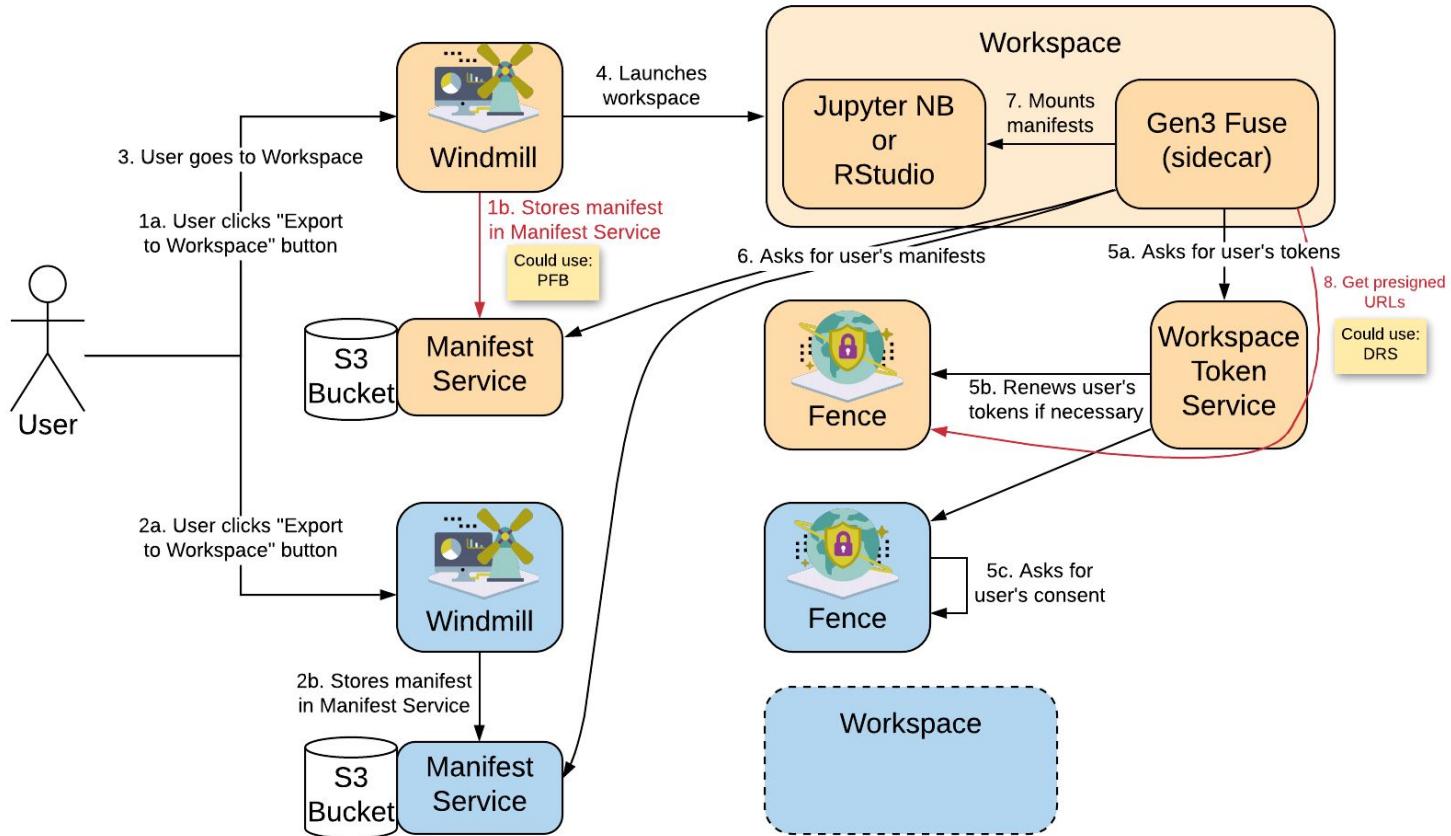
1. Update the Gen3 Fuse sidecar to enable mounting files from multiple commons
2. Update the Workspace Token Service to automate the retrieval of user credentials and tokens
3. Update the Manifest Service to accept manifests from different commons
4. Create a Gen3 Fuse compatible manifest of the files from Cavatica (Kids First GUIDs)

Use Case Update: NHLBI BioData Catalyst + Kids First DRC

Presenter: Gina Kuffel

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez

Multi-Commons Workspace





Use Case Update: NHLBI BioData Catalyst + Kids First DRC



Presenter: Garrett Rupp

Researchers: Robert Grossman, Dmitry Grigoryev, Kyle Hernandez

GAPS:

- Mechanism to transfer phenotype data from non-Gen3 portals (KF DRC) to Gen3 Workspaces undefined
- Unified Gen3 Workspace environment across the NIH ecosystem unavailable

STANDARDS FUTURE WORK:

- Standard manifest would improve passing cohorts within and between ICs, including Gen3 instances of each IC, especially for transfer of metadata

SCIENCE FUTURE WORK:

- Identify de novo mutations in the Kids First dataset
- Trace accumulations of these de novo mutations in the TopMed dataset

INTEROP FUTURE WORK:

- Implementation of PFB hand-off mechanism to import and export data from a Gen3 Workspace
- *Potential Implementation of unified Gen3 Workspace environment across the NIH ecosystem*
- DRS implementation based on GA4GH emerging specs
- Ongoing RAS integration

Use Case Update:

NHGRI AnVIL + Kids First DRC + NHLBI BioData Catalyst

Presenter: Alisa Manning, PhD Researchers: Alisa Manning, Brian O'Connor, Timothy Majarian

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

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

Data

- PCGC (Gen3) **[CFDR]**
- PCGC_CHD (Gen3) **[BDC]**
- FHS (Gen3) **[BDC]**
- JHS (Gen3) **[BDC]**
- GTEx (Terra) **[AnVIL]**

Platforms

- Cavatica **[CFDR]**
- Terra **[BDC]**
- Terra **[AnVIL]**

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

Presenter: Alisa Manning, PhD **Researchers:** Alisa Manning, Brian O'Conner, Timothy Majarian

Write analysis plan / pipeline to address **SCIENTIFIC QUESTION:** Investigate genetic factors related to congenital heart defects in a study design that uses healthy controls from two NHLBI cohorts. Perform pooled analysis on AnVIL powered by Terra.

Analysis Steps / Milestones	Timeline	Gaps
Coordinate with data providers and the NIH Systems Interoperation Working Group to test needed datasets are accessible via the GA4GH DRS 1.1 standard and metadata is accessible in the PFB format.	April 1 – June 30, 2020	
Data Sciences Platform at Broad to test auth account linking with Kids First, AnVIL, and Bio Data Catalyst.	July 1 - September 30, 2020	Auth account linking with KFDR, AnVIL, and BDC.
Data Sciences Platform at Broad to test data accessibility through DRS 1.1 , ensuring the Terra platform powering AnVIL and Bio Data Catalyst can access necessary datasets.	July 1 - September 30, 2020	Implementation of DRS 1.1
Implement analysis plan on the AnVIL/BD Cat workspace environment powered by Terra. Note: Implementing the analysis plan is work that is funded by a separate funding stream, and is noted in this proposal as an external dependency that must be completed before the final deliverable of this project can be completed.	September 30, 2020 – June 30, 2021	
Write tutorials/resources Write report documenting problems and proposed solutions for platform developers.	October 15, 2020 – June 30, 2021	



General Interoperability Solution

Use Case Gaps

Technical Gaps

- Next Steps
- Timeline

Goals for 2020

Bob Grossman / Garrett Rupp

Brian O'Connor

Valentina DiFrancesco



General Interoperability Solution - Use Case Gaps

Presenter: Garrett Rupp

Our Use Cases keep the science community at the heart of our work, but...

...our current solutions are use-case specific, and not generalizable for the research community.

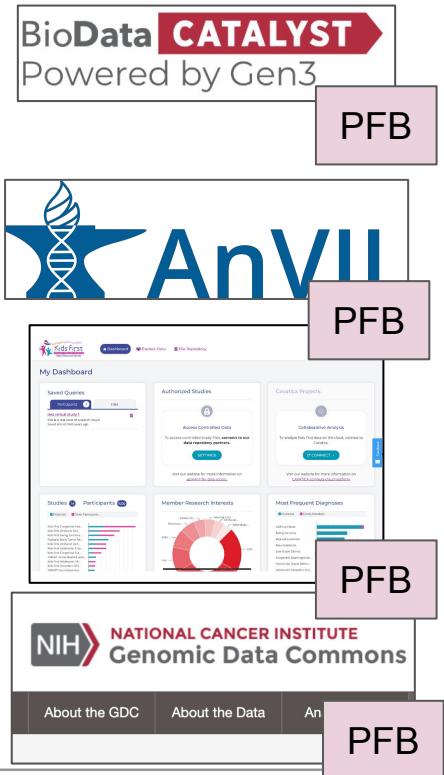
To drive solutions valuable to all users, we must:

- **Standardize handoff** mechanisms across systems
- Continue to **Generalize handoff** across systems. For NCPI, there could be many permutations of a Portal to Workspace handoff ([enjoy a fun diagram here](#))
- **Define security requirements** and agreements across systems, not projects, within NIH
- **Generate User Stories/Journeys**, in addition to use cases
- **Prioritize** activities and use cases across all teams, so resources and timelines are aligned
- **Fund activities** in alignment with their prioritization
- **Reprioritize** funded work within each IC

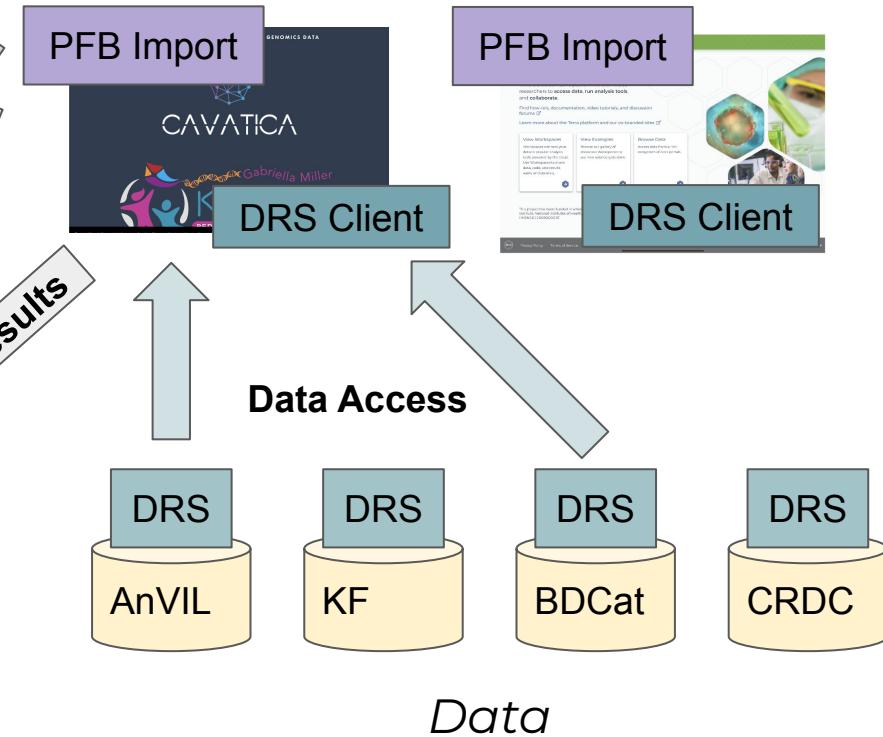
General Interoperability Solution - Initial Tech Goal

Presenter: Brian O'Connor

Portals



Workspaces





General Interoperability Solution - Current Tech Gaps

Presenter: Brian O'Connor

Current Technical Gaps:

Given the information we described in the Systems Interoperation update (and illustrated in the use case updates), what are the **technical gaps** between how things are done today and where we want to be soon?

- **Search Result Handoff: PFB**, we need to have 4 portals try and prototype PFB as the search result/DRS URI handoff mechanism so we can iterate and improve
- **Data Access: DRS 1.1** is just landing, we need DRS implemented across 4 Gen3 instances and DRS 1.1 in our workspaces (Terra, SBG, Gen3 notebooks)
- **Auth**: We need to understand how we use **Gen3 tokens now** and how we **migrate to RAS**
- **Work Plans**: We need to ensure this **key work is in groups' work plans**
- **Technical Timelines**: We need to **align multiple groups' timelines** so we all implement DRS/PFB at a point that will help our use cases.

We need agreement on a collective timeline for a technical “safe harbor”

General Interoperability Next Steps - PFB

Presenter: Brian O'Connor

PFB: A common mechanism to hand off search results from
Kids First, AnVIL, BDCat, and CRDC

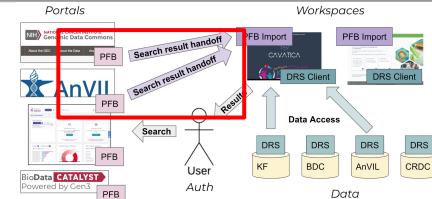
Portals that support passing search results as PFB:

	AnVIL Portals	BDC Portals	CRDC Portals	GMKF Portals
PFB output	Gen3 to Terra supported upon PROD deployment	Gen3 to Terra supported Gen3 to SB in progress	ISB, SB, Terra: Not supported PDC, ICDC, CTDC: Not supported	KF Data Portal: Prototyped PFB with PCGC. Roadmap is focused on FHIR for Portal/CAVATICA.

Workspace environments that can receive PFB-based search results:

	Gen3 (BDCat, AnVIL, CRDC)	Seven Bridges (BDCat, CRDC, KF)	Terra (BDCat, AnVIL, CRDC)
PFB input	PFB workspace ingestion in development	BDC: Working with Gen3 on windmill integration (~Q3) CRDC & KF can inherit functionality.	UI redirect flow for receiving a PFB file and creating a workspace. This is used in BD Cat in production

Opportunity to improve PFB based on detailed analysis during prior 6 m; minimize custom work for Portals
Risks PFB not a GA4GH standard (yet) which may slow adoption; PFB tooling currently focused on Gen3



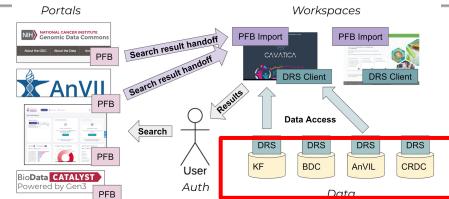
General Interoperability Next Steps - DRS

Presenter: Brian O'Connor

DRS Service:

A common mechanism to access data across Kids First, AnVIL, BDCat, and CRDC

U. Chicago is running instances for NIH interoperability; others are also implementing DRS (Terra, SB)



	AnVIL Gen3 Instance	BDC Gen3 Instance	CRDC Gen3 Instance	GMKF Gen3 Instance	Gen3 platform as a whole	Terra	Seven Bridges
DRS 1.x	DRS 1.0 rollout scheduled April; availability gated by ISA	DOS supported; DRS 1.0 rollout to be scheduled	DOS supported; DRS 1.0 rollout to be scheduled	DOS supported; DRS 1.0 rollout to be scheduled	DRS 1.0 Supported*	DRS compatible server in dev (Terra Data Repo)	Servers (CGC, Cavatica) in prod

DRS 1.0 and 1.1 are the same from a server perspective

Terra and SB are included here for the use case of interoperability of derived files

DRS is fairly low risk now and is a high value opportunity for a common data access mechanism

*Current systems implements DRS 1.0 spec with additional prefixes, with bundles available in May

General Interoperability Next Steps - DRS

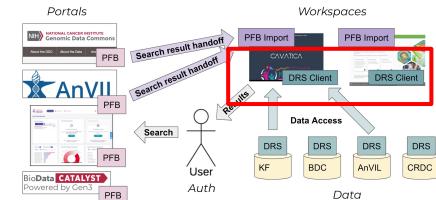
Presenter: Brian O'Connor

DRS Client:

Supporting the ability to access data via DRS 1.0/1.1 in a workspace environment

	Gen3	Seven Bridges	Terra
DRS 1.0	DRS 1.0 supported within Gen3 workspaces	Functional prototypes; DRS client in dev.	Supports a pre-release DRS 1.0 (used for BDCat, TCGA, HCA)
DRS 1.1	DRS 1.1 client planned upon spec finalization	DRS 1.1 client planned.	Evaluating proposal, will update Terra to use DRS 1.1 once available

This table shows the workspace environments that support DRS 1.0 and status of DRS 1.1. DRS 1.1 doesn't change the service, *it changes DRS clients to support GUID compact identifiers* like those used in Gen3 environments for AnVIL, Kids First, BDCat, and CRDC. Opportunity is DRS 1.1 gives ability to work with Data GUIDs and other services. Risk in the time it takes to approve and release standards.

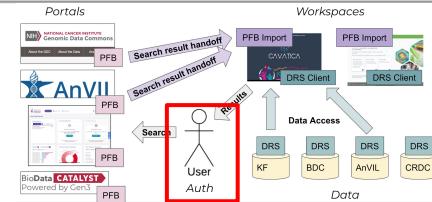


General Interoperability Next Steps - AAI

Presenter: Brian O'Connor

RAS:

A common auth mechanism from NIH. This will identify users and also provide information about the consent groups they have access to. Data servers (Gen3), workspace environments, and portals need to support this over the next 6-12 months.



	UChicago	Broad	Seven Bridges	CHOP
RAS / AAI	<p><i>Currently evaluating RAS documentation and prototyping against RAS and GA4GH Passport and Visa spec;</i></p> <p><i>Coordinating co-development with NIH RAS team</i></p>			

Opportunity is a consistent auth mechanism across NIH projects

General Interoperability Solution - Tech Safe Harbor

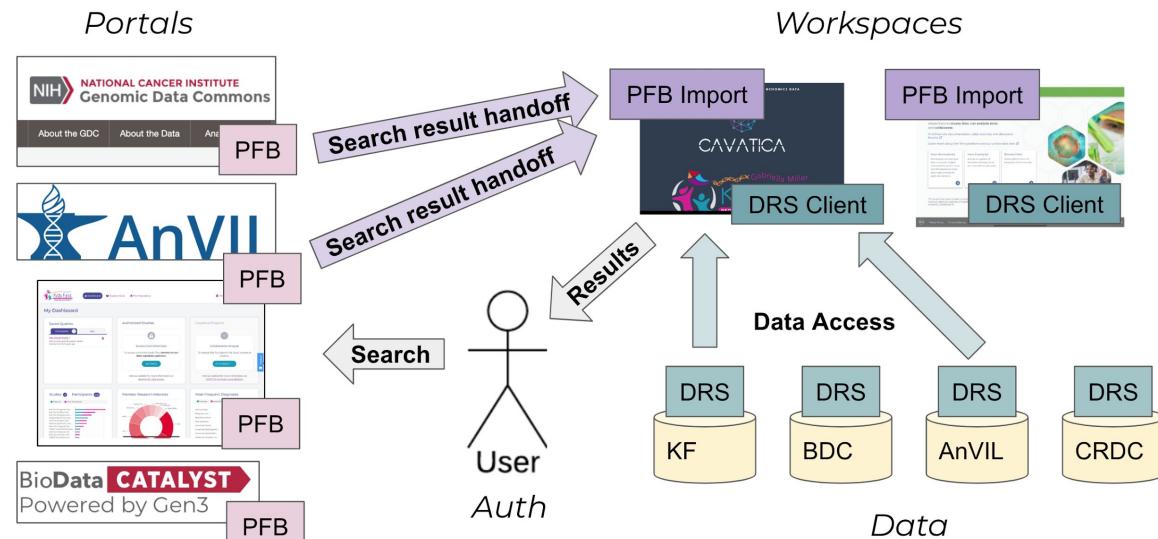
Presenter: Brian O'Connor

Our Technical “Safe Harbor”

Filling these technical gaps doesn't *finish* interop but it gives us a *solid foundation* to improve the efficiency and accessibility of the use cases through improved interop.

What is the timeline?

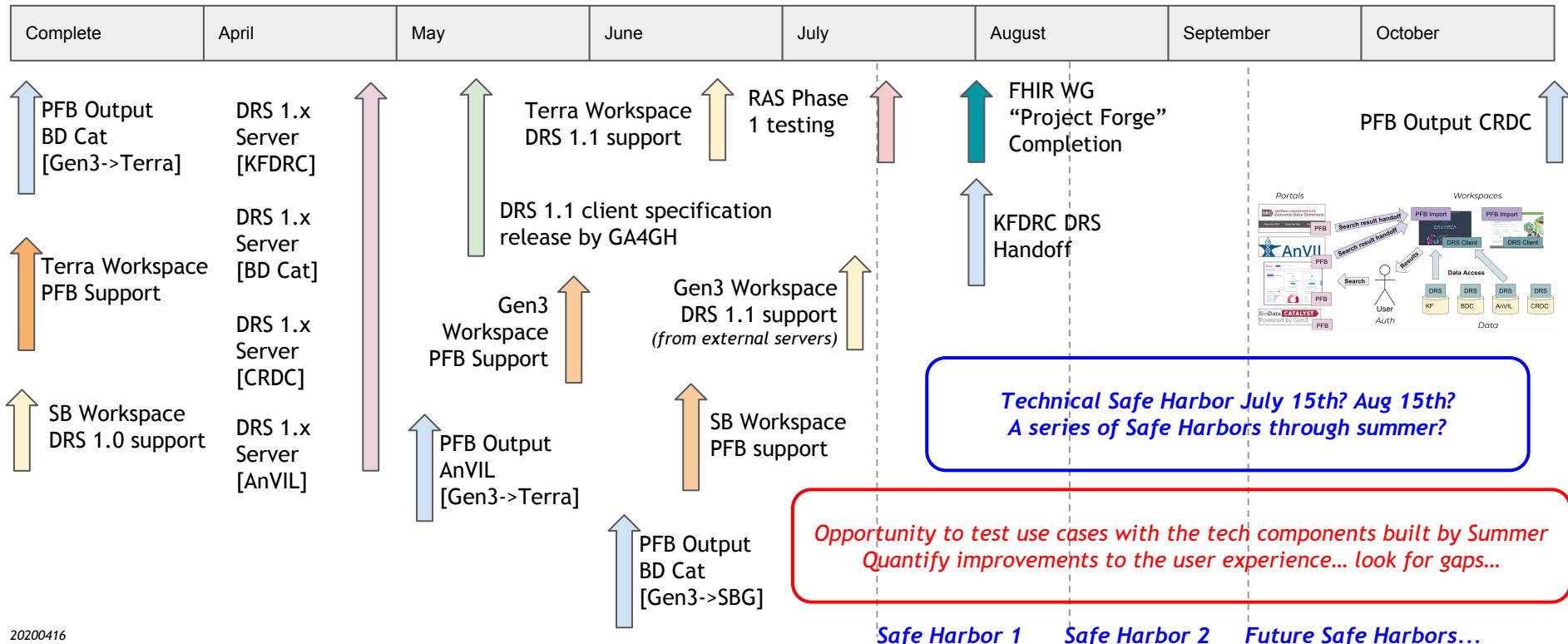
Can we align our project timelines?



General Interoperability Solution - Tech Timeline

Presenter: Brian O'Connor

Given the technical gaps, what might a timeline be for filling these?



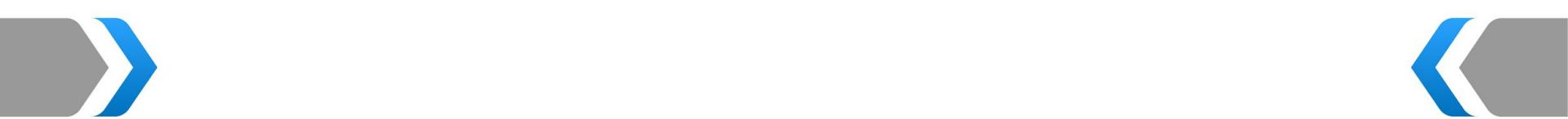


General Interoperability Solution - Goals for Oct 2020

Presenters: Valentina Di Francesco

NIH Input Needed

- Taken 1) technical and 2) scientific gaps, what are priorities from the perspective of the IC leadership?
 - How to prioritize **use cases**?
 - How do we **align our individual timelines**?
 - How to **balance** interoperability efforts versus local system priorities?
 - *Translate the use cases and technical timeline to our **goals for 2020***
 - What are our safe harbors over 2020?
 - How do we measure success?
 - How do we translate interoperability technical achievements and progress on use cases to broad researcher success?
 - Ratification of the **Interoperability Principles**
 - Developing appropriate **metrics** to evaluate adherence to the Interop Principles
- Addressing **other interoperability challenges**: general adoption of DUOS/Passport, common metadata models, managing costs, data security requirements in a federated system, platform branding



Closing

Questions?

All

Next Steps

Rich Silva

Closing Remarks / Thank You!

Anthony Philippakis

