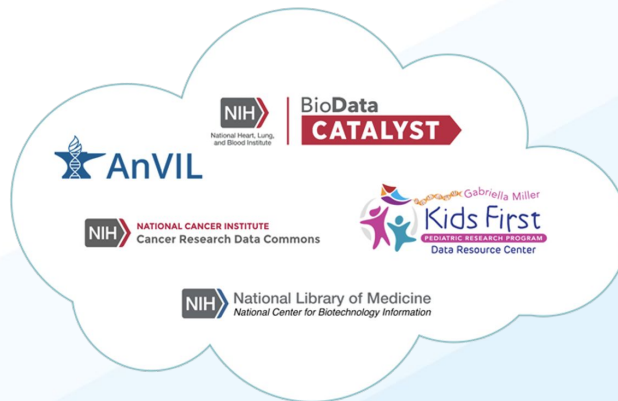


Welcome to Day 1...

NIH Cloud Platforms Interoperability Fall 2021 Workshop

We'll be starting shortly!



Welcome

Stan Ahalt, Patrick Patton

Virtual Meeting Roles (Patton)



Role	Purpose	Assignee & Slack	
Maestro: Mute Master, Raised-Hand Monitor, & Security	Master of Zoom Ceremonies. Contact Amanda for questions about Zoom issues, breakout rooms, or other general questions or if you notice suspicious activity.	@Amanda Miller (amiller@renci.org)	
Screen Sharing	Will share screen and advance slides.	@Julie Hayes	
Slide Content	Will update slide content throughout the meeting.	@Sarah Davis	
Moderator	Moderator listed for each agenda item. Moderator will prompt slide transitions during presentations and foster productive conversation during discussions.	Becky Boyles (@rboyles)	Stan Ahalt (@stan)
Plenary Notetakers	All are encouraged to add comments to the Homepage and Meeting Notes		
Q&A Monitor	Monitor questions in #oct_workshop Slack channel as well as Zoom Chat. Share Action Items, Decisions, and Outstanding Questions from Slack and Zoom to the Homepage and Meeting Notes	@Patrick Patton @Paul Kerr @Allie Gartland-Gray	@Joe Asare @Tom Madden @John Cheadle
Time Watcher	Will try to keep us on time while still allowing room for important conversations.	@Sarah Davis	



Questions during the event? (Patton)



Verbal Questions: There will be time for questions throughout the meeting. If you want to verbally ask a question, use the Zoom feature to "raise your hand" and the host will enable your audio and then call on you to ask your question.

Zoom Chat: You can type questions via Zoom Chat throughout the meeting. Paul Kerr, Patrick Patton, Joe Asare, Allie Gartland-Gray, Tom Madden and John Cheadle will share questions from Slack and Zoom chat into the [Homepage and Meeting Notes](#).

Slack: Questions can be asked throughout the meeting by using the [#oct_workshop](#) Slack channel. We encourage anyone to write questions, comments, answers, or discussion in Slack at any time. If you have not received an invitation to [#oct_workshop](#), please email amiller@renci.org.



The latest version

Want the ability to move independently between breakout sessions?

We updated the meeting settings to allow attendees to move freely between the breakout rooms. **This setting requires the latest version of Zoom.**

- [Follow these instructions](#) or
- Watch this how-to video here: <https://youtu.be/E7zERcVLUBM>



Registration (Patton)



BDC3 will reach out to attendees who have not yet registered to ensure they [register via the form](#) (bit.ly/NCPI2021_Register).

Note that future invitation lists are determined using past registration lists.



BDCatalyst Statement of Conduct (Ahalt)



The BioData Catalyst Consortium is dedicated to **providing a harassment-free experience for everyone**, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of community members in any form. Sexual language and imagery is generally not appropriate for any venue, including meetings, presentations, or discussions.



BDCatalyst “Santa Cruz Rules of Engagement”:

- Do not shy away from identifying problems & risks
- Be candid
- Be heard
 - Identify an ally or motivate via Slack
 - Reach out to a Contact for particular topic(s) - Slack or email bd3@renci.org if you don't know the Contact
- Be polite
 - If you are a “talker” remember to give others time/space to talk - if you are “quiet”, take advantage of any opening
 - Add your comments/ideas to notes if you don't find space to talk!

Connecting Data, Enhancing Software...What Does a Data Ecosystem Look Like?

Susan Gregurick

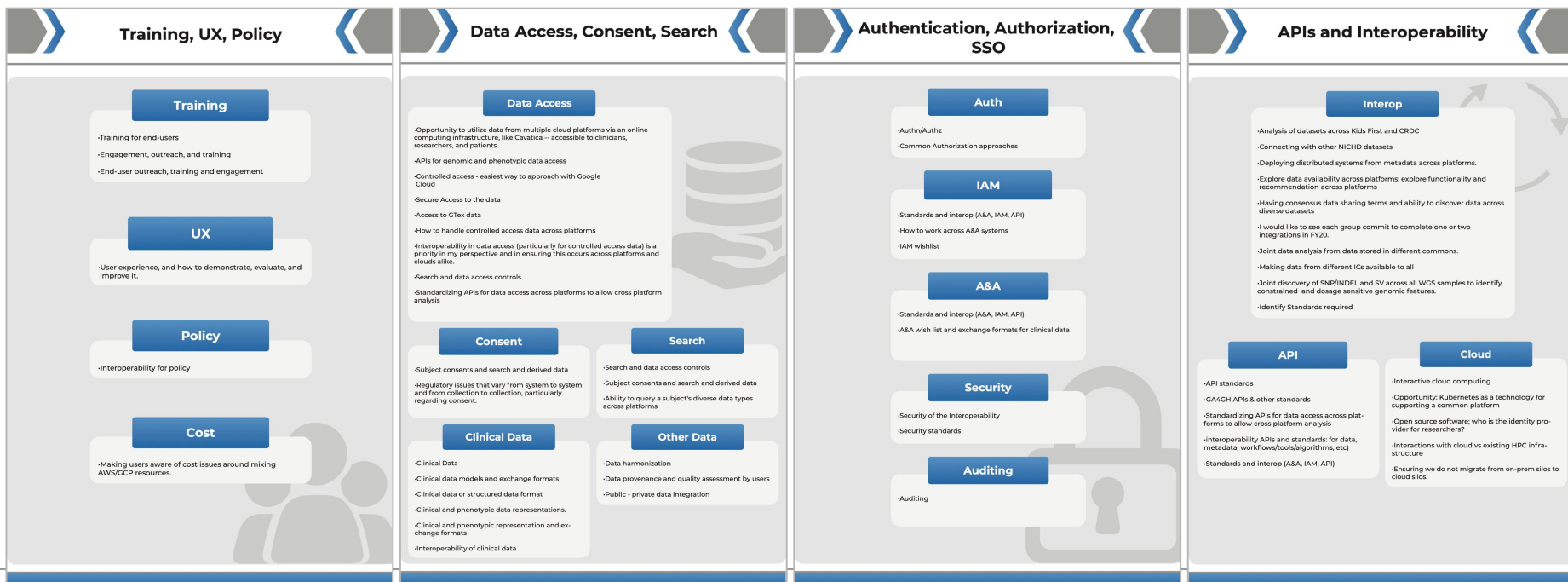
Goals Day 1:

Calibrate, Catalog, Identify Gaps/Challenges

Stan Ahalt

NCPI: Marking Progress (Ahalt)

- 2-years since first NCPI meeting in Chapel Hill, which focused on brainstorming the world of potential activities.





Since then... (Ahalt)



- 5 working groups moving forward on policy and development
- Multiple use cases driving progress
 - E.g. BDCatalyst used funds from ODSS and NHLBI to support development of interoperable AuthZN methods, search capabilities, semantic harmonization, and cross-platform compute on Kids First and AnVIL
 - More updates on driving use cases on Day 2

See all the good work accomplished to date in the [Working Group Executive Summaries](#).

What Does a Data Ecosystem Look Like?

Data, Software enhanced to support the FAIR and CARE Principles

Plan prospectively on how you will handle data;

Repository's ability to easily share data, metadata, and enhance findability across repositories

Software engineering and best practices enhanced for data science

Enhance an open community to work and communicate on software engineering

Cloud-enabled data analytics platforms can cross siloed boundaries, enable greater usability for researchers

- Participants in studies easily findable, data disambiguated
- Sustainability, sharing data, making available metadata and standards more compatible across systems

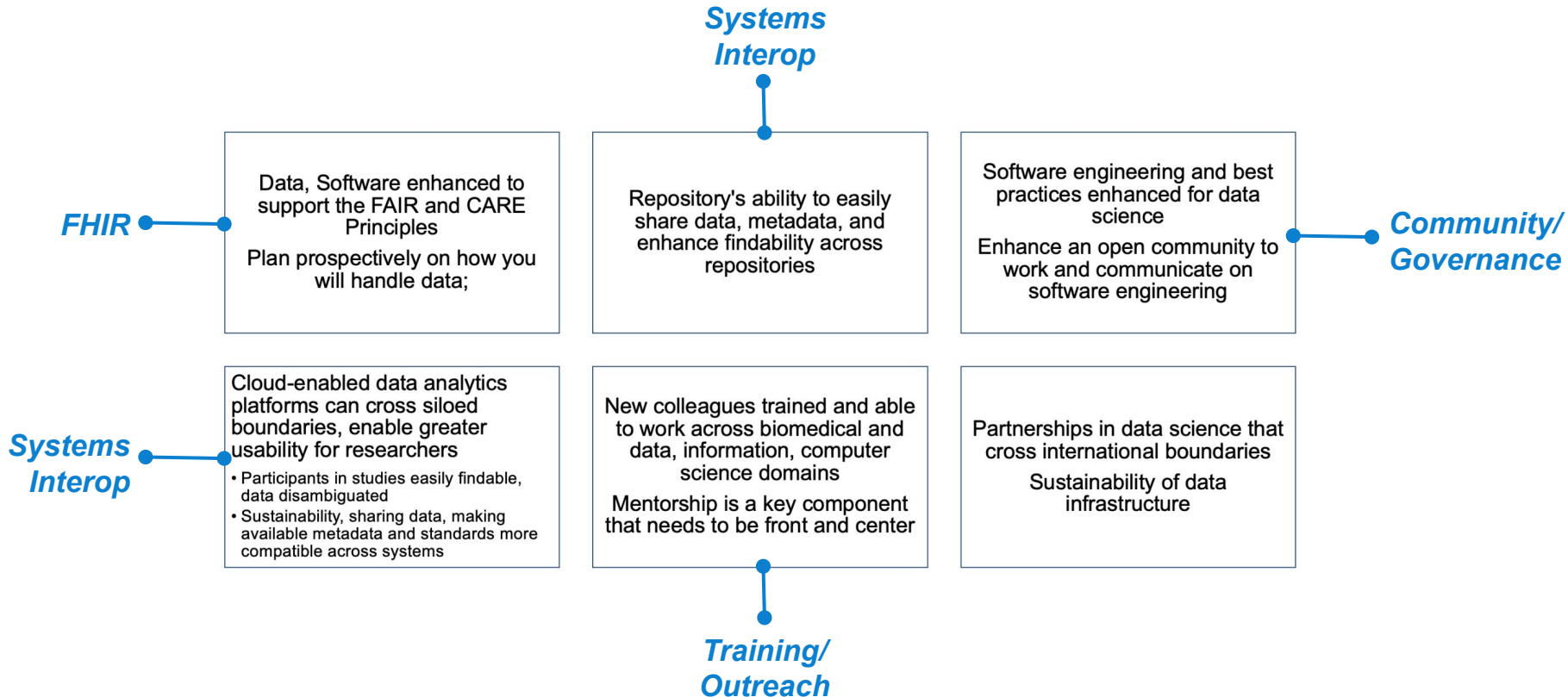
New colleagues trained and able to work across biomedical and data, information, computer science domains

Mentorship is a key component that needs to be front and center

Partnerships in data science that cross international boundaries

Sustainability of data infrastructure

NCPI Working Groups (Ahalt)





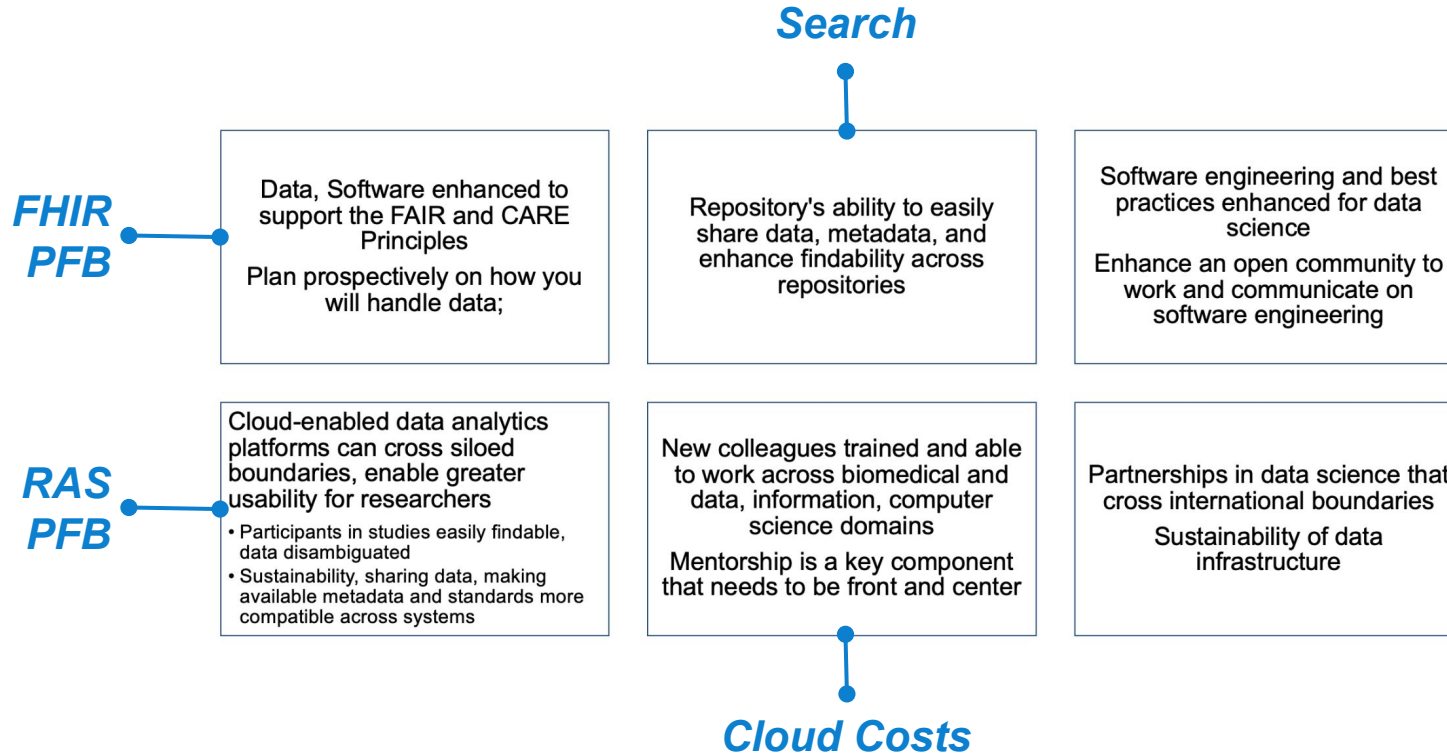
Workshop Goals: Getting to “done” (Ahalt)



- Meeting Agenda is focused on actionable key topics to help reach the ODSS goals
 - RAS
 - PFB
 - FHIR
 - End User Cloud Costs
 - Search
- Catch-all Other Interoperability Efforts gathers other activities that we are working on and what's coming next



Key Topics (Ahalt)





Workshop Goals: Getting to “done” (Ahalt)



- How can we **move the needle forward** on each key topic?
- What is the status of **use cases driving progress**?
- Where are the **gaps** for these topics that might need new use cases?
- What are **policy and development blockers** and how can we unblock them?
- What are the **next key pieces** that will help reach NIH goals?

Agenda: Day 1 All times ET (Ahalt)

Time	Activity	Owner	Links
11:00-11:05am	Welcome	Stan Ahalt, Patrick Patton	Slides Notes
11:05-11:40am	Connecting Data, Enhancing Software...What Does a Data Ecosystem Look Like?	Susan Gregurick	Slides Notes
11:40-11:50am	Goals Day 1: Calibrate, Catalog, Identify Gaps/Challenges	Stan Ahalt	Slides Notes
11:50 -12:15pm	Demo of Successful Federated Use Case (from search to FHIR to workspace)	Brian O'Connor, Jack DiGiovanna, Robert Carroll	Slides Notes
12:15-1:00pm	Updates on Key Topics (Part 1) •PFB (10 min) (Grossman) •FHIR (15 min) (Carroll) •RAS (20 min) (O'Connor)	Moderator: Becky Boyles	Slides Notes
1:00-1:45pm	Lunch Break		
1:15-1:45pm	Lunch Breakout 1: Discuss Gaps and Decide on Concrete Next Steps •RAS and data access (O'Connor)	Brian O'Connor	Slides Notes
1:45-2:35pm	Updates on Key Topics (Part 2) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt)	Moderator: Becky Boyles	Slides Notes
2:35-3:05pm	Breakout Session 2: Discuss Gaps and Decide on Concrete Next Steps •PFB (VanTol) and FHIR (Carroll) •Other Interoperability Efforts (Ahalt)	Robert Carroll, Stan Ahalt	Slides Notes
3:10-3:15pm	Break Plan for Day 2	Becky Boyles	Slides Notes
3:10-4:00pm	Breakout Session 3: Discuss Gaps and Decide on Concrete Next Steps •End-user Cloud Costs (Schatz) •Search (Rogers) (EasyRetro)	Michael Schatz, David Rogers	Slides Notes
Day 2: Wednesday, October 6			

Agenda: Day 2 All times ET (Ahalt)

Time	Activity	Owner	Links
11:00-11:10am	Welcome and Goals Day 2: Synthesize next steps, driving use cases, determine NIH/NCPI priorities	Stan Ahalt	Slides Notes
11:10-12:40pm	Breakout Report Backs and Discussion <ul style="list-style-type: none"> •PFB (10 min) (Grossman) •FHIR (10 min) (Carroll) •RAS (20 min) (O'Connor) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt) 	Moderator: Becky Boyles	Slides Notes
12:40-12:50pm	GA4GH Relationship	Brian O'Connor	Slides Notes
12:50-2:00pm	Lunch Break		
1:30pm-2:00pm	NIH Breakout: NIH Coordination Working Group Discussion of Priority Next Steps	NIH Only (via separate invitation)	
2:00-2:15pm	Use Case Overview: The Journey of a NCPI Use Case	Asiyah Lin	Slides Notes
2:15-3:20pm	Review of Current Scientific Use Cases	Moderator: Valentina Di Francesco	Slides Notes
2:15-2:30pm	Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems	Owen Hirschi	Slides Notes
2:30-2:50pm	Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA	Valerie Cotton, Allison Heath	Slides Notes
2:50-3:05pm	Genetic Sex as a Biological Variable and X-inactivation	Melissa Wilson	Slides Notes
3:05-3:20pm	Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra	Simran Makwana	Slides Notes
3:20-4:00pm	Synthesize Goals and Next Steps for the next 6 Months, with focus on driving use cases	Stan Ahalt, Jon Kaltman	Slides Notes

Meeting Deliverable: [NCPI Glossary](#) (Ahalt)

- While we often use the same words, we sometimes use them to mean different things.
- We hope this [Glossary](#) will be a concrete deliverable at the end of the meeting to help us coalesce around common definitions and/or highlight differences.
- Please review and add your definitions to listed words or add new words

Glossary

Metadata

Semantic

Search

API

Portal

Proof of Concept

Pilot

AuthN/AuthZ

Data Stewards

[Add your word here]

Demo of Successful Federated Use Case: From Search to FHIR to Workspace

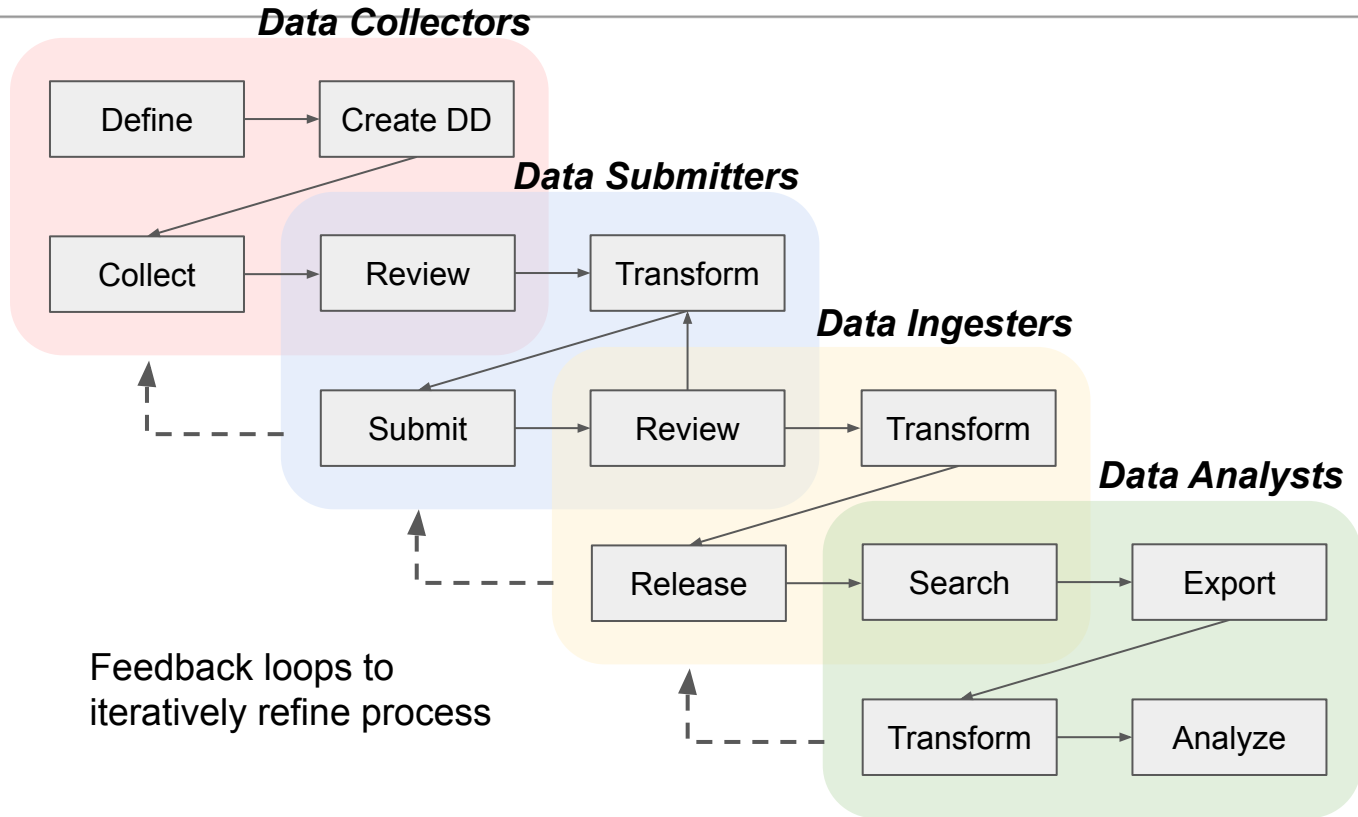
Brian O'Connor, Jack DiGiovanna, Robert Carroll



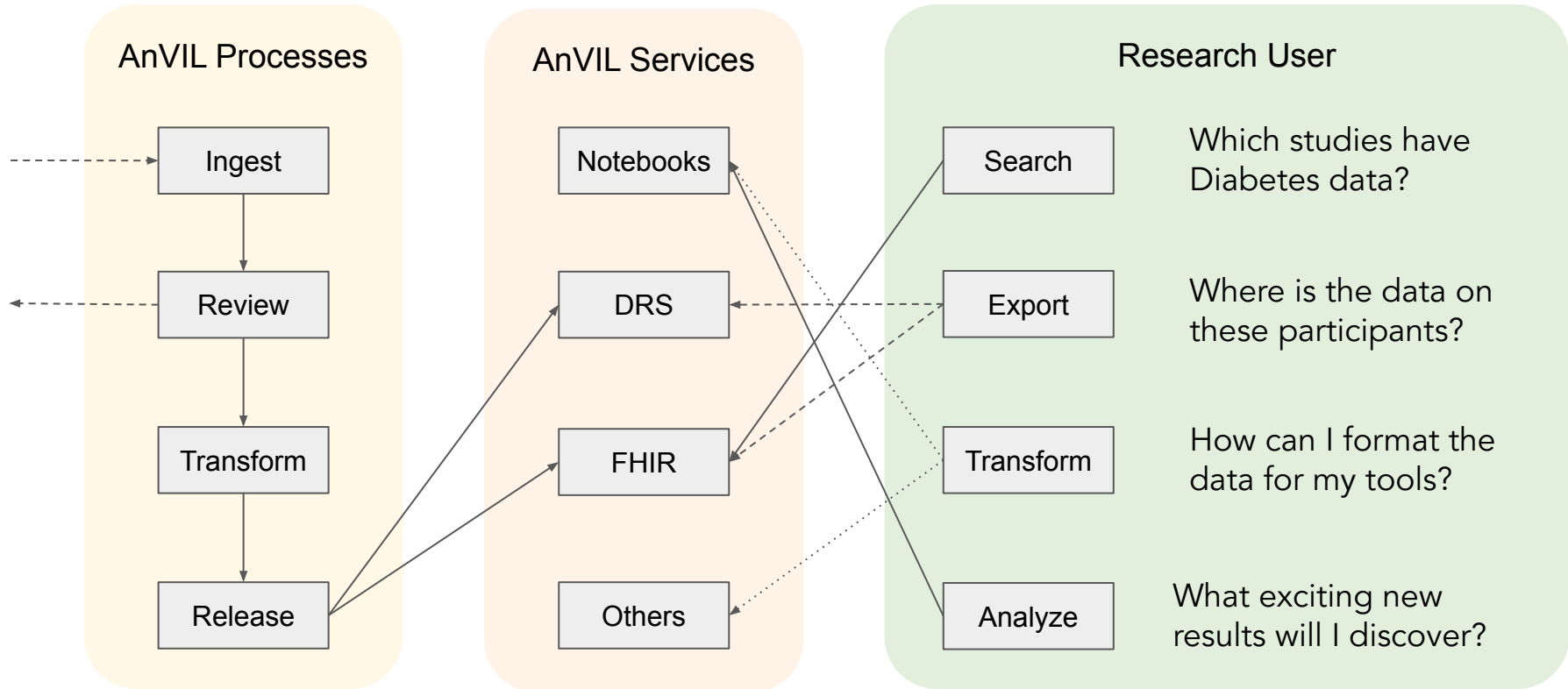
FHIR & Search

Data Access & Compute

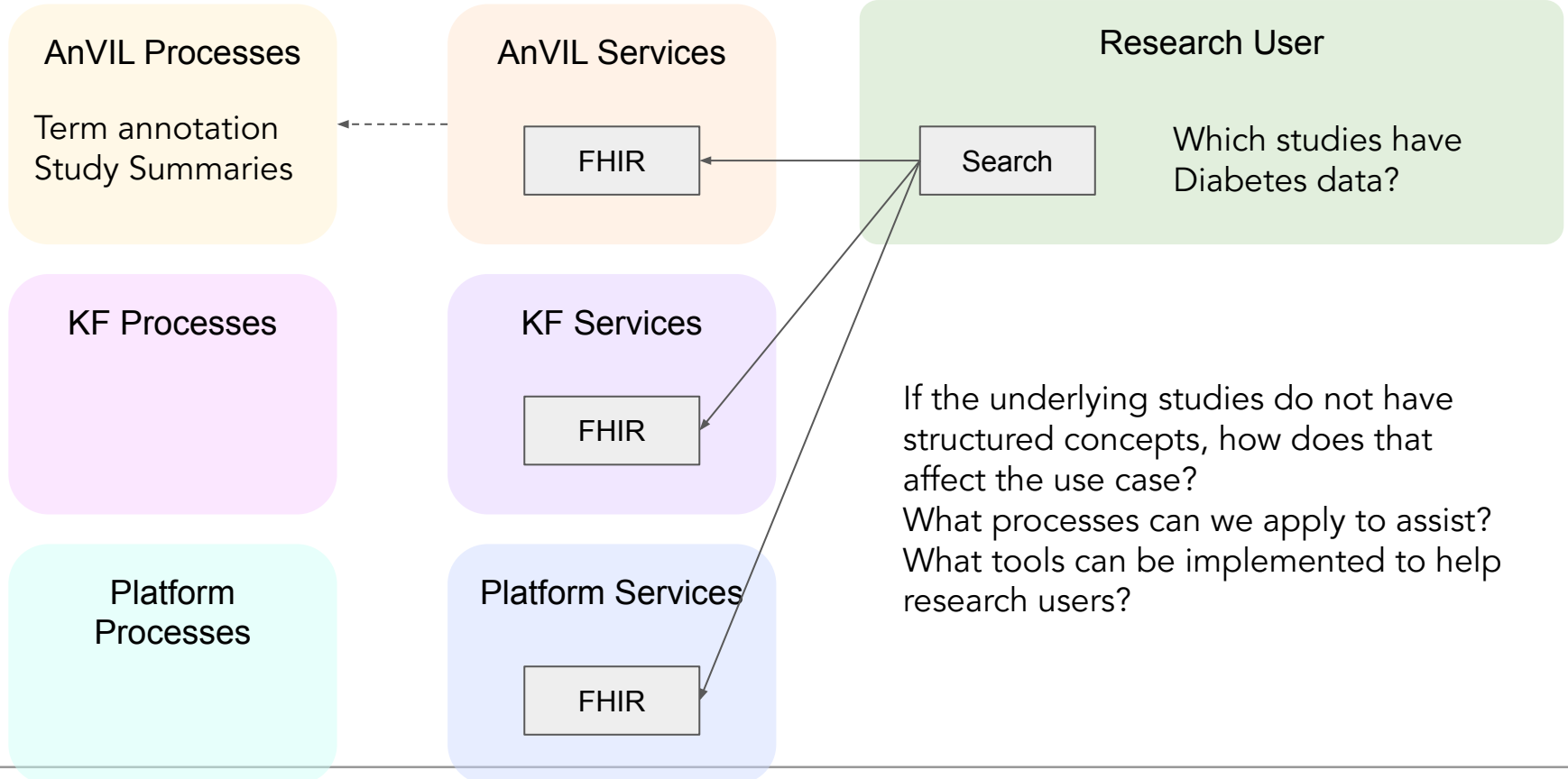
Data Life Cycle



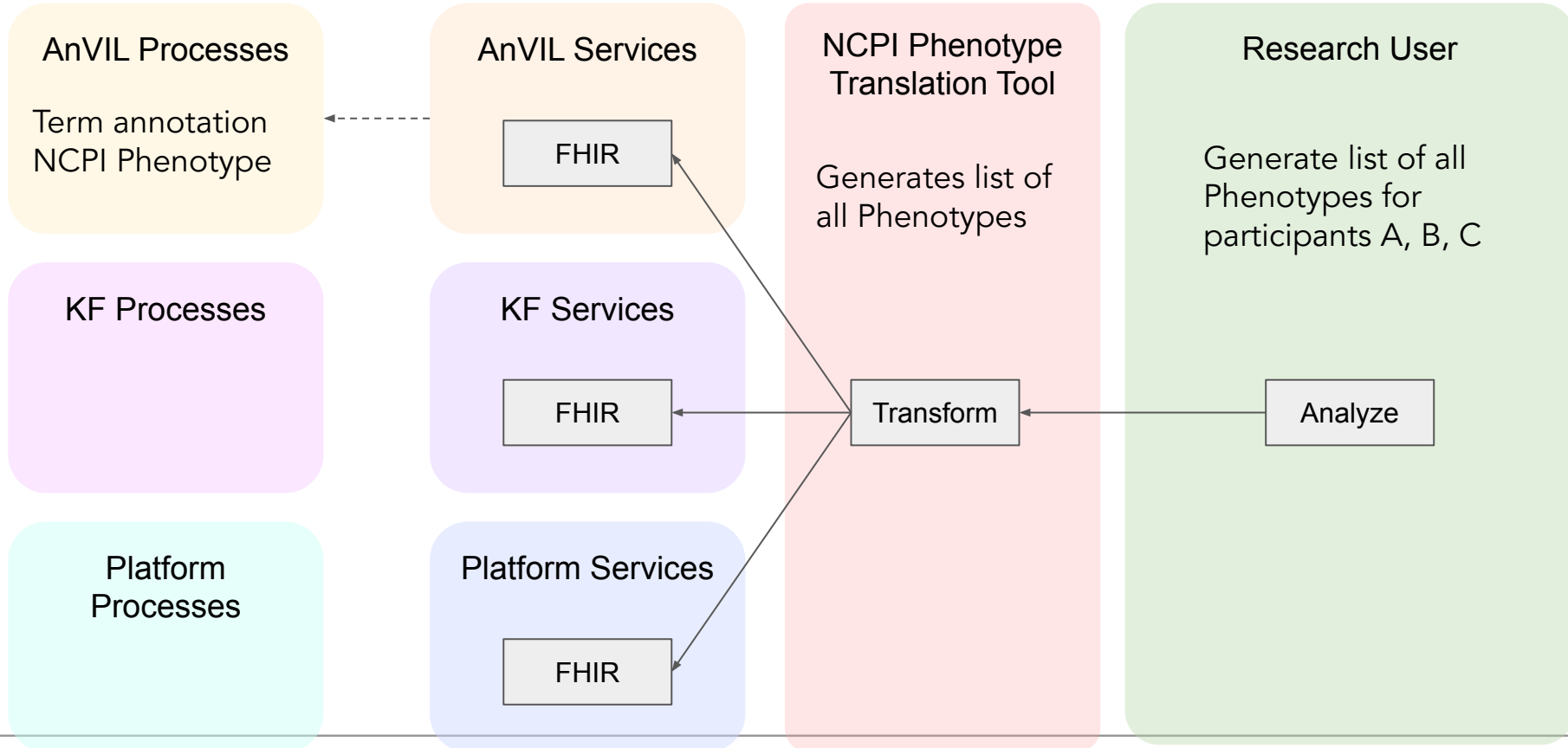
Platforms and Research Users



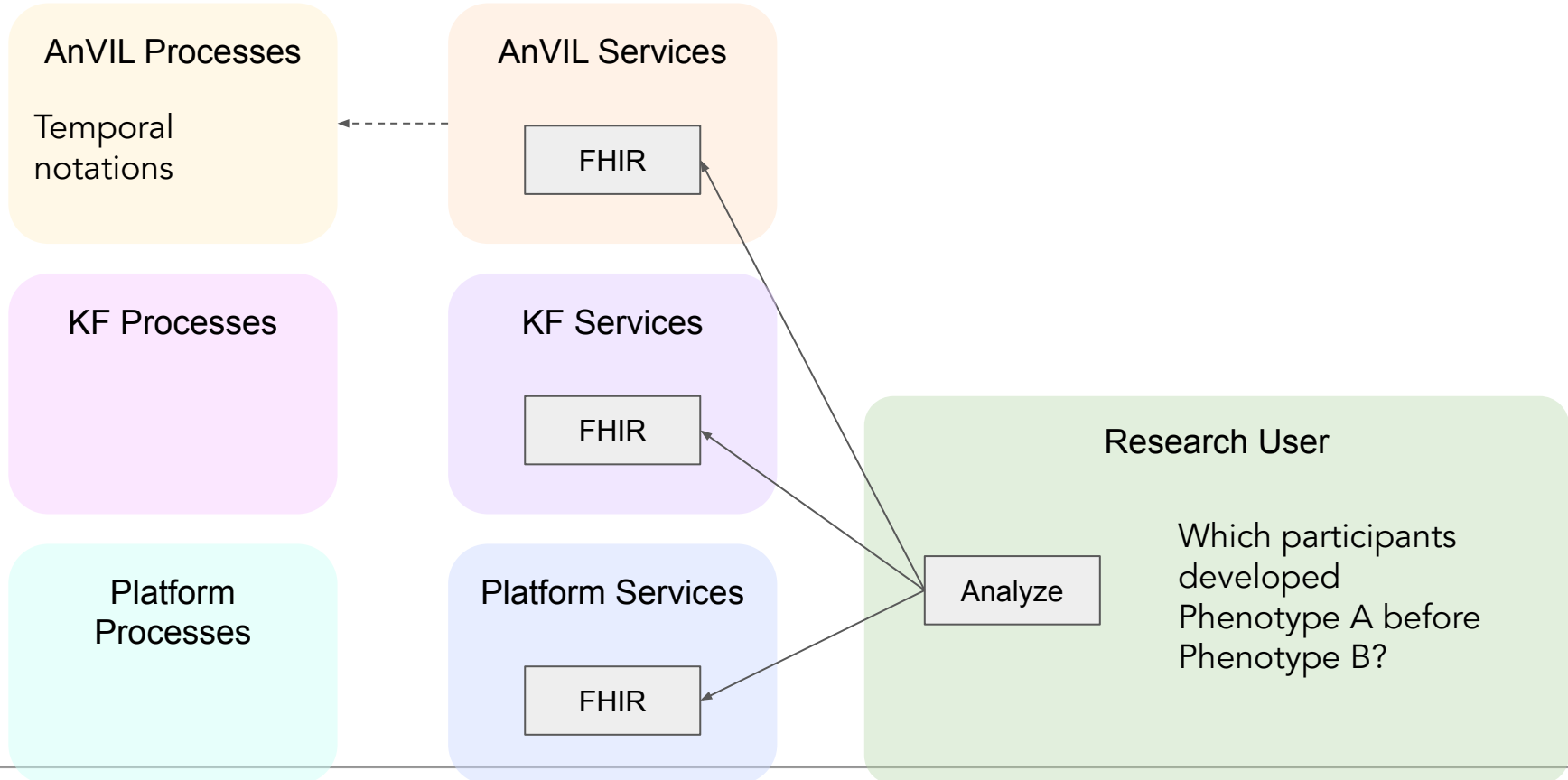
Study summary use cases



Additional tool / service layers



Platforms and Research Users



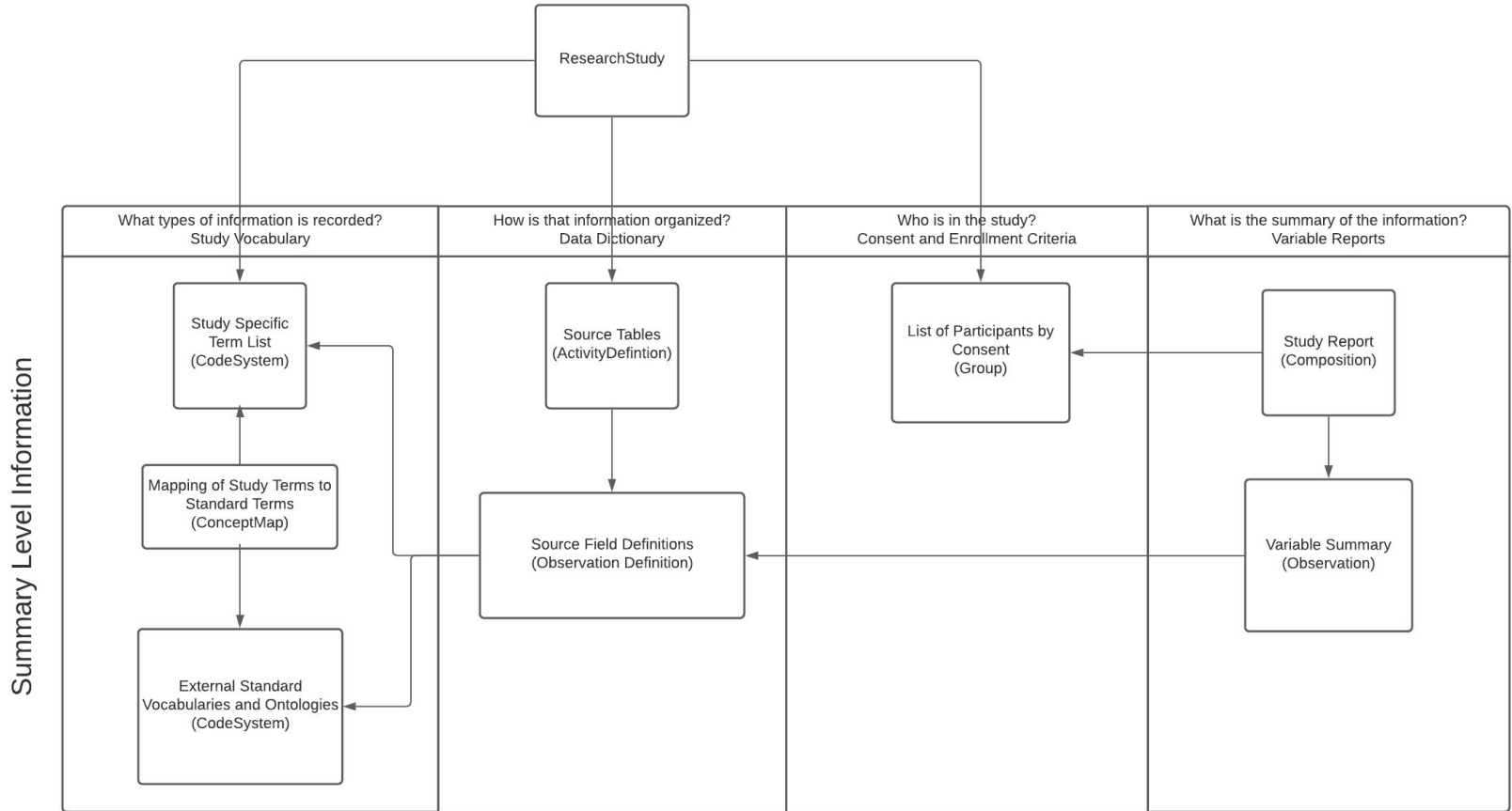


Representing Study Data



- Providing detailed study metadata is very important to understanding the data that's presented.
- dbGaP has set the standard for information that's available, and they are working towards "modernizing" the representation.
- This is currently organized in FHIR, but it's using a custom extension approach.
- We have built a proposal using more FHIR native approaches that should enable easy lift-over for existing data, programmatic definitions for new data, and structured links within the metadata.
- This model is developed in the context of researchers accessing existing research data.

Big picture





- FHIR Example: <https://github.com/anvilproject/DD-On-FHIR>
- Study Summary Tool:
<https://github.com/NIH-NCPI/ncpi-study-summary-generation-tool>
- Study Browser Tool:
<https://github.com/NIH-NCPI/ncpi-fhir-study-summary-browser>



Demo!



- Using the NCPI FHIR Implementation guide, we have several studies loaded into FHIR servers.
 - AnVIL internal test server on Google Healthcare API
 - KF development server running Smiles CDR on AWS
- Eric Torstenson developed and ran a ResearchStudy summary tool, which generated summary objects that could be made available publicly.
- I've written a quick Shiny app that looks at those summaries to generate some interactive content.
- Live demo if possible

NCPI Study Summary FHIR Browser

[=> ResearchStudy Browser](#)
[Study Phenotypes Browser](#)
[Configuration](#)

Select a study:

Show entries

Search:

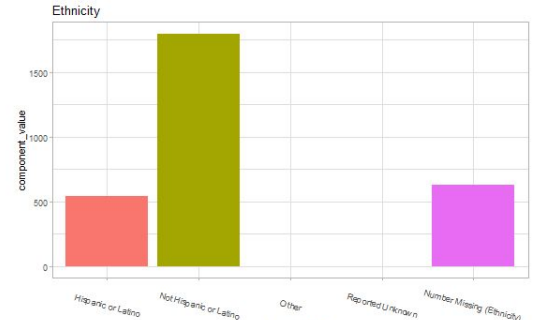
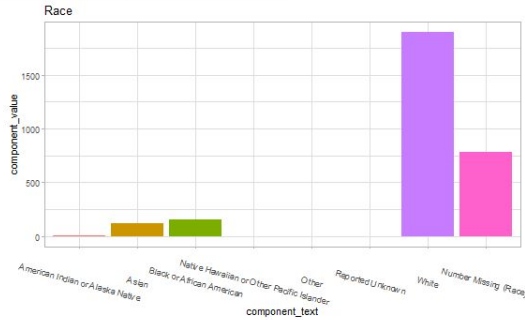
Study Title	Participants
Baylor Hopkins Center for Mendelian Genomics (BH CMG)	1621
National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet Program: The Gabriella Miller Kids First Pediatric Research Program of the Pediatric Cardiac Genetics Consortium (PCGC)	2966
University of Washington Center for Mendelian Genomics (UW-CMG)	2802
Yale Center for Mendelian Genomics (Y CMG)	6979

Showing 1 to 4 of 4 entries

Previous Next

National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet Program: The Gabriella Miller Kids First Pediatric Research Program of the Pediatric Cardiac Genetics Consortium (PCGC)

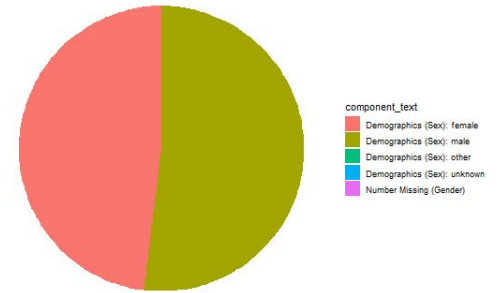
Study ID: f8fe498c-faa4-46eb-81b1-10231dac52d7



■ American Indian or Alaska Native ■ Black or African American ■ Other ■ WI
■ Asian ■ Native Hawaiian or Other Pacific Islander ■ Reported Unknown ■ Nui

Category	Participants
Demographics (Race): American Indian or Alaska Native	5
Demographics (Race): Asian	119
Demographics (Race): Black or African American	159
Demographics (Race): Native Hawaiian or Other Pacific Islander	2
Demographics (Race): White	1897
Demographics (Race): Other	0
Demographics (Race): Reported Unknown	0
Number Missing (Race)	784

Category	Participant
Demographics (Sex): male	154
Demographics (Sex): female	142
Demographics (Sex): other	
Demographics (Sex): unknown	
Number Missing (Gender)	



NCPI Study Summary FHIR Browser

National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet Program: The Gabriella Miller Kids First Pediatric Research Program of the Pediatric Cardiac Genetics Consortium (PCGC)

Study ID: f8fe498c-faa4-46eb-81b1-10231dac52d7

List of groups in this study:

Show 10 entries Search:

Group Name	Participants
SD_PREASA7S-complete	2966

Showing 1 to 1 of 1 entries Previous 1 Next

Phenotype Summary

Show 10 entries Search:

Phenotype	Phenotype Present	Phenotype Absent	Number Missing Phenotype	Phenotype Reported Unknown
Conotruncal Left-sided Lesion	545	191	2230	0
Abnormal Ventricular Septum	419	309	2230	8
Abnormal Ventriculo-arterial Connection	337	396	2230	3
Abnormal Pulmonary Valve	301	421	2230	14
Abnormal Atrial Septum	291	365	2230	80
Abnormal Aorta	285	387	2230	64
Abnormal Right Ventricle	243	491	2230	2
Abnormal Aortic Valve	227	394	2230	115
Abnormal Mitral Valve	194	539	2230	3
Left Ventricular Outflow Tract Obstruction	158	578	2230	0

Showing 1 to 10 of 367 entries Previous 1 2 3 4 5 ... 37 Next

NCPI Study Summary FHIR Browser

ResearchStudy Browser Study Phenotypes Browser Configuration

Select a study:

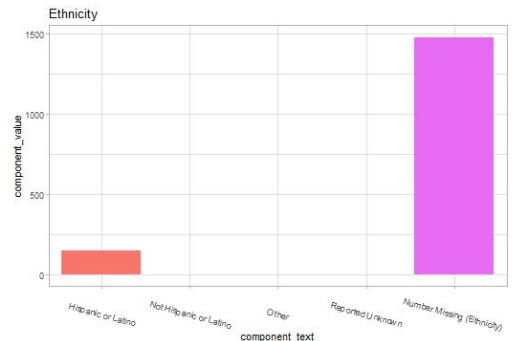
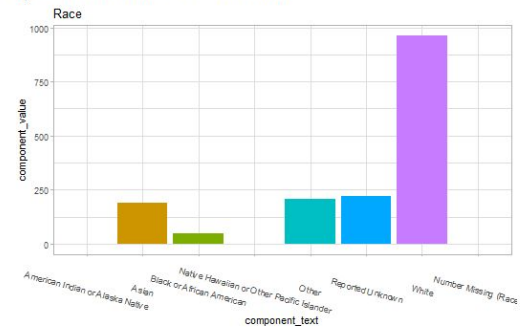
Show **10** entries Search:

Study Title	Participants
Baylor Hopkins Center for Mendelian Genomics (BH CMG)	1621
National Heart, Lung, and Blood Institute (NHLBI) Bench to Bassinet Program: The Gabriella Miller Kids First Pediatric Research Program of the Pediatric Cardiac Genetics Consortium (PCGC)	2966
University of Washington Center for Mendelian Genomics (UW-CMG)	2802
Yale Center for Mendelian Genomics (Y CMG)	6979

Showing 1 to 4 of 4 entries Previous **1** Next

Baylor Hopkins Center for Mendelian Genomics (BH CMG)

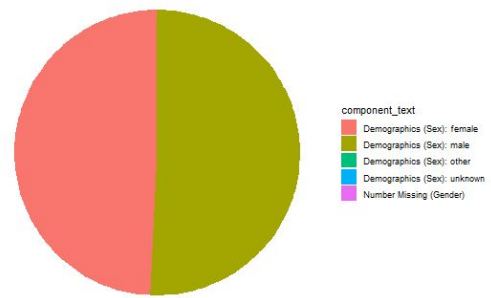
Study ID: cdbdac2b-ff44-4c30-bdaf-e0b806851713



Legend for Race: American Indian or Alaska Native (red), Asian (orange), Black or African American (green), Native Hawaiian or Other Pacific Islander (teal), Other (blue), Reported Unknown (cyan), White (purple), Number Missing (Race) (pink).
 Legend for Ethnicity: Hispanic or Latino (red), Not Hispanic or Latino (green), Other (teal), Reported Unknown (blue), Number Missing (Ethnicity) (purple).

Category	Participants
Demographics (Race): American Indian or Alaska Native	0
Demographics (Race): Asian	189
Demographics (Race): Black or African American	46
Demographics (Race): Native Hawaiian or Other Pacific Islander	0
Demographics (Race): White	961
Demographics (Race): Other	207
Demographics (Race): Reported Unknown	218
Number Missing (Race)	0

Category	Participants
Demographics (Sex): male	~1621
Demographics (Sex): female	~162
Demographics (Sex): other	0
Demographics (Sex): unknown	0
Number Missing (Gender)	0



NCPI Study Summary FHIR Browser

=> ResearchStudy Browser Study Phenotypes Browser Configuration

Baylor Hopkins Center for Mendelian Genomics (BH CMG)

Study ID: cdbdac2b-ff44-4c30-bdaf-e0b806851713

List of groups in this study:

Show 10 entries

Search:

Group Name	Participants
HMB-IRB-NPU	804
HMB-NPU	817
BH_CMG-complete	1621

Showing 1 to 3 of 3 entries

Previous 1 Next

Phenotype Summary

Show 10 entries

Search:

Phenotype	Phenotype Present	Phenotype Absent	Number Missing Phenotype
Global developmental delay	79	0	1542
Scoliosis	76	0	1545
Joint laxity	61	0	1560
Microcephaly	56	0	1565
Hypotonia	51	0	1570
Seizure	46	0	1575
Expressive language delay	45	0	1576
Decreased body weight	43	0	1578
Proportionate short stature	40	0	1581
High palate	39	1	1581

Showing 1 to 10 of 1,312 entries

Previous 1 2 3 4 5 ... 132 Next

NCPI Study Summary FHIR Browser

Baylor Hopkins Center for Mendelian Genomics (BH CMG)

Study ID: cdbdac2b-ff44-4c30-bdaf-e0b806851713

List of groups in this study:

Show 10 entries

Search:

Group Name	Participants
HMB-IRB-NPU	804
HMB-NPU	817
BH_CMG-complete	1621

Showing 1 to 3 of 3 entries

Previous 1 Next

Phenotype Summary

Show 10 entries

Search:

Phenotype	Phenotype Present	Phenotype Absent	Number Missing Phenotype
Global developmental delay	42	0	775
Scoliois	41	0	776
Microcephaly	34	0	783
Seizure	28	0	789
Intellectual disability	22	0	795
Expressive language delay	22	0	795
Peripheral neuropathy	22	0	795
Recurrent infections	21	0	796
Intellectual disability, moderate	15	0	802
Abnormality of the face	15	0	802

Showing 1 to 10 of 727 entries

Previous 1 2 3 4 5 ... 73 Next

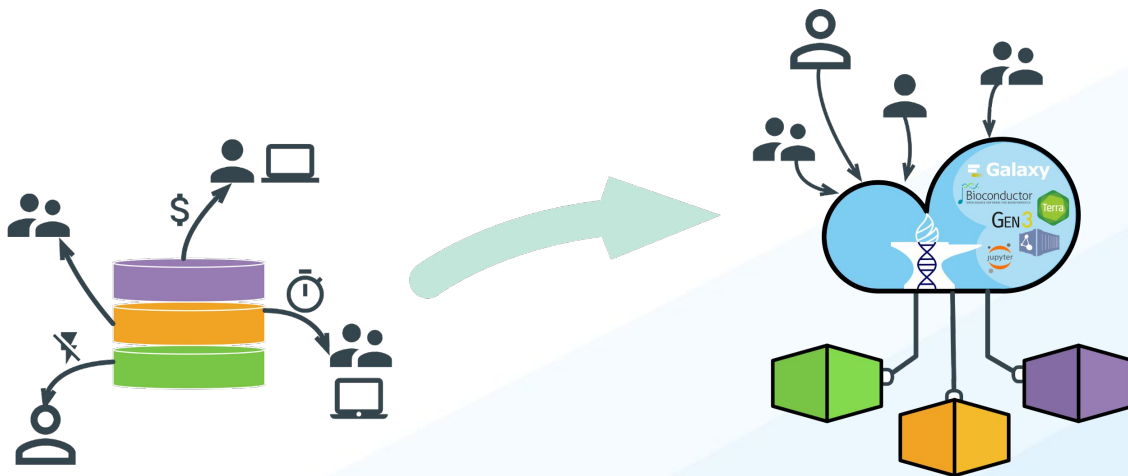


FHIR & Search

Data Access & Compute

Inverting the Model of Genomic Data Sharing

*AnVIL, BioData Catalyst, CRDC, and GMKF have
~11PB of data accessible on the cloud for ~831K participants*



Traditional: Bring data to the researcher

Goal: Bring researcher to the data



NCPI Systems Interoperation WG



The [NCPI Systems Interoperation Working Group](#) spearheads technical improvements to the NCPI participating cloud-based platforms that enable improved interoperability.



<https://anvilproject.org/ncpi>



NCPI Systems Interoperation Working Group -- Use Cases

About

This is our document to capture new use cases as they emerge. Please add yours below. The first five use cases can be found in the Systems Interoperation working group [Charter](#).

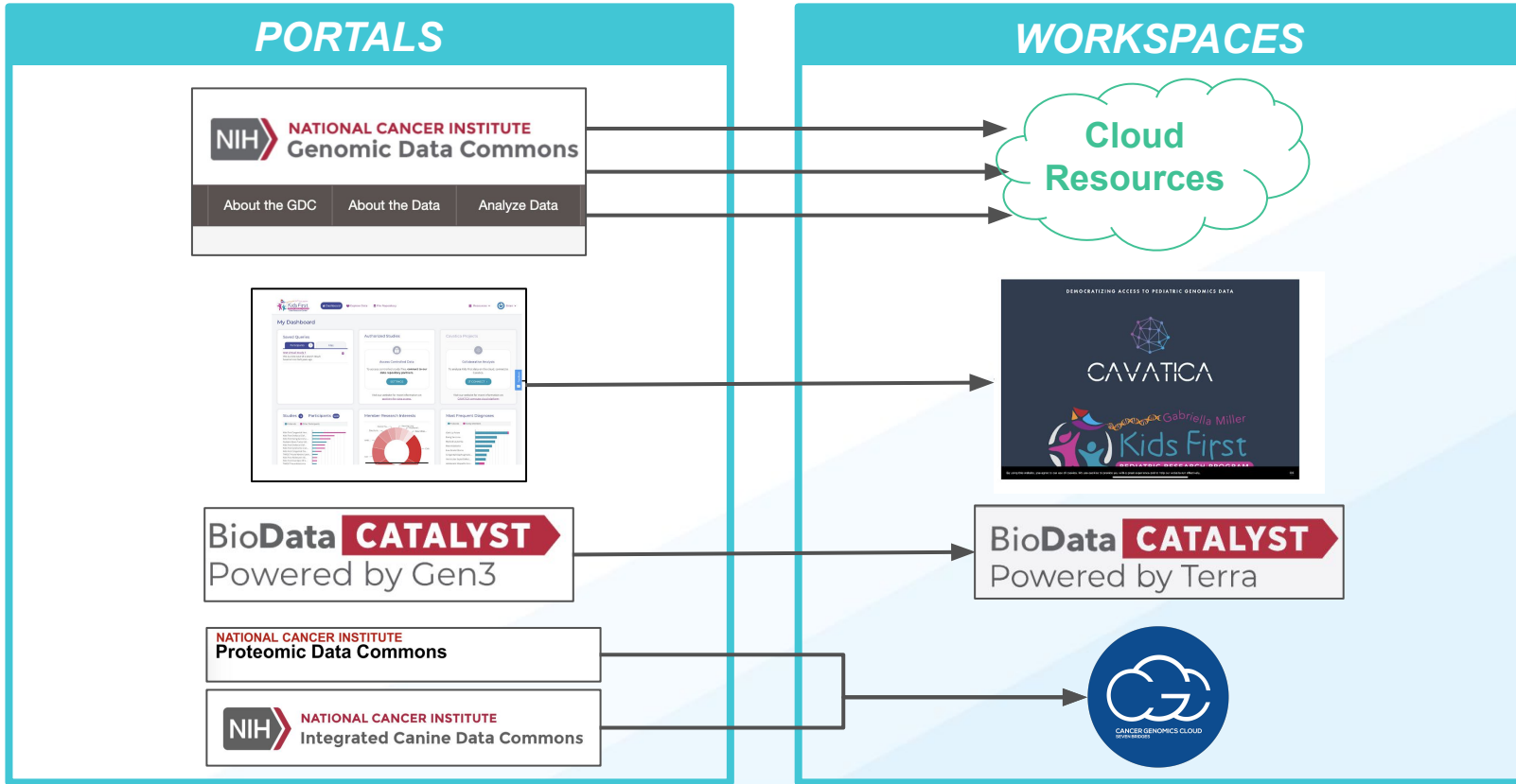
Version 2.0.0 of the charter will cover our work in 2021.

Version	Date	Description
1.0.0	1/17/2020	Initial version, focused on establishing researcher use cases and work in progress. Approved by: <ul style="list-style-type: none">• CRDC – Tanja Davidsen• Kids First – James Coulombe• AnVIL – Ken Wiley and Valentina di Francesco• BD Catalyst – Jonathan Kaltman (approved 1.0.0 on 1/21)
2.0.0	1/2021	pending

*We worked with multiple researchers to define **11 driver use cases** for our work*

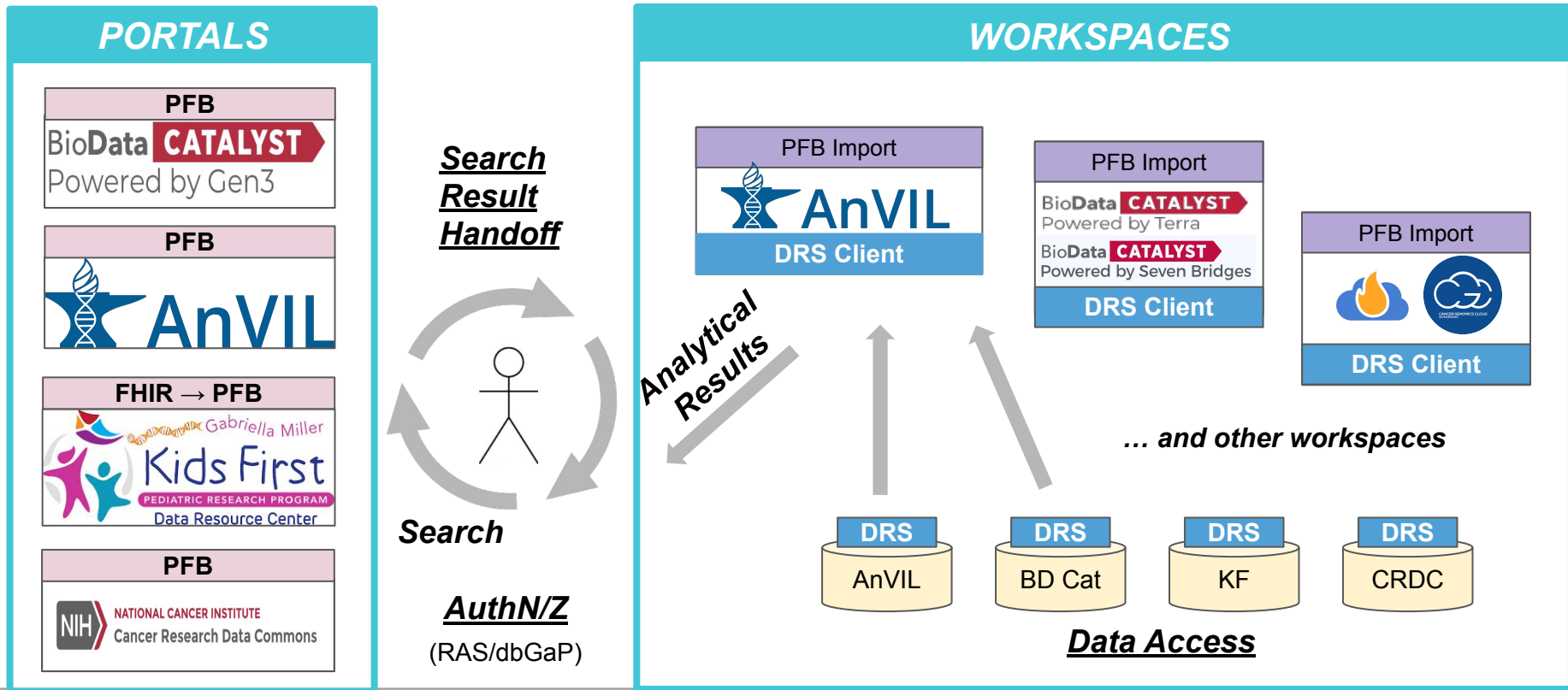
When NCPI Sys Interop Started (Jan 2020)

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



Our Vision for Interoperability

Data portals connect to any workspaces (inter-IC), workspace access data (inter-IC)





3 Key Standards in NCPI Systems Interop



Search Result Handoff:
PFB (FHIR and Manifests)

Data Access: GA4GH DRS

Auth: RAS GA4GH Passports for
AuthN/Z



NCPI Sys Interop's Progress



2020

- **MOUs/ISAs** for RAS and system interconnects
- **PFB for data handoff from portals to workspaces** (BDCat & AnVIL)
- **DRS for data access** to AnVIL, BDCat, Kids First, and CRDC
- Progress on **Researcher Use Cases**

2021

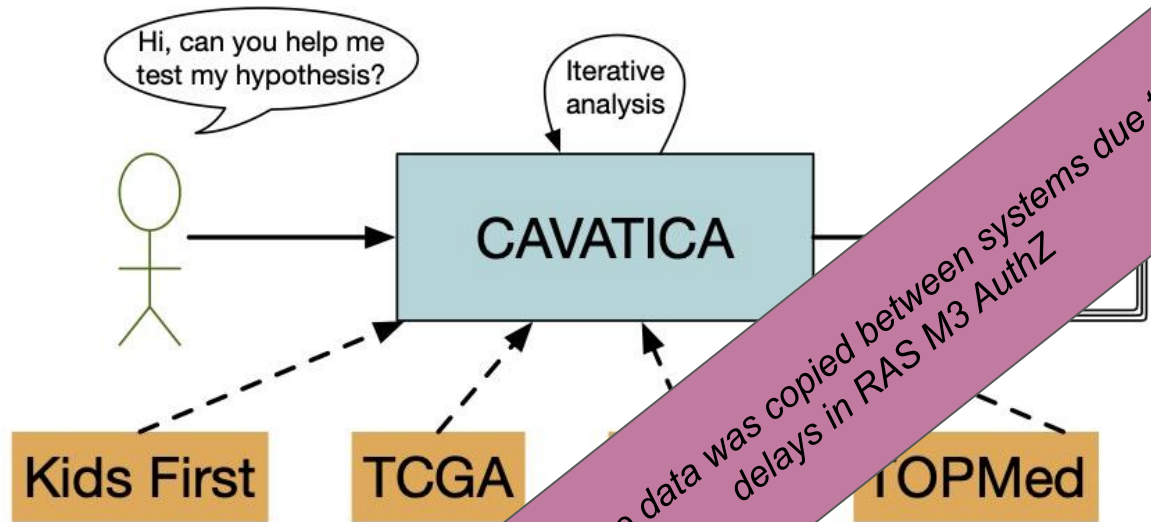
- **NIH RAS for authentication**
- **GA4GH standards** evolved (Passports, DRS, etc)
- More systems working on **PFB handoff** (PDC)
- Prototyping **FHIR** → **PFB bridge**
- More **workspaces supporting more DRS servers**
- **RAS Passports for Authorization** designed
- **Researcher Use Cases** finishing/expanding

Use Case Success Stories

Use Case #5: Wilson McKerrow et al. LINE1 analysis on the CGC spanned Proteomics Data Commons, TCGA, and GTEx

Use Case #1B: Deanne Taylor et al. PCGC analysis on CAVATICA and BDC powered by SB spanned the PCGC data governed by Kids First and PCGC data governed by TOPMed.

Proof of concept: KF, TCGA, GTEx, and TOPMed data in CAVATICA April 2021



UX not yet optimal

Fast (<1
AuthN) 1
but clear
improve
(RAS), a
user bas



Submit Data | Documentation | mpingram@uchicago.edu | Logout

The AnVIL

Dictionary | Exploration | Workspace | Profile

Data | File | Downloadable

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

Filters

Sequencing | **Projects** | Subject | Sample

Collapse all

Project Id 1 selected

- open_access-1000Genomes 3,202
- open_access-Cleversafe_demo 21

Anvil Project Id

- no data 3,202

Project dbGaP Accession Nu...

- open 3,202

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

Export to BioData Catalyst | Export to CGC | Export to CAVATICA

Subjects 3,202

Sex

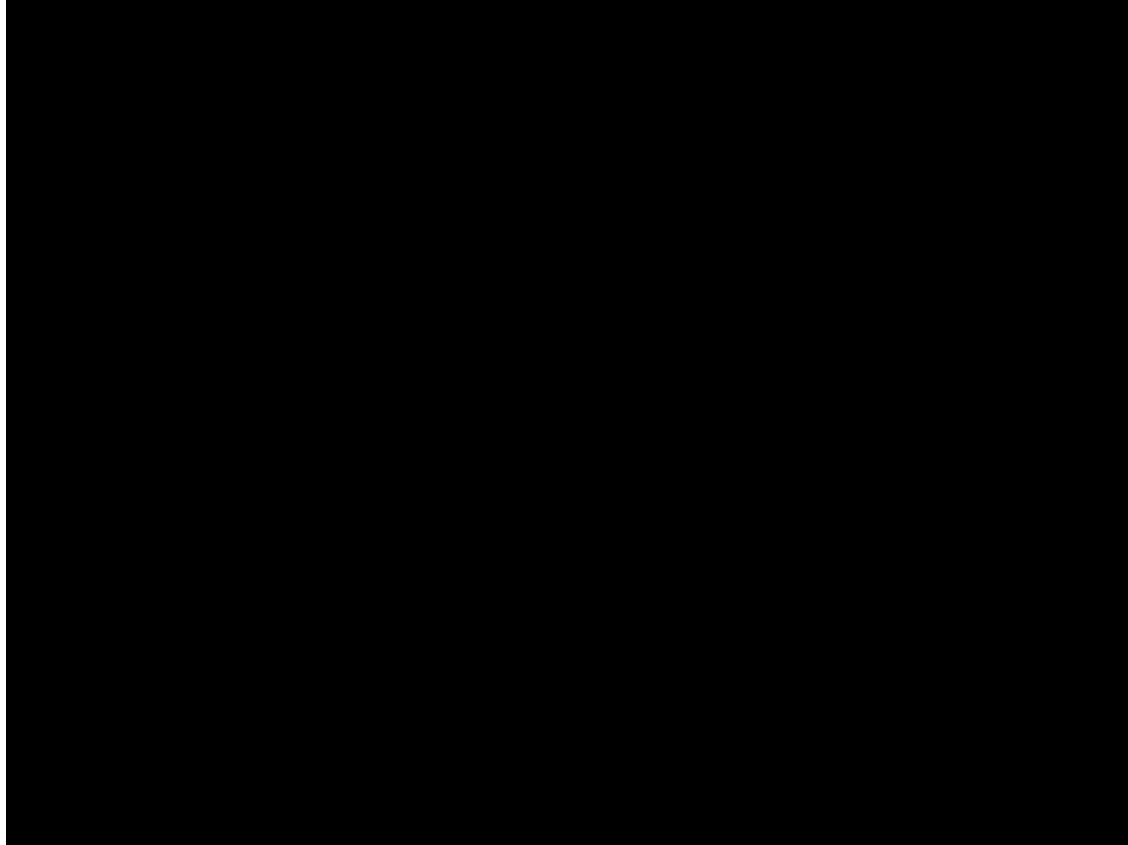
Female	1,271 (39.7%)
Male	1,233 (38.5%)
no data	698 (21.8%)

Ancestry

no data 100%



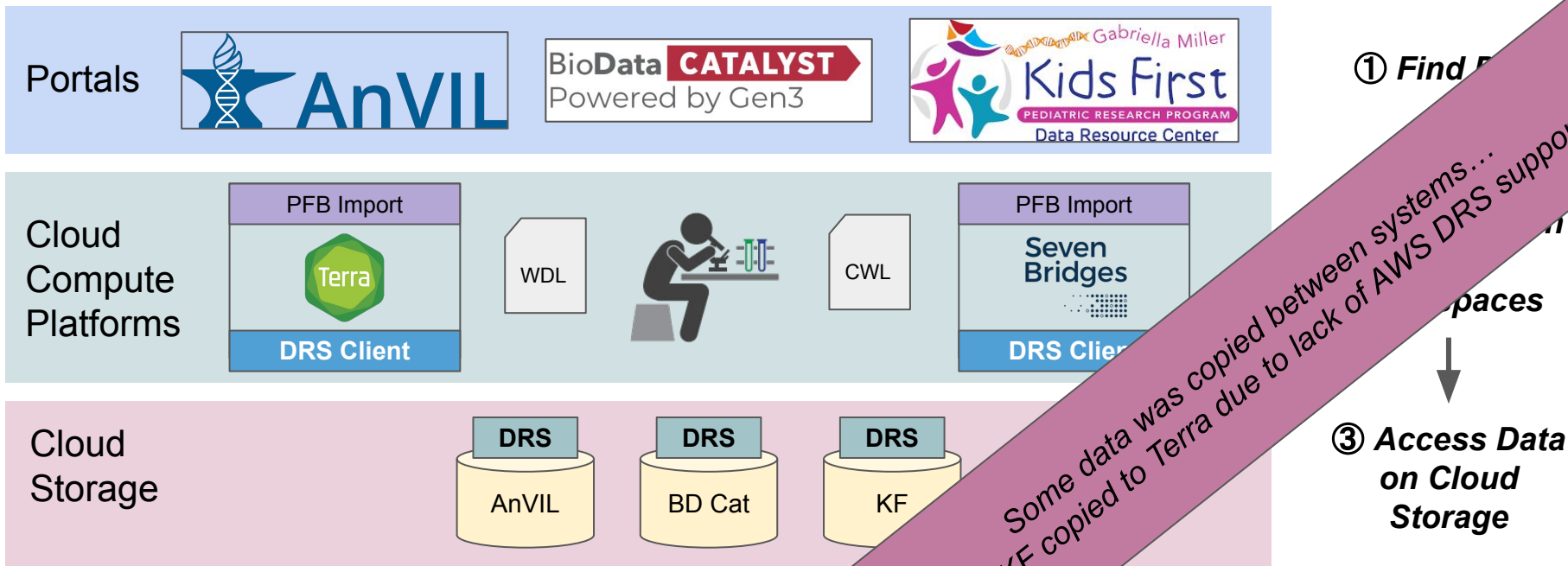
Improved UX



Use Case Success Stories

Use Case #7: Tim Majarian's cross dataset analysis for Congenital Heart Disease

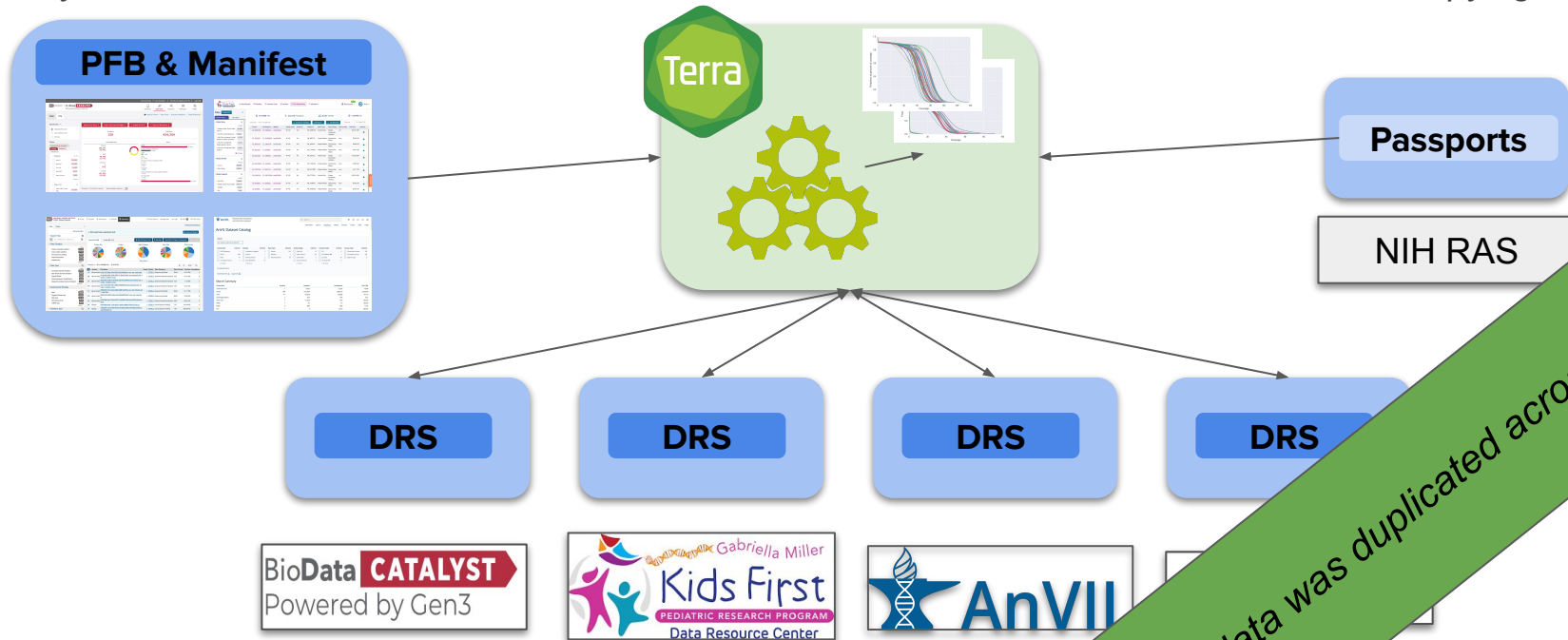
"We performed an association analysis, interrogating the effect of rare exonic variation on CHD risk at a fraction of the cost that would have otherwise been incurred without these interoperability tools."



Our Latest Use Case Success Story

Use Case #11: Melissa Wilson's use case examining Sex as a Biological Variable

Assessing the state of X and Y chromosome calling, we created a Terra workspace referencing AnVIL, BioData Catalyst, CRDC, and Kids First datasets. We used DRS to access data on demand without copying.



No data was duplicated across systems!



Priorities for 2022



Finish RAS Milestone 3

Multiple account links

NHLBI BioData Catalyst Framework Services
Username: BRIANDCONNOR
Link Expiration: Oct 28, 2021, 12:39 PM
Renew [↗](#) | Unlink

NCI CRDC Framework Services
Username: BRIANDCONNOR
Link Expiration: Oct 11, 2021, 6:19 PM
Renew [↗](#) | Unlink

NHGRI AnVIL Data Commons Framework Services
Username: boconnor@broadinstitute.org
Link Expiration: Oct 26, 2021, 6:20 PM
Renew [↗](#) | Unlink

Kids First DRC
Username: BRIANDCONNOR
Link Expiration: Oct 29, 2021, 11:58 AM
Renew [↗](#) | Unlink



*Simplify connecting
data source through
single, RAS
identity/authorization*

A single RAS-based account link

NIH Account [i](#) via RAS
Username: BRIANDCONNOR
Link Expiration: Oct 26, 2021, 6:16 PM
Renew [↗](#)

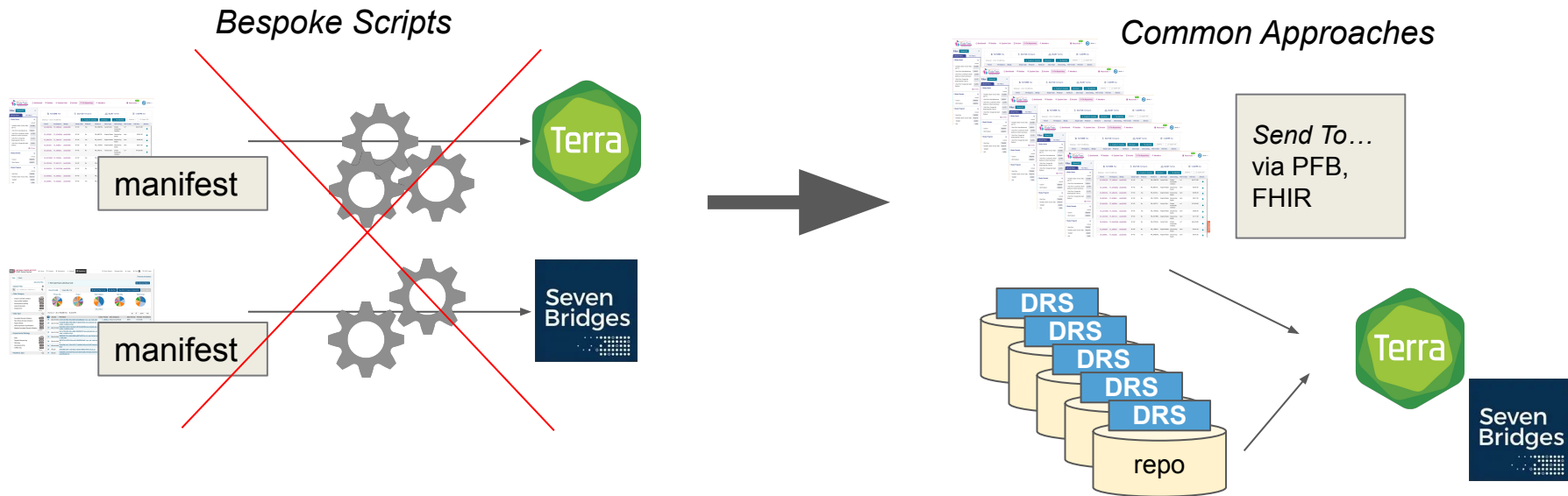
Resources

Authorized to access >

Not authorized > [i](#)



Connect more portals + data repositories

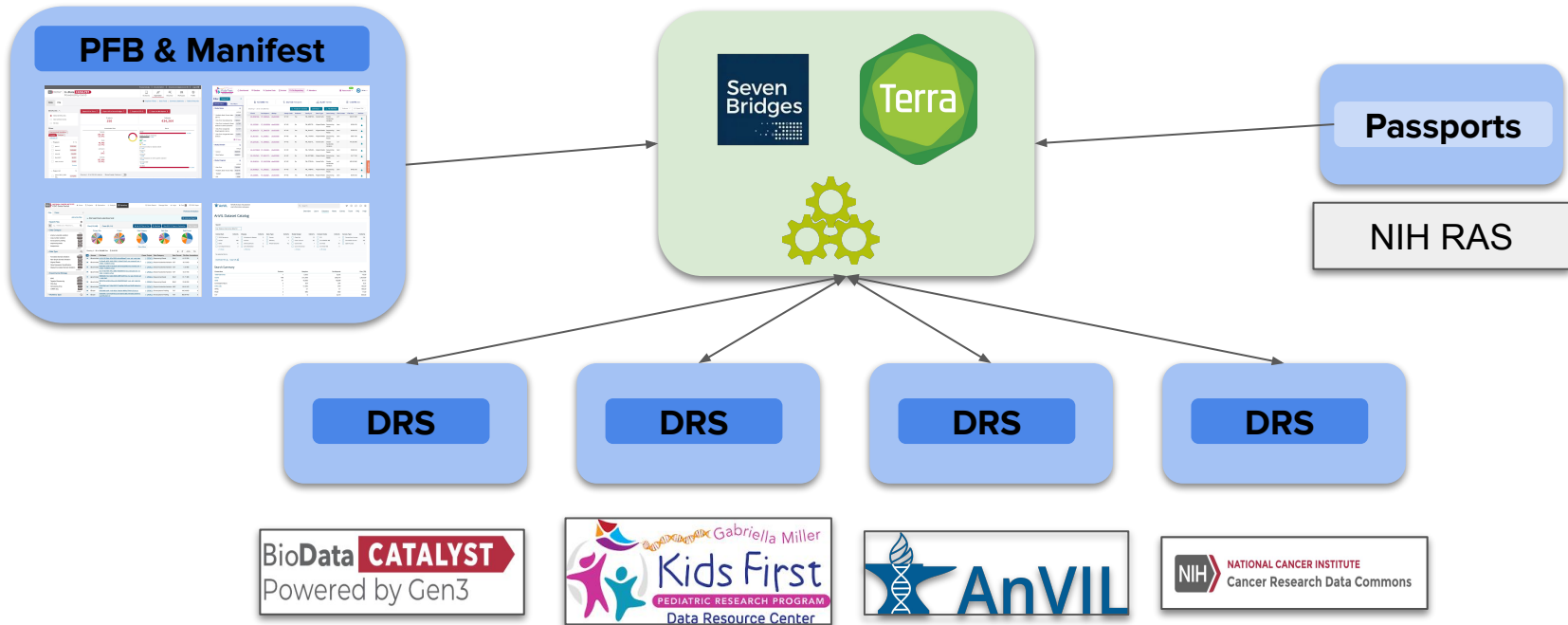


Users love the ability to "send to..." a workspace, add to more portals, make it easy to add new data repositories

Priorities for 2022

Tell Users!

Work with Outreach to let users know they can work with 11PB of data in these platforms today!





NCPI Systems Interoperation WG



Thank you to everyone that has made NCPI Systems Interoperation possible!!

Please consider joining our meetings, you can find more information at:

<https://anvilproject.org/ncpi>



Updates on Key Topics

Part 1

PFB, FHIR and RAS

Becky Boyles, Moderator



Reminder - What is PFB?



- The Portable Format for Biomedical Data (PFB) is a self-contained, self-describing, application independent **bulk format** for clinical, phenotype or other structured data.
- It is based upon **Avro**
- It encapsulates:
 - Data model / data dictionary
 - The bulk data itself
 - Pointers to third party controlled vocabularies for data elements
- It started based upon Gen3's graphical data model, but you can define PFB formats for **any data model**, relational, graph model, etc.
- For data versioning, support for multiple platforms and applications, long term support for data, it is helpful to have a self-contained bulk format



	Updates	Gaps/Next Steps
Gen3	<ul style="list-style-type: none"> • Gen3 is formulating a feature around functionality to allow external users to download study level PFBs via an API. The intent is to create a process to host PFB files that can be downloaded via DRS URIs externally existing data ingestion and submission would continue. An added ability to handoff PFBs would be available with this feature • Working closely with other groups to improve interop testing and Quality around PFB handoffs • Adding ability to export PFBs to 3 new destinations. These options will be provided in BioData Catalyst (export to export to CGC, CAVATICA, BDC powered by SBC) and in AnVIL (CGC and CAVATICA) supporting greater interoperability 	<ul style="list-style-type: none"> • Complete review of feature document and design for providing ability to download study level PFBs. Plan for work to implement. • Continue commitment to Quality by supporting interop testing for various test cases around PFB handoffs.
Seven Bridges	<p>Seven Bridges is developing interop solution to enable a user to send PFB from Gen3 systems outside of BioData Catalyst (like AnVIL) to BioData Catalyst Powered by Seven Bridges.</p>	<p>Pilot users to test the feature</p>



PFB - 2 (Grossman)



	Updates	Gaps/Next Steps
Terra	<ul style="list-style-type: none">• Terra currently supports PFB import from the AnVIL and BioData Catalyst data portals• Continued support of PFB as additional portals support the convention, currently working with the PDC portal	<ul style="list-style-type: none">• Adding automation for ETL process of PFB in FHIR• Adding automated transfer of PFB onto Healthcare API (FHIR)
Kids First	<p>FHAVRO: A generic Java library for converting FHIR resources into Avro and vice-versa. Avro schemas are obtained from project's FHIR implementation guide.</p> <ul style="list-style-type: none">• Enable developers to manage FHIR resources using the well-established Avro software ecosystem (e.g. Spark, Kafka)• Open source: https://github.com/Ferlab-Ste-Justine/fhavro Apache License 2.0• Current status: in active development	<ul style="list-style-type: none">• Generating Avro schema from a profile• Generating schema from NCPI implementation guide



	Updates	Gaps/Next Steps
NCBI	NCBI pioneered portable genomic data in 2011 with VDB, the foundation of SRA storage. It is a schema-driven columnar store with high compression, capable of representing any type of data, and organized into transportable units. The SRA uses these to model sequencing runs that was initially used across the INSDC.	Gap: Having provided APIs to access VDB, many tool vendors have not yet updated. Next Step: The VDB team is ready to guide tool vendors who are now willing to update in their adoption.
NCPI Outreach	Linking to documentation of PFB at https://anvilproject.org/ncpi/technologies	Keep PFB documentation up to date and expand as needed.

	Updates	Gaps/Next Steps
AnVIL	<p>FHIR Services</p> <ul style="list-style-type: none"> ● Proof of concept AnVIL FHIR server setup ● Currently access only for AnVIL dev team as development continues <p>FHIR Model</p> <ul style="list-style-type: none"> ● Implemented transform to NCPI Model for CMG data ● Developed and implemented pilot release of a Study and Summary level model. ● REDCap FHIR module has been updated to support export resources at minimum requirements. 	<ul style="list-style-type: none"> ● Wider release of FHIR server will need to wait until Terra picks up managed service for FHIR <ul style="list-style-type: none"> ● Team engaged with Terra engineering ● Project plan in place to create process to configure service and ensure authorization ● Continued onboarding of existing datasets ● Testing and refinement of Study and Summary level model
BDCatalyst	<ul style="list-style-type: none"> ● Loaded four test datasets using Bulk FHIR and built data ingestion pipelines to test Bulk FHIR standard in PIC-SURE ● Prototyped FHIR server deployments in both the Google and Azure clouds ● Prototyped conversion of synthetic HL7v2 and C-CDA documents into FHIR using Azure tools ● Prototyped FHIR to PFB export 	<ul style="list-style-type: none"> ● Continue to load appropriate data from FHIR sources in PIC-SURE ● Expand to use more real data sources and use cases (not test data) ● Longer term, ensure appropriate data is accessible via FHIR as determined by the BDCatalyst project.

	Updates	Gaps/Next Steps
Kids First	<p>1. NIH NCPI FHIR Implementation Guide: https://nih-ncpi.github.io/ncpi-fhir-ig/index.html</p> <p>2. Loaded five projects released on dbGap & KFDRRC:</p> <ul style="list-style-type: none"> ● Kids First: Enchondromatoses (SD_7NQ9151J): 285 Patients; 289 Specimens; 5,952 DocumentReferences ● Kids First: Congenital Heart Defects (SD_PREASA7S): 2,966 Patients; 2,987 Specimens; 16,506 DocumentReferences ● TARGET: Neuroblastoma (SD_YNSSAPHE): 277 Patients, 614 Specimens; 3,380 DocumentReferences ● Kids First: Familial Leukemia (SD_W0V965XZ): 620 Patients; 373 Specimens; 3,076 DocumentReferences ● Pediatric Brain Tumor Atlas - Children's Brain Tumor Tissue Consortium (SD_BHJXBDQK): 4,170 Patients; 48,240 Specimens; 43,004 DocumentReferences 	<ul style="list-style-type: none"> ● Replicating dbGaP's ResearchSubject model especially for curating various aggregate counts ● Developing a genomics module for sequencing and genomic workflow using Task and Observation ● Sustainable AuthN/AuthZ: The current AuthN/AuthZ flow requires an expiry ALB cookie and the acquisition of a cookie needs to be done manually. We therefore plan to implement OAuth2/OIDC setup supported via Keycloak. ● Exploring RAS-FHIR integration with Kurt R (UDN use case) ● Pedigree: Observation vs FamilyMemberHistory ● Phenotype: Condition vs Observation

	Updates	Gaps/Next Steps
<p>NCBI dbGaP API</p>	<p>Overview: 1800 Studies comprising approx. 3 million subjects, 370,000 variables and 2.5 billion observations.</p> <p>Study level meta-data: The NCBI dbGaP FHIR API provide access to all of dbGaP studies meta data. Users can search using multiple criteria including study title, sponsor, type (prospective, longitudinal, cohort, case-control), keyword, condition, and many others. https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/ResearchStudy</p> <p>Variable level data: An initial FHIR research database populated with synthetic data from a few representative studies is in place for the development team and some limited beta testers to better understand the technology and how to best represent dbGaP data as a collaboration with NLM Research Data Finder.</p> <p>RAS Integration: A working prototype of RAS access mechanisms is expected to be completed in Q1 FY22 for testing with dbGaP control-access consent group and to allow authorized users to reach de-identified research.</p>	<ul style="list-style-type: none"> • Current work is scaling up the servers and test loading more than 200 million observations. The ultimate goal is to provide seamless access to all of dbGaP metadata and phenotypic observations. • Continue development and testing to improve server performance. performance is a problem with big datasets such as dbGaP in native FHIR servers. • Integration of RAS will continue to be challenging due to constraints working with existing databases with different authorization systems • NCBI will continue to collaborate with LHC NLM to map and standardize the variable data. dbGaP variables have inconsistent and irregular labels that will require substantial effort to harmonize. • Continue with integration coordination with NLM Research Data Finder https://lhcfirms.nlm.nih.gov/fhir/research-data-finder/ and NCPI dataset catalog NCPI Data NCPI (anvilproject.org)

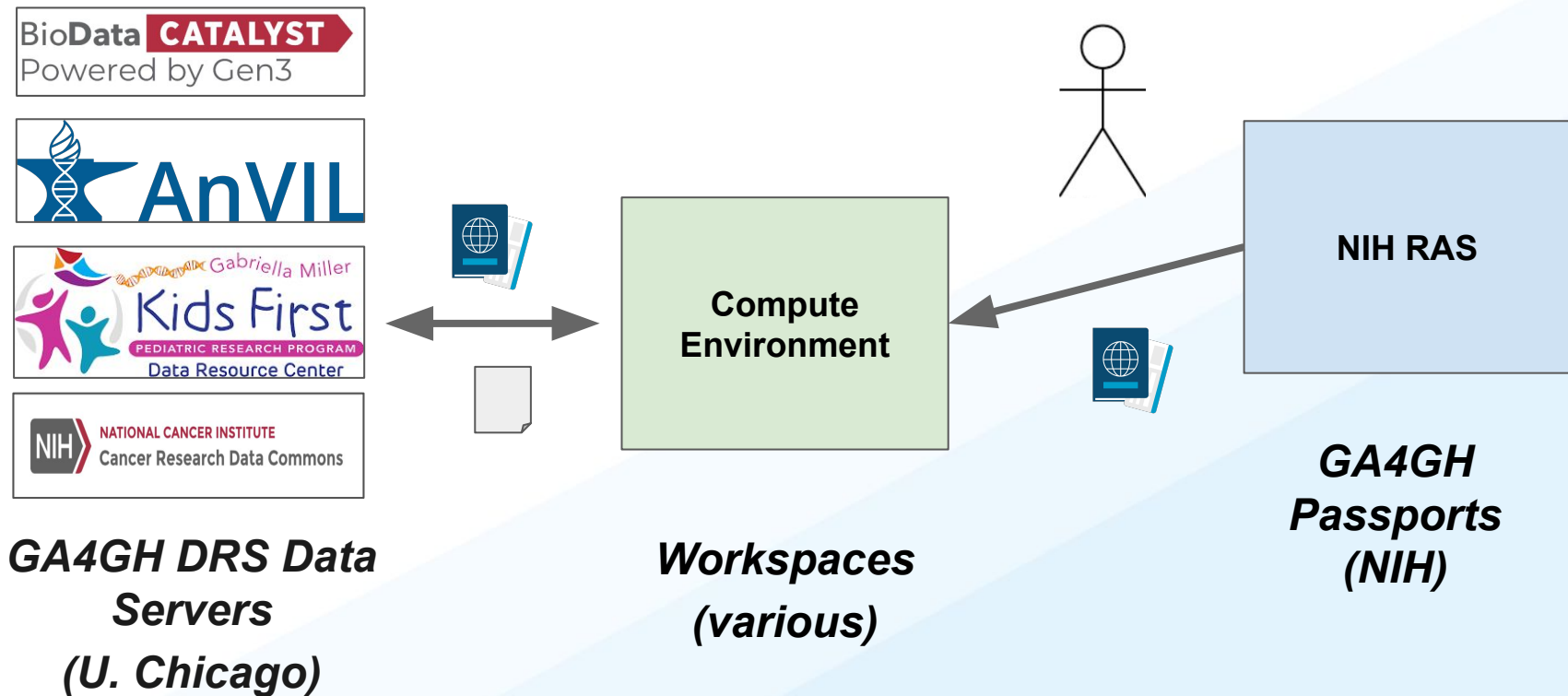


FHIR - 4 (Carroll)



	Updates	Gaps/Next Steps
NCPI Outreach	Linking to documentation of FHIR at https://anvilproject.org/ncpi/technologies	Keep FHIR documentation up to date and expand as needed.

RAS Update

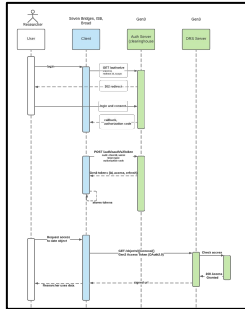


RAS Key Docs & Milestones

- **RAS design work** across a variety of teams and projects to date:
 - See: [RAS Authn/Authz "Milestone 3" Design with GA4GH Passports](#)
- Groups coordinated a 3 milestone plan:
 - **Milestone 1** : Login with RAS ✓
 - ~~**Milestone 2** : Gen3 uses RAS Visas as the authorization information instead of dbGaP telemetry files~~ Skipping this
 - **Milestone 3** : RAS Passport Visas can be used directly to access data resources, Central Fence is enabled by consistency across IC stacks
→ *designed in Q2-Q3 and now on an implementation timeline*

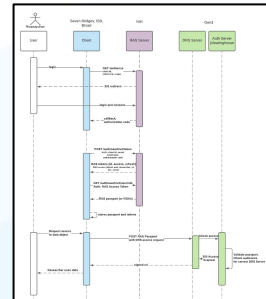
Summary of Milestone 3

- We've worked with Kids First, CRDC, [AnVIL](#) and [BDCat](#) to converge on a common approach for Milestone 3
- We've tried to help by putting together a [summary of two preferred approaches](#) and collaboratively address concerns... *goal is to add ability to access data with passports rather than taking away previous approach*



1: Current Gen3 Approach

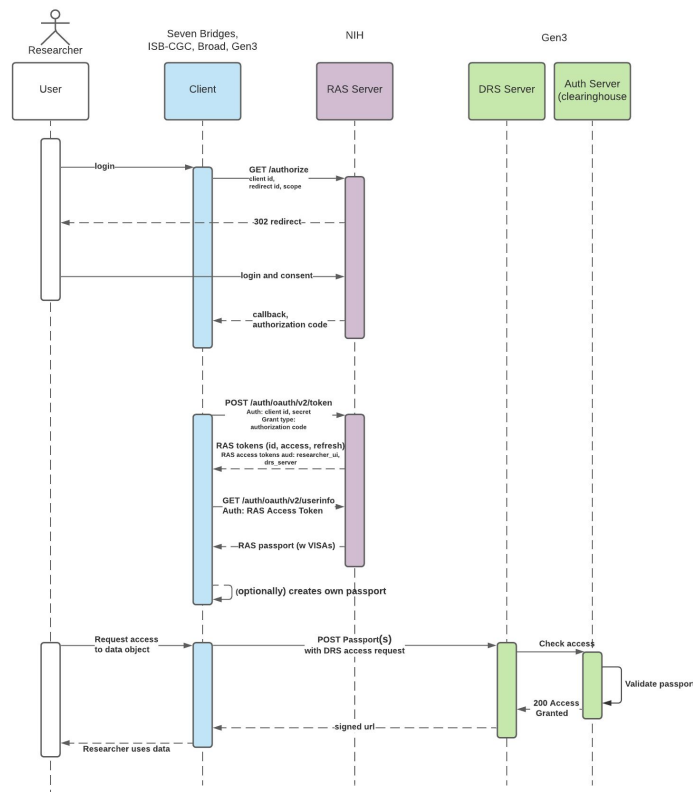
&



2: New Passport Approach

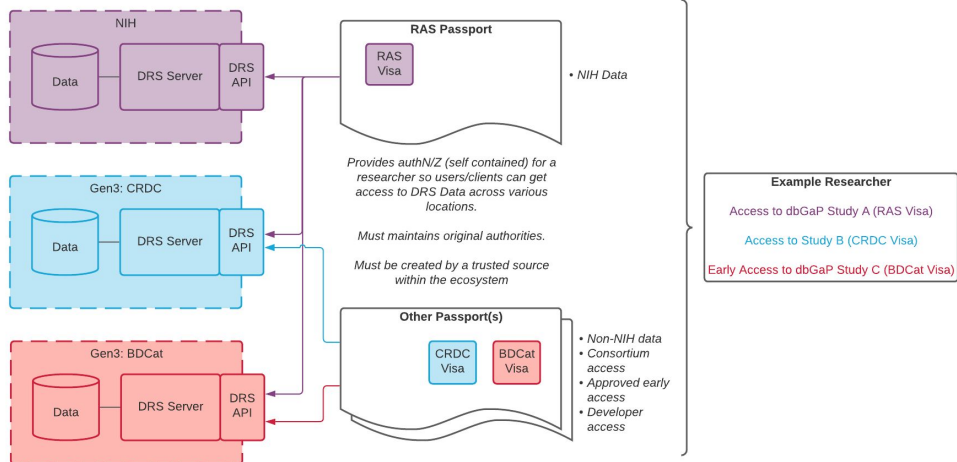
New Passports Approach

- Systems can interact with RAS directly, using RAS GA4GH Passports + Visas to access data from DRS data servers such as Gen3:
 - Client request RAS Passport directly from RAS
 - ~~Client repackages Passport while keeping RAS Visas intact~~
 - Passport is passed to DRS server in DRS data access request
 - DRS verifies and sends back a signed url
- Significantly improves the user experience for interoperability across NIH IC “stacks” by requiring just a single “account linking” with RAS (instead of multiple as is done today)
- RAS Passport + Visas then open up datasets that the users are authorized to use across systems like AnVIL, BDCat, CRDC, and GMKF as approved by researchers’ SO... *this is transformative for interop!!*



What We Learned Along the Way

- Use the Passport from RAS unmodified (don't repackage)
 - Other passport brokers may use repackaged passports for developer/consortium access lists but don't mix with RAS visas
 - **DRS 1.2** now supports sending multiple, complete passports in a DRS data access request



What We Learned Along the Way

- **Requirement for mutual TLS authentication for client verification**
 - *AnVIL, BioData Catalyst, CRDC, and GMKF indicated this is required*

Platform	Policy Requires Client Verification
AnVIL	Yes
BioData Catalyst	Yes
CRDC	Yes
GMKF	Yes

What We Learned Along the Way

- **Teams agreed to a timeline/plan**
 - *Implementation of staging/dev by U. Chicago before end of 2021*
 - *Workspace platforms implementing/testing by end of Q1 2022*
 - **See signatures of platform architects, POs, and security team members**

3.2*	<p>Use RAS V1.1+ Passports for Data Access at DRS Servers</p> <p>*This is a significant architecture change, including</p> <ul style="list-style-type: none"> • API Level Support for acceptance and validation of v1.1 passport(s) against DRS API as an alternate 	<ul style="list-style-type: none"> • Clients can POST the full RAS v1.1 Passport to get controlled-access data from a DRS endpoint • Gen3 DRS Server uses GA4GH claims clearinghouse to validate unmodified RAS passports and visas for authorization decisions 	See below for 3.2.1 and 3.2.3 target dates
	<p>means of authentication and authorization</p> <ul style="list-style-type: none"> • Parsing, validation, and interpretation of visas contained within v1.1 Passport(s) for means of realtime authorization upon data access requests • Caching support for scalability of average researcher workflows supporting thousands of data access requests in a short time frame • Final authorization decision by clearinghouse by aggregating information from parsing/interpretation of passport(s)/visa(s) and making a decision for controlled access data 		
3.2.1	<p>Minimum Viable Product with support for RAS V1.1 Passports in Gen3 DRS endpoints</p> <p>*does not include full integration tests nor performance support. <i>These are usually performed before</i></p>	<ul style="list-style-type: none"> • Stand up a development environment for clients to connect to, populated with mock control NIH data • An MVP deployment into respective development environment for clients of Gen3 to test respective flows (i.e. authorized users based on passport are returned a signed URL to data) 	Target Date: 12/07/21
	rolling to environments		
3.2.2	Load testing and profiling of 3.2.1 support	<ul style="list-style-type: none"> • Validate performance is comparable to current support via OIDC and OAuth 2 tokens 	Start date 12/1/21
3.2.3	Performance improvements based on results of 3.2.2 and subsequent load testing	<ul style="list-style-type: none"> • Performance improvements to ensure support is comparable to current support, as done via OIDC and OAuth 2 tokens 	Target Date 02/18/22
3.3	Mutual TLS Support as a mechanism for client authentication for controlled egress	<ul style="list-style-type: none"> • Client authentication so that systems know which client is presenting the RAS Passport to their DRS endpoint <p>*This is a system requirement for AnVIL and BDC *This support is not a RAS requirement but it is RAS recommended</p>	Target Date: 12/17/21

What We Learned Along the Way

- **U. Chicago has shared a detailed technical plan with the RAS and other teams**
 - *Milestone 2 not needed*
 - *Gen3 plan needed for building clearinghouse function in G3FS*
 - *"The RAS team does not need to review another version of this technical planning document."*
 - *We are ready to provide support on the clearinghouse design as needed. "*

Gen3 RAS Authorization/Authentication: Milestone 3 Requirements and Design

Version 1.0 (2021-09-17)

Table of Contents

Overview	1
Purpose and Scope	2
Key Capabilities	2
Technical Requirements and Design	4
RAS Milestone 3.1: Use RAS v1.1 passports for user authorization, instead of RAS v1.0	4
RAS Milestone 3.2 Use RAS V1.1 Passports for Data Access at DRS Servers	7
API Design	10
RAS Milestone 3.2.4	11
RAS Milestone 3.4 Support to create G3FS passports V1.1 and visas for custom data access	11
Timelines	14



	Updates	Gaps/Next Steps
Gen3	<ul style="list-style-type: none">• Full implementation of RAS for AuthN• Continued discussion of milestones 3 to enable full AuthZ using RAS passports directly• Required SIA completed by U Chicago and signed off by CBIIT security	<ul style="list-style-type: none">• Full consensus on plans for milestone 3 to enable use of RAS passports• Implementation of tasks established in the milestone 3 document
Seven Bridges	<ul style="list-style-type: none">• Seven Bridges working to get RAS passports directly from RAS as part of UDN/NCBI/SRA use case.• Full implementation of RAS for AuthN• ISAs signed with all relevant systems• Approval of current RAS milestone 3 plans	<ul style="list-style-type: none">• Upstream implementation of RAS for AuthZ and use of RAS passports in all systems• Rapid and robust SOP for support of end users



	Updates	Gaps/Next Steps
Terra	Terra is code complete on a new service the External Credentials Manager (ECM) which can obtain a RAS-issued Passport from RAS using the RAS v1.1 passport specification and can monitor to identify expiring visas and request updated visas. This is currently on Terra's non-production environment and will not be deployed or utilized until Gen3's DRS work is complete and Terra has completed the work for sending a RAS Passport to a Gen3 DRS server for data access.	Terra is currently planning the development work for sending a RAS Passport to a Gen3 DRS server for data access.
PIC-SURE	Compatibility with Gen3 RAS based authentication, still using Gen3 based authorization.	Leveraging RAS Passports for Authorization once available.



	Updates	Gaps/Next Steps
NCBI	NCBI has released a RAS Clearinghouse v1 service that processes RAS 1.1 passport tokens. This service is used by the dbGaP DRS v1.0+ service, with extended features for processing passports that have now been incorporated into DRS v1.2. Online dbGaP genomic data stored in the SRA can be reached by POSTing a RAS passport and DRS id to the DRS service.	<ul style="list-style-type: none">• Minor adjustments to bring dbGaP DRS in line with v1.2• Pilot passport v1.2 in support of FHIR• Provide externally accessible pre-release support for developers
NCPI Outreach	Linking to documentation of RAS at https://anvilproject.org/ncpi/technologies	Keep RAS documentation up to date and expand as needed.

Lunch Break 1:00-1:45pm ET
(and RAS Breakout 1:15-1:45pm ET)



Breakout 1 Instructions (Patton)



We will open the **RAS** breakout room in this same Zoom. This breakout will last from 1:15-1:45 p.m. ET

- If you have downloaded the latest version of Zoom ([instructions](#) and [how-to video](#)), you can move yourself into your preferred room.
- Otherwise, request that meeting host move you to your room.

We will have breakout rooms for other key topics this afternoon.

[Breakout Report Backs](#) will be first on the agenda for Day 2.

We will reconvene in the plenary session at 1:45 p.m. ET.

Updates on Key Topics

Part 2

End-User Cloud Costs, Search and Other Interoperability Efforts

Becky Boyles, Moderator

End-User Cloud Costs - 1 (Schatz)

	Updates	Gaps/Next Steps
CRDC	<ul style="list-style-type: none">• Continued use of \$300 pilot credits for new users• Continued use of benchmarking data for published tools• File archiving on AWS added to Seven Bridges for users to save on storage costs• FireCloud has expanded its tools for cloud cost estimation with in-app cost reporting. Users can now see cost incurred on a per submission and per workflow basis	<ul style="list-style-type: none">• Continued documentation on AWS and Google costs, breakdown of costs for each analysis per NCI Cloud Resource• New tutorial coming soon describing how to estimate cloud costs
AnVIL	<ul style="list-style-type: none">• AnVIL Cloud Credits (AC2) Program initiated• Offering \$300 in cloud credits• Developed a cloud cost budget justification spreadsheet & template• Improved cloud cost calculations released on AnVIL/Terra• Started a project to empirically measure the costs of popular genomics tools: talk	<ul style="list-style-type: none">• Continuation and possible expansion of the AnVIL Cloud Credits (AC2) Program• Develop a technical report on empirical cloud costs by Summer 2022

End-User Cloud Costs - 2 (Schatz)

	Updates	Gaps/Next Steps
BDCatalyst	<ul style="list-style-type: none">• Offering \$500 cloud credits through the NHLBI Cloud Credits program with an opportunity to request more funds for Heart, Lung & Blood research.• September Cloud Costs Community Hour: notes, slides, and Youtube videos available• Published documentation on managing team costs, estimating workflow costs, and setting up spend alerts on Terra and estimating total cloud costs on SBG• Launch of Project Per WorkSpace (PPWS) on Terra to support improved cost reporting functionality	<ul style="list-style-type: none">• STRIDES enhancing Dashboard to provide users with more visibility and timeliness into cloud credits available and spent.• Seven Bridges investigating adding file archival options on Google Cloud.• Identify and evaluate solutions for BYO cloud credits
Kids First	<ul style="list-style-type: none">• All new users receive \$100 in pilot credits to explore CAVATICA (Seven Bridges).• Training materials related to costs are available through the Kids First DRC Help Center and Cavatica's Support Documents.• Wrote a report assessing the successes and lessons learned of a 2.5 year long pilot cloud credits program	<ul style="list-style-type: none">• Finalize new guidelines and training materials to launch a public Cloud Credits program for Kids First users based on our own pilot's conclusions and recommendations of other NCPI platforms.

End-User Cloud Costs - 3 (Schatz)

	Updates	Gaps/Next Steps
NCBI	<ul style="list-style-type: none">• No egress costs for SRA datasets stored in AWS Open Data (ODP) buckets and GCP Public Dataset Program• No egress costs to access SRA data on AWS or GCP from within the respective cloud compute environments, if running from the correct regions; compute in the cloud is at user expense• NCBI's Cloud Data Delivery Service provides free “thaw” from cold storage and delivery to users’ buckets of SRA data in cold storage classes on AWS and GCP; per-user limits apply• Example user costs can be found here	<ul style="list-style-type: none">• All public SRA data is in AWS ODP, but more controlled access (dbGaP) sequence data is coming



	Updates	Gaps/Next Steps
CRDC	<ul style="list-style-type: none">• Seven Bridges Cancer Genomics Cloud (CGC) complete UI for integration of Cancer Data Service (CDS) datasets now available• Ongoing work to harmonize metadata and identifier standards across the CRDC to better enable search• Cancer Data Aggregator (CDA) Release 1 launch to enable query of Genomic Data Commons (GDC) and Proteomic Data Commons (PDC) open access data• UAT testing of CDA Release 1• Collaborated with Center for Cancer Data Harmonization (CCDH) team on end-to-end workflow demonstration of CRDC interoperability use cases	<ul style="list-style-type: none">• Integration of Seven Bridges CGC and the CDA via Jupyter notebook• Continued efforts at harmonizing metadata and identifier standards across CRDC• CDA Release 2 launch scheduled to connect GDC, PDC and Imaging Data Commons (IDC) open access data• Obtain Authority to Operate (ATO) to publicly launch CDA API and enable controlled-access data query• Integrate with Cloud Resources (cloud analysis platforms) that provide CDA front end interface for UAT

Search - 2 (Rogers)

	Updates	Gaps/Next Steps
BDCatalyst	<ul style="list-style-type: none">• Open Access search for phenotypic (PIC-SURE) and genomic data (Seven Bridges) prior to authorization• File and variable level search (phenotypic and genomic data) from User Interface on PIC-SURE• File and variable level search (phenotypic and genomic data) from PIC-SURE API on Terra and Seven Bridges• Semantic, public, full text, variable granularity search of TOPMed phenotypic concepts and dbGaP studies with explainable results, provenance in biomedical knowledge graphs, links to peer reviewed literature, and preliminary harmonization. (Dug)• Open Access Study-Level Search prior to Login through Gen3 Discovery Page• Subject, Study, and File Level Search (w/ secure limiting of results prior to authorization) through Gen3 Exploration Page• Exposed search API's for metadata, file object records, and genomic/phenotype data (GraphQL)	<ul style="list-style-type: none">• Development of search use cases and personas• Development of integrated search strategy• Interoperability of handoff of search results to analysis workspaces across ecosystems - finding key use cases to drive development

Search - 3 (Rogers)

	Updates	Gaps/Next Steps
Kids First	<ul style="list-style-type: none">• All clinical, phenotypic, demographic, file data searchable [both faceted and text] in registered-tier Portal that anyone can access (if they agree to click-through terms). Key non-Kids First datasets are also searchable in the Portal (interoperability with TARGET, CBTN etc). Filters applied to dynamic visualizations to build cohorts of multiple datasets and ability to identify children affected with multiple conditions (e.g., cancer & birth defects).• All source data (as provided by submitters) made searchable in addition to “harmonized” HPO, MONDO, and NCIt terms• Variant Database enables search of variants with annotations from ClinVar, TOPMed, Gnomad and the ability to identify which datasets include that variant and aggregated phenotypes.• All studies are searchable in dbGaP and grouped in an umbrella BioProject Study• (see slide about FHIR)	<ul style="list-style-type: none">• Back-end API is migrating from custom data service to a FHIR-based data service for interoperability with Portal, CAVATICA, and other tools.• 6 out of 22 Kids First studies loaded into the Variant Database, more to be loaded• Improvement to Variant Database in the Portal and the launching of the Variant Workbench (controlled access tool) which using table/matrix formats to find participants of interest and run analyses on a SPARK cluster.



Search - 4 (Rogers)



	Updates	Gaps/Next Steps
AnVIL	<ul style="list-style-type: none">• AnVILproject.org displays the dataset catalog• Dataset Catalog - Newly added detail page for each study populated via dbGaP FHIR API and Terra.• Dataset Catalog - Deep link from the study page to dbGap “Request Access” page preserving the study context.• Gen3 - Subject, Study, and File Level Search (w/ secure limiting of results prior to authorization) through Gen3 Exploration Page• Gen3 - Exposed search API’s for metadata, file object records, and genomic/phenotype data (GraphQL)	<ul style="list-style-type: none">• UX Research / Dataset Catalog UI updates to make the study detail pages more informative and useful.

Search - 5 (Rogers)

	Updates	Gaps/Next Steps
NCBI	<p>Studies and Metadata: Web / SOLR faceted search for: https://www.ncbi.nlm.nih.gov/gap/advanced_search/</p> <p>FHIR Research Study resource: https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/ResearchStudy</p> <p>Sequence Runs: Sequence Read Archive: "controlled"[Access] - SRA - NCBI (nih.gov)</p>	<p>Gap: Coordinated Sequence catalog/API format</p> <p>Next steps: Linking access request system to search interfaces.</p> <p>Adding metadata and phenotypic data as RAS enabled FHIR API</p>
NCPI-portal	<ul style="list-style-type: none">● Held a “Mini Hackathon” for Dataset Catalog API integration AnVIL, BDC, CRDC db GaP studies all refreshed automatically via APIs.● Newly added study descriptions for each dataset with data from dbGaP● Deep link to dbGap “Request Access” from each study.	<ul style="list-style-type: none">● Refresh KF dbGaP studies via API● Read non dbGaP studies via API.● UX research / Incremental UI updates.● Explore deeper integration with platform APIs and search engines.● POC of integration with Dug semantic search.

Other Interoperability Efforts - 1 (Ahalt)

	Updates	Gaps/Next Steps
CRDC	<ul style="list-style-type: none">• DRS Client added to CGC (now has both Server and Client)• Push button connection between CGC, BDC, and Cavatica utilizing DRS endpoints• Ability to connect to any open DRS endpoint or add known DRS endpoints	<ul style="list-style-type: none">• Broad FireCloud using DRS to integrate Proteomics Data Commons datasets
AnVIL	<ul style="list-style-type: none">• Forward looking work at workflow interoperability• Forward looking work at utilizing generic "app" definitions for extending the AnVIL platform (e.g. expanding the Leo service that powers Galaxy integration)	<ul style="list-style-type: none">• Continue to collaborate with BDCatalyst and other NCPI teams on the development of the "app" interface and extension of the Leo component to support it
BDCatalyst	<ul style="list-style-type: none">• Imaging: POC Nifti ingestion workflow• New co-leads of Tools & Apps WG, proposed tiered approach for establishing criteria to support V3PAs• Established Tool Trust Tiger Team (T4) to address data protections and workflow credibility standards• Ongoing discussion between PIC-SURE and AnVIL	<ul style="list-style-type: none">• Identification of use cases to drive next interop efforts• Test RAS using incoming SRA data (PCGC)• Continued exploration in imaging access and analysis, eg ability to support new image formats

Other Interoperability Efforts - 2 (Ahalt)

	Updates	Gaps/Next Steps
Kids First	<p>Active cross-platform use cases include:</p> <ul style="list-style-type: none">● CFDE (Kids First and HubMap: running common workflows, integrated knowledge graph; multiple DCCs: develop FHIR profiles for CFDE human datasets; exploring CAVATICA use)● INCLUDE (Data Hub launching in March) - interop on genomic data of children with DS & leukemia and CHD● CARING for Children with COVID: share pediatric COVID clinical data through FHIR API for ImmPort and BioData Catalyst to interoperate with. <p>Additional use cases of interest:</p> <ul style="list-style-type: none">● FaceBase - craniofacial birth defects data, human and model - also other model organism databases● ABCD/HBCD (NDA) - pediatric genomics and imaging● RDCRN - exploring CAVATICA use	<p>New DATA Scholar starts 9/27, she will engage users, document use cases, propose, test and implement solutions, coordinate with NCPI and stakeholders</p>

Other Interoperability Efforts - 3 (Ahalt)

	Updates	Gaps/Next Steps
NCBI	<p>All dbGaP Approvals delivered to RAS</p> <p>dbGaP DRS in Public See: dbGaP DRS Documentation</p> <p>IDX service in Public See: SRA IDX Documentation</p> <p>FHIR Research Study API See: https://dbgap-api.ncbi.nlm.nih.gov/fhir/x1/ResearchStudy</p>	<p>dbGaP File Selector and SRA Run Selector configured for RAS Auth (in development)</p>

**Breakouts
PFB & FHIR,
Other Interoperability Efforts**



Breakout 2 Instructions (Patton)



We will open the **PFB/FHIR and and Other Interoperability Efforts** breakout rooms in this same Zoom. These breakouts will last until **3:10 p.m. ET**

- If you have downloaded the latest version of Zoom ([instructions](#) and [how-to video](#)), you can move yourself into your preferred room.
- Otherwise, request that meeting host move you to your room.

We will then have a brief break and open the breakout rooms for End-user Cloud Costs, Search, and Other Interoperability Efforts.

[Breakout Report Backs](#) will be first on the agenda for Day 2.

Plan for Day 2

Becky Boyles

Agenda: Day 2 All times ET (Ahalt)

Time	Activity	Owner	Links
11:00-11:10am	Welcome and Goals Day 2: Synthesize next steps, driving use cases, determine NIH/NCPI priorities	Stan Ahalt	Slides Notes
11:10-12:40pm	Breakout Report Backs and Discussion <ul style="list-style-type: none"> •PFB (10 min) (Grossman) •FHIR (10 min) (Carroll) •RAS (20 min) (O'Connor) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt) 	Moderator: Becky Boyles	Slides Notes
12:40-12:50pm	GA4GH Relationship	Brian O'Connor	Slides Notes
12:50-2:00pm	Lunch Break		
1:30pm-2:00pm	NIH Breakout: NIH Coordination Working Group Discussion of Priority Next Steps	NIH Only (via separate invitation)	
2:00-2:15pm	Use Case Overview: The Journey of a NCPI Use Case	Asiyah Lin	Slides Notes
2:15-3:20pm	Review of Current Scientific Use Cases	Moderator: Valentina Di Francesco	Slides Notes
2:15-2:30pm	Genetic Sex as a Biological Variable and X-inactivation	Melissa Wilson	Slides Notes
2:30-2:50pm	Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA	Valerie Cotton, Allison Heath	Slides Notes
2:50-3:05pm	Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems	Owen Hirschi	Slides Notes
3:05-3:20pm	Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra	Simran Makwana	Slides Notes
3:20-4:00pm	Synthesize Goals and Next Steps for the next 6 Months, with focus on driving use cases	Stan Ahalt, Jon Kaltman	Slides Notes



Meeting Deliverable: NCPI Glossary



- Remember to keep populating the NCPI Glossary with new words or additional definitions
- We hope this Glossary will be a concrete deliverable at the end of the meeting to help us coalesce around common definitions and/or highlight differences.

Breakouts

End-user Cloud Costs, Search



Breakout 3 Instructions (Patton)



We will open the **End-user Cloud Costs and Search** breakout rooms in this same Zoom. These breakouts will last until **4:00pm ET**

- If you have downloaded the latest version of Zoom ([instructions](#) and [how-to video](#)), you can move yourself into your preferred room.
- Otherwise, request that meeting host move you to your room.

[Breakout Report Backs](#) are on the agenda at 11:10am ET on Day 2.

Welcome to Day 2...

NIH Cloud Platforms Interoperability Fall 2021 Workshop

We'll be starting shortly!



Welcome and Goals Day 2:

**Synthesize next steps, driving use cases, determine
NIH/NCPI priorities**

Stan Ahalt

Virtual Meeting Roles (Patton)

Role	Purpose	Assignee & Slack	
Maestro: Mute Master, Raised-Hand Monitor, & Security	Master of Zoom Ceremonies. Contact Amanda for questions about Zoom issues, breakout rooms, or other general questions or if you notice suspicious activity.	@Amanda Miller (amiller@renci.org)	
Screen Sharing	Will share screen and advance slides.	@Julie Hayes	
Slide Content	Will update slide content throughout the meeting.	@Sarah Davis	
Moderator	Moderator listed for each agenda item. Moderator will prompt slide transitions during presentations and foster productive conversation during discussions.	Becky Boyles (@rboyles)	Stan Ahalt (@stan)
Plenary Notetakers	All are encouraged to add comments to the Homepage and Meeting Notes		
Q&A Monitor	Monitor questions in #oct_workshop Slack channel as well as Zoom Chat. Share Action Items, Decisions, and Outstanding Questions from Slack and Zoom to the Homepage and Meeting Notes	@Patrick Patton @Paul Kerr @Allie Gartland Gray	@Joe Asare @Tom Madden @John Cheadle
Time Watcher	Will try to keep us on time while still allowing room for important conversations.	@Sarah Davis	



Questions during the event? (Patton)



Verbal Questions: There will be time for questions throughout the meeting. If you want to verbally ask a question, use the Zoom feature to "raise your hand" and the host will enable your audio and then call on you to ask your question.

Zoom Chat: You can type questions via Zoom Chat throughout the meeting. Paul Kerr, Patrick Patton, Joe Asare, Allie Gartland-Gray, Tom Madden and John Cheadle will share questions from Slack and Zoom chat into the [Homepage and Meeting Notes](#).

Slack: Questions can be asked throughout the meeting by using the [#oct_workshop](#) Slack channel. We encourage anyone to write questions, comments, answers, or discussion in Slack at any time. If you have not received an invitation to [#oct_workshop](#), please email amiller@renci.org.



The latest version

Want the ability to move independently between breakout sessions?

We updated the meeting settings to allow attendees to move freely between the breakout rooms. **This setting requires the latest version of Zoom.**

- [Follow these instructions](#) or
- Watch this how-to video here: <https://youtu.be/E7zERcVLUBM>



BDCatalyst Statement of Conduct (Ahalt)



The BioData Catalyst Consortium is dedicated to **providing a harassment-free experience for everyone**, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, or religion (or lack thereof). We do not tolerate harassment of community members in any form. Sexual language and imagery is generally not appropriate for any venue, including meetings, presentations, or discussions.

BDCatalyst “Santa Cruz Rules of Engagement”:

- Do not shy away from identifying problems & risks
- Be candid
- Be heard
 - Identify an ally or motivate via Slack
 - Reach out to a Contact for particular topic(s) - Slack or email bdc3@renci.org if you don't know the Contact
- Be polite
 - Please use your full name on zoom. (* new addition! *)
 - If you are a “talker” remember to give others time/space to talk - if you are “quiet”, take advantage of any opening
 - Add your comments/ideas to notes if you don't find space to talk!

Agenda: Day 2 All times ET (Ahalt)

Time	Activity	Owner	Links
11:00-11:10am	Welcome and Goals Day 2: Synthesize next steps, driving use cases, determine NIH/NCPI priorities	Stan Ahalt	Slides Notes
11:10-12:40pm	Breakout Report Backs and Discussion <ul style="list-style-type: none"> •PFB (10 min) (Grossman) •FHIR (10 min) (Carroll) •RAS (20 min) (O'Connor) •End-User Cloud Costs (20 min) (Schatz) •Search (20 min) (Rogers) •Other Interoperability Efforts (10 min) (Ahalt) 	Moderator: Becky Boyles	Slides Notes
12:40-12:50pm	GA4GH Relationship	Brian O'Connor	Slides Notes
12:50-2:00pm	Lunch Break		
1:30pm-2:00pm	NIH Breakout: NIH Coordination Working Group Discussion of Priority Next Steps	NIH Only (via separate invitation)	
2:00-2:15pm	Use Case Overview: The Journey of a NCPI Use Case	Asiyah Lin	Slides Notes
2:15-3:20pm	Review of Current Scientific Use Cases	Moderator: Valentina Di Francesco	Slides Notes
2:15-2:30pm	Genetic Sex as a Biological Variable and X-inactivation	Melissa Wilson	Slides Notes
2:30-2:50pm	Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA	Valerie Cotton, Allison Heath	Slides Notes
2:50-3:05pm	Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems	Owen Hirschi	Slides Notes
3:05-3:20pm	Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra	Simran Makwana	Slides Notes
3:20-4:00pm	Synthesize Goals and Next Steps for the next 6 Months, with focus on driving use cases	Stan Ahalt, Jon Kaltman	Slides Notes



Next Steps, Next Steps, Next Steps

- What do we hear in the Breakout Report Backs and Use Case Updates that highlight or clarify what we need to do **next**?
- How do we distill those potential next steps to the **priority next steps**?



Meeting Deliverable: NCPI Glossary



- Remember to keep populating the NCPI Glossary with new words or additional definitions
- We hope this Glossary will be a concrete deliverable at the end of the meeting to help us coalesce around common definitions and/or highlight differences.

Breakout Report Backs

Becky Boyles, Moderator



What are gaps and/or key blockers for creating interoperability across platforms?

- Recall PFB supports different data models
 - With “PFB Light,” we have defined [some standard attributes](#) (13 attributes to define identify required BAM/CRAM files in a manifest) - solves a basic interoperability problem
 - Other PFB models used in NCPI to transfer clinical/phenotype data from Gen3 to a cloud platform
- Identifying next set of use cases for interoperability that includes **both data objects** (e.g. BAM/CRAM files) **and structured data** (e.g. clinical/phenotype data)
 - We have FHIR use case but only a use use for PFB Light



PFB Interop Trade-Offs



- Selecting user-define virtual cohorts in a portal, computes PFB on the fly (which can take time) **vs** also supporting precomputed PFB for predefined studies
- Agreeing to one data model for PFB **vs** supporting arbitrary models that must be parsed by the cloud platform that imports the PFB



PFB - Gaps and/or Key Blocks



- Confusion about what PFB is / is not
- Clarifying differences and similarities between PFB, VDB and other self-contained, self-describing encapsulation file formats and FHIR



PFB - Actionable Next Steps



What are actionable next steps to take in the next six months (including existing or potential driving use cases)?

- Document distinguishing PFB use cases
 - NCPI “Light PFB” for exchanging “manifest information” about research subjects in a cohort and associated BAM/CRAM files (13 fields)
 - Exporting and Importing full clinical/phenotype data and associated data model for a study
 - Exporting and Importing self-describing “AI/ML ready” datasets
- Demonstration of using precomputed PFB file containing data data for a research publication with full clinical/phenotype and DRS references to external BAM/CRAM files that is exchanged across two or more NCPI systems (will focus on AI/ML ready data)



What are gaps and/or key blockers for creating interoperability across platforms?

- Adoption across platforms
- Lack of clear documentation of uses of FHIR
- Need a map to communicate what the goals are and what the limits are

What are actionable next steps to take in the next six months (including existing or potential driving use cases)?

- Align on Research Study and metadata v1 representation
 - To help facilitate Portal and Search activities
- Develop and promulgate a set of milestones around services/use cases/limitations, and work with platforms to identify roadmaps for these opportunities

Quick FHIR use cases (Carroll)

FHIR includes a data model, vocabulary tools, and service layers (eg, REST API)

- Ingest of EHR data- [Federal Mandates for EHRs to support FHIR](#)
- Ingest of other data, e.g. with [REDCap module](#) or [CDEs](#)
- Vocabulary tools that support existing standards and custom or local definitions
- We can represent existing study data in a structured way “as is”
- We can represent study data in a robust, harmonized way to provide service guarantees to platforms and users
- Options for server implementations of a global standard we don't have to invent
 - [Google](#) (AnVIL?), [AWS](#), [Azure](#), [IBM*](#), [smile CDR](#) (KF) / [HAPI*](#), [firely](#), and more (* open source)
- Exchange data from disparate systems in a common way (even if content is not harmonized)
- Capacity to represent Study Summary and Study Metadata
- Capacity to reference external files, eg DRS URIs, with file metadata

What are gaps and/or key blockers for creating interoperability across platforms?
Specifically, we focused on **risks** for "milestone 3", the use of RAS for authorization:

- Timeline for 1) testing environment 2) production release
 - December for testing
 - End of Q1 2022 for workspaces and production DRS servers
- Architecture of services vs. implementation details
- Performance of RAS passports for data access with DRS
- Single sign on experience (maybe a longer term topic)



Beyond "milestone 3":

- Performance, batch operations, requester pays → updated DRS 1.3 and beyond
- Derived data authorization inheritance
- Securing other APIs (e.g. FHIR) with Passports
- Consortium users and repackaged Passports from non-RAS brokers for this purpose
- Working with other IAM systems and partners, international collaborations with groups like Elixir and standards groups like GA4GH

What are actionable next steps to take in the next six months (including existing or potential driving use cases)? ***A proposal:***

- Meet our "milestone 3" goals, top priority
- Begin planning "milestone 4"
 - Performance
 - Derived data
 - Securing other APIs (FHIR) with Passports
 - Consortium users and repackaged Passports
- Reach out to Passport partners beyond RAS
 - Working with other IAM systems and partners, international collaborations with groups like Elixir and standards groups like GA4GH
 - How would we access data from systems beyond those accessible with RAS Passports?



What are gaps and/or key blockers for creating interoperability across platforms?

- Cloud cost model is an enormous cultural shift
 - Institutional resources are “free”; anxiety over runaway costs; difficult to budget; complex payment
- Be mindful of both direct costs (e.g. storage, compute, egress) and overhead (e.g. admin, initialization)
 - “Free credits” are expensive; need to emphasize the advantages & make platforms easier to use
- A consumable model for analysis costs
 - Sequencing assays range from very routine (e.g. WGS w/ predictable protocols & costs) to highly experimental (e.g. 1st-gen Single Cell w/ very unpredictable protocols & costs)
 - Most NCPI computing now is highly experimental => Need to transition into a consumable model

What are actionable next steps to take in the next six months?

- Budget templates & guides; standardization language for grants endorsed by NCPI
- Draw out end-to-end user stories: upload, analysis, egress/distribution, maintenance, payment, accounts
- Aggregate cost modeling efforts across NCPI into a unified “database”
- Long term: Free tier for NCPI (Google Colab, AWS free tier); codeathon to optimize workflows; funding



Gaps	Next Steps
<p data-bbox="46 281 900 314">Understanding cross-platform personas and use cases.</p> <p data-bbox="46 358 521 390">Finding Studies (priority order)</p> <ul data-bbox="77 436 900 698" style="list-style-type: none"><li data-bbox="77 436 900 502">● Understanding how the data is consented and how to apply for access.<li data-bbox="77 513 521 546">● Searching over phenotype.<li data-bbox="77 556 676 589">● Searching by experimental metadata.<li data-bbox="77 600 656 633">● Searching by subject demographics.<li data-bbox="77 644 900 698">● Determining if a given genotype is present in a given dataset before having access. <p data-bbox="46 742 540 775">Building Cohorts (priority order)</p> <ul data-bbox="77 819 900 928" style="list-style-type: none"><li data-bbox="77 819 900 851">● Finding and gaining access to different search portals.<li data-bbox="77 862 900 928">● Portals lacking “send to workspace env’ buttons to easily take search results to analysis platforms. <p data-bbox="46 971 301 1004">Easy Retro Board</p>	<p data-bbox="981 281 1537 314">Form a search working group and ...</p> <ul data-bbox="973 358 1858 928" style="list-style-type: none"><li data-bbox="973 358 1858 467">● Conduct UX research to determine personas and use cases for search from actual users. Determine who how to source users e.g. BDC Fellows.<li data-bbox="973 478 1804 587">● Create a list of search components and APIs used in the NCPI platforms, demonstrate how to use, and collect feedback.<li data-bbox="973 598 1812 707">● Create a search taxonomy to define the different kinds of search used/envisioned to inform an integrated search roadmap.<li data-bbox="973 718 1761 784">● Link back to studies in context from the NCPI dataset catalog.<li data-bbox="973 794 1630 860">● Generate input for the upcoming search RFI NOT-OD-21-187.<li data-bbox="973 871 1325 904">● Explain data consents.<li data-bbox="973 915 1704 936">● Explore integrating FHIR into the search strategy.



Other Interoperability Efforts (Ahalt)



What are gaps and/or key blockers for creating interoperability across platforms?

- We need defining use cases from real-world researchers to help us identify the next steps for increased ecosystem Interoperability. Interestingly, there is a significant demand!
- Search across platforms is essential - and fortunately, we are making progress.

What are actionable next steps to take in the next six months (including existing or potential driving use cases)?

- Seek out real-world researchers and identify the next generation of users who want new Interoperability features.
- Look into the feasibility of standardizing how Tools/Apps are deployed across ecosystems to encourage portability.
- Develop methods for publishing completed use cases so that researchers can replicate them locally for training purposes / scientific verification. Include YouTube videos!
- Look for opportunities to create and deliver training on interoperable problems and methods.

GA4GH Relationship

Brian O'Connor

Mission: *Enable genomic data sharing for the benefit of human health*

The GA4GH is a policy-framing and **technical standards-setting** organization, seeking to enable responsible genomic data sharing within a human rights framework.

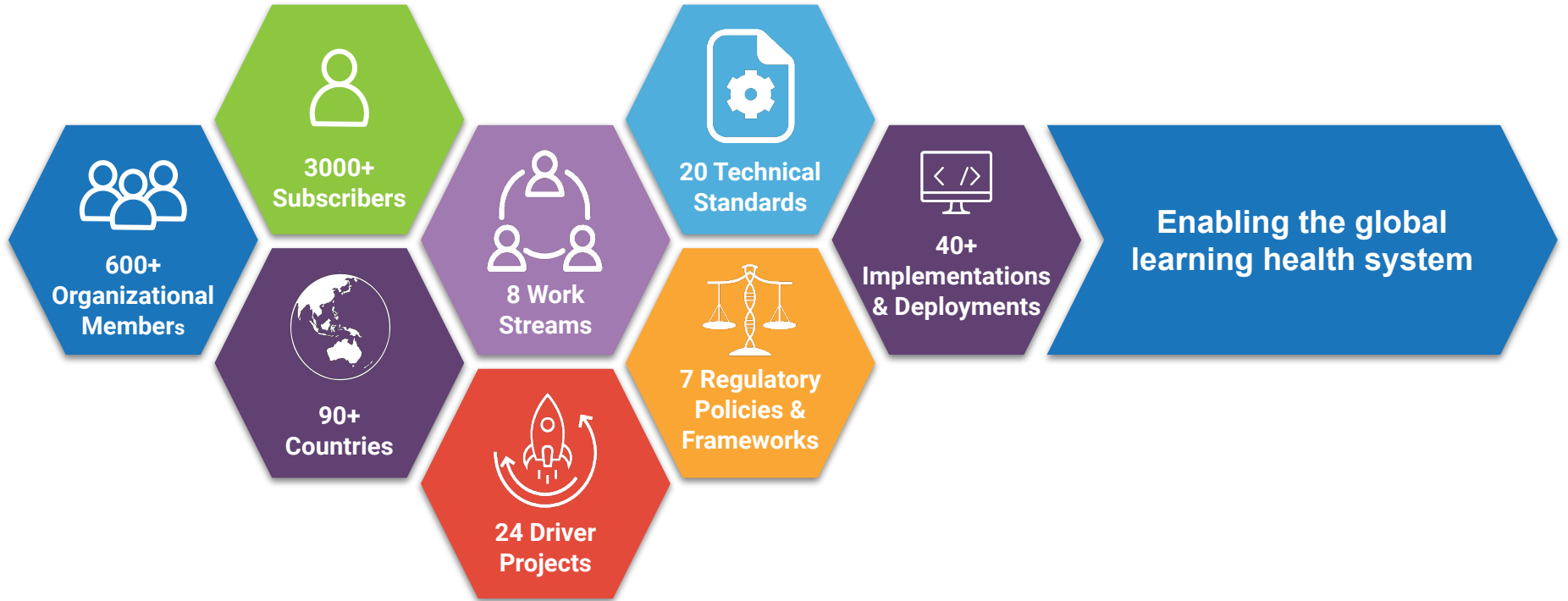


Global Alliance
for Genomics & Health

<https://ga4gh.org>



The GA4GH Ecosystem



The GA4GH Work Process



Work Streams

GA4GH Work Streams develop standards, tools, and frameworks that are designed to overcome technical and regulatory hurdles to international genomic data-sharing.

[VIEW WORK STREAMS](#)



Driver Projects

GA4GH Driver Projects are real-world genomic data initiatives sourced from around the globe that provide guidance on GA4GH standards development.

[VIEW DRIVER PROJECTS](#)



Technical Alignment Sub-Committee

The Technical Alignment Sub-Committee (TASC) provides mechanisms and recommendations to create internal consistency and technical alignment across GA4GH Work Streams and product deliverables. TASC serves as a central decision-making group, documenting and communicating these decisions across multiple stakeholders.

[LEARN MORE](#)



Partner Engagement

The GA4GH Partner Engagement initiative facilitates two-way dialogue with the international community, including national initiatives, major health care centres, and patient advocacy groups.

[CONTACT](#)

The GA4GH Work Process

		Real-World Driver Projects									
Technical Work Streams	Discovery	✓		✓		✓		✓			
	Large-Scale Genomics		✓		✓		✓		✓		
	Data Use & Researcher IDs	✓		✓		✓	✓				✓
	Cloud		✓	✓					✓		
	Genomic Knowledge Standards		✓				✓	✓	✓		
	Clinical & Phenotypic Data Capture	✓			✓	✓	✓				✓
Foundational Work Streams	Regulatory & Ethics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Data Security	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Partner Engagement

GA4GH Vision for Interoperability





The GA4GH Driver Projects



GA4GH Driver Projects are real-world genomic data initiatives that help guide our development efforts and pilot our tools. Stakeholders around the globe advocate, mandate, implement, and use our frameworks and standards in their local contexts.



**NATIONAL
CANCER
INSTITUTE**

**National Cancer Institute
Cancer Research Data
Commons (NCI CRDC)**



NATIONAL CANCER INSTITUTE
Genomic Data Commons

**National Cancer Institute
Genomic Data Commons (NCI
GDC)**



National Heart, Lung,
and Blood Institute

**Trans-Omics for Precision
Medicine (TOPMed)**

And many others...



GA4GH Standards Used by NCPI



- See the full collection at <https://www.ga4gh.org/genomic-data-toolkit/> and <https://www.ga4gh.org/genomic-data-toolkit/data-security-toolkit/>
- Passports and Authentication & Authorization Infrastructure (AAI)
- Data Repository Service (DRS)
- Tool Registry Service (TRS) (used by workspaces)
- Various file formats maintained by the GA4GH
 - CRAM
 - SAM/BAM
 - VCF/BCF
- *Others?*



New Opportunities with GA4GH



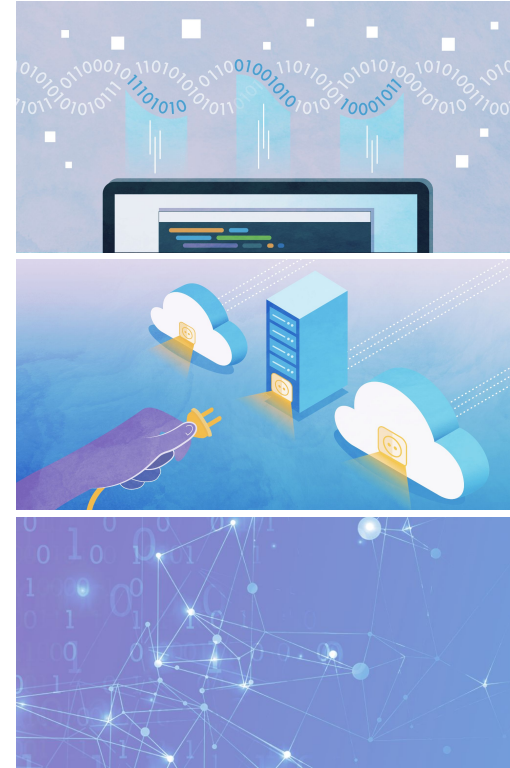
- *What are new opportunities for collaborating with GA4GH?*
- New API Possibilities
 - Data Connect → search API
 - Data Use Ontology (DUO) → describing data use restrictions
 - Phenopackets → relationship with FHIR for example
 - Task Execution Service (TES)/Workflow Execution Service (WES) → federated compute
 - Service Registry → advertise our services
 - *Explore the options [here...](#)*
- API adjacent and working groups
 - Starter Kit → trying out APIs
 - Technical Alignment Sub-Committee (TASC) → Building tooling for Work Streams
 - Federated Analysis Systems Project (FASP) → testing use cases with Drivers
- *Are there new standards we want to propose? E.g. PFB to Discovery?*



GA4GH Starter Kit



- Reference server implementation suite of GA4GH API specs (DRS & WES right now)
- Simplicity and versatility of setup
 - local laptop, HPC, cloud
- Technical on-ramp for:
 - Individuals new to GA4GH
 - Organizations exploring GA4GH on non-cloud native architectures
- Modular - Run APIs tailored to use case

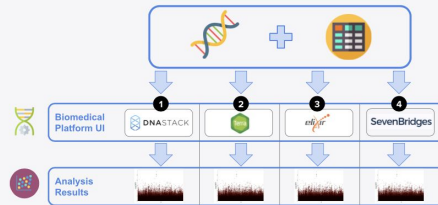


<https://bit.ly/starterkit-slides>

- GA4GH Federated Analysis Systems Project
Working with Driver Projects to demonstrate GA4GH standards
→ **great opportunity to collaborate on researcher use cases**
→ **we are already participating in this e.g. use case #7**

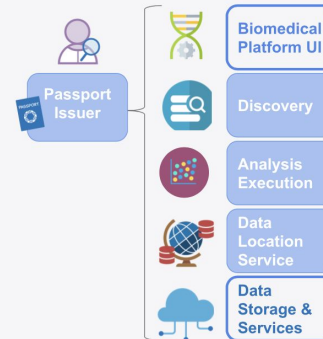
Horizontal Demo

Reliably produce the same research results
regardless of your choice of platform



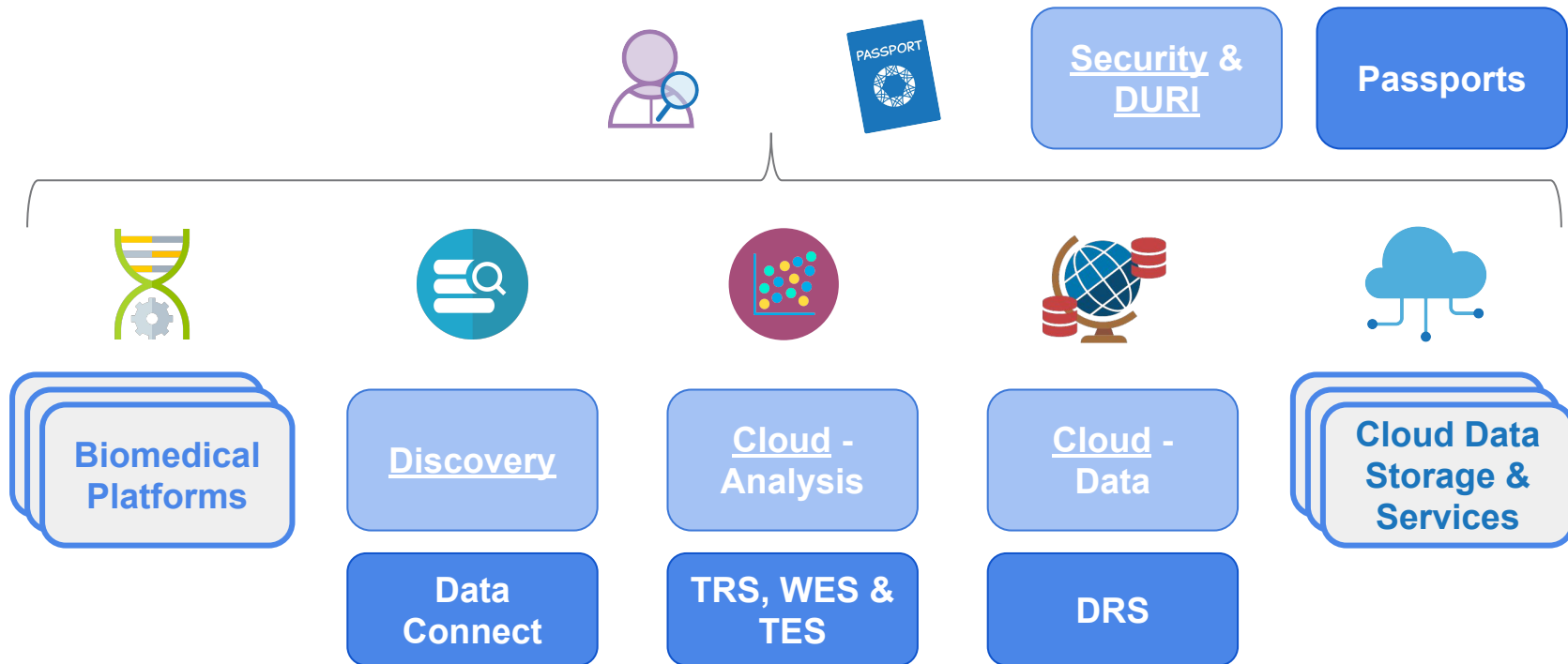
Vertical Demo

Deep integration of a comprehensive suite of
standards implemented by a single institution



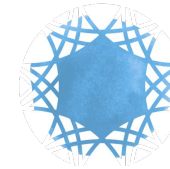
What Might this Look Like?

- See the full collection at <https://www.ga4gh.org/genomic-data-toolkit/>



Engagement Opportunities

- [GA4GH Connect](#) Oct 12-14, register [here](#)
 - Opportunities for collaboration across Work Streams and Driver Projects and for contributors to advance work on the GA4GH Strategic Roadmap
- [Genomics in Health Implementation Forum](#) (GHIF) Nov 16-17, register [here](#)
 - Genomics in Health Implementation Forum (GHIF) aims to support accurate data interpretation, diagnosis, and innovative solutions through global cooperation in data sharing and clinical implementation of genomics.
- FASP Regular Bi-Weekly [Meetings](#)
- GA4GH Equity, Diversity, and Inclusion (EDI) Advisory Group → info@ga4gh.org



GENOMICS IN HEALTH
IMPLEMENTATION FORUM

Lunch Break

12:50 p.m. - 2:00 p.m.

1:30 - 2:00 p.m.

Breakout, by invitation only:
NIH Coordination Working Group
Discussion of Priority Next Steps

Use Case Overview: The Journey of a NCPI Use Case

Asiyah Yu Lin

The Journey of a NCPI Use Case

From a seed to a forest

Asiyah Y. Lin

The 5th NCPI Workshop

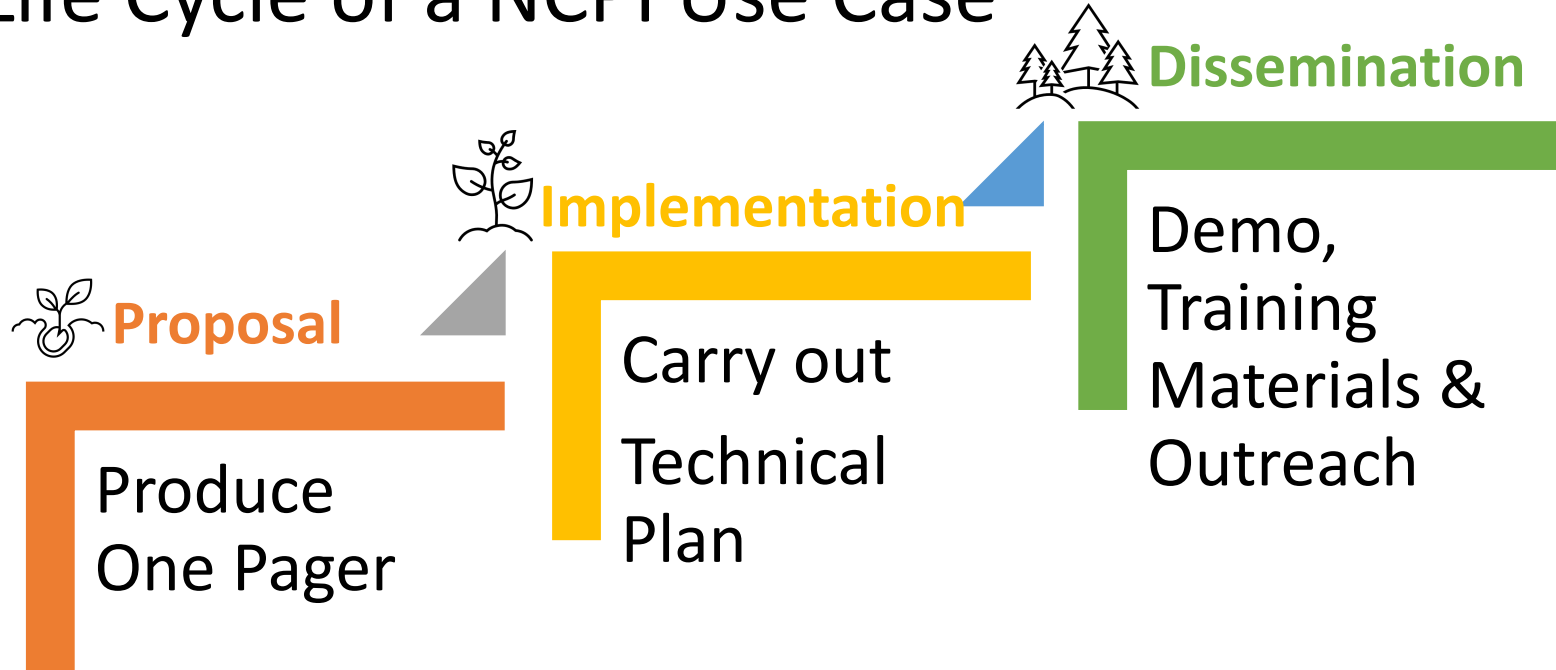
Oct 6, 2021

What is a NCPI Use Case?

- Access and integration of data from NCPI platforms ($n \geq 2$) is needed to answer a scientific question
- Interoperability demos
 - Search datasets from 2+ NCPI platforms
 - Access data from 2+ NCPI platforms for analysis in one workspace
 - Portable software and tools across 2+ NCPI platforms
 - More examples ...
- Ultimate goal: to drive the development of NCPI interoperability technology specification

*NCPI platforms: platforms support AnVIL, BDCat, CRDC, Kids First, NCBI.

Life Cycle of a NCPI Use Case



Proposal Phase



- NIH staff or a researcher identifies a potential scientific use case.
- In collaboration with NCPI WG Leads and platform PIs:
 - Identify scientific lead
 - Identify platform lead (a.k.a. interoperability tech lead)
 - Develop the interoperability plan and challenges
 - Identify funding resources
- Develop one pager.
- NIH Coordination WG keeps the one pager for documentation and management purposes.

Interoperability between Kids First/CAVATICA and SRA's copy of the Undiagnosed Diseases

NCPI Use Case Details

Status: NCBI actively moving all files (BAMs) to hot AWS/SRA storage. Files become immediately available as they are moved into S3. Next steps: Seven Bridges development work to obtain RAS and present them to NCBI/SRA DRS server to access files in CAVATICA workspaces.

Platform contact for genomic interop: Michele Mattioni and Kurt Rodarmer

Platform contact for FHIR structuring: TBD (one from dbGaP, one from Kids First)

Researcher contact: ~~TBD~~—assigned to Adam Resnick to resolve [Lisa Bastarache](#)

Dataset: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs00123
and ~~[insert Kids First datasets once determined, listed here:~~

~~<https://commonfund.nih.gov/kidsfirst/x01projects>]~~

NCPI use case link for genomic data interop: https://github.com/NIH-NCPI/NCPI_use_case

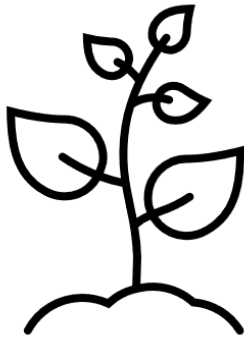
NCPI use case link for FHIR structuring of phenotypes: https://github.com/NIH-NCPI/NCPI_use_case_tracker/issues/18

Summary:

- **Goal:** enable co-analysis of Kids First data (BAMs/CRAMs + phenotype data) with U (and other) phenotype data) in Seven Bridges CAVATICA. This requires 1) search/finding the data

One pager
Example

Implementation Phase:



- Scientific and platform leads coordinate with the System Interoperability WG to carry out the technical plan.
- Scientific and platform leads are responsible for reporting implementation progress.
- Demos at the bi-annual workshop.
- Provide updates on the GitHub Use Case Tracker.
- May become inactive use cases if no progress is made.

Dissemination Phase:





















- A NCPI use case is completed with a demo of the implemented interoperability technical plan (**Note:** completion of the research plan is not necessary).
- Work with Outreach WG (Dave Rogers) to develop training materials:
 1. Training videos
 2. Necessary documentations
 3. Any publications (if relevant)
- Reach out and educate users to implement and grow the user community!

Training
video
example

Demo of Search Result Hand-off



<https://anvilproject.org/ncpi#demo-of-search-result-hand-off>

- 🕒 FHIR UC1: ResearchStudies representation in rare disease (CMGs & Kids First)**
#16 opened 22 days ago by cottonva  Needs One Pager
- 🕒 UC 13: Leverage functionally equivalent pipelines for long-reads data on different systems** **one pager done**   2
#15 opened on Jul 13 by jackDiGi  Ready to develop
- 🕒 UC 12 - (Xihong) Whole Genome Sequencing Association Analysis pipeline** **one pager done**  3
#12 opened on Jun 29 by NoopDog  On Hold
- 🕒 UC 11. (Wilson) Sex as a Biological Variable** **one pager done**  3
#11 opened on Jun 29 by NoopDog  Ready to develop
- 🕒 UC 10. SRA & Kids First DRC for Kids First & UDN co-analysis** **one pager done**   2
#10 opened on Jun 29 by NoopDog  Ready to develop
- 🕒 UC 9. Whole slide images** **need one pager**  2
#9 opened on Jun 29 by NoopDog  Ready to develop
- 🕒 UC 8. PIC-SURE API search of clinical and genomic data available from Seven Bridges Platform** **need one pager**  2
#8 opened on Jun 29 by NoopDog  Ready to develop
- 🕒 7. NHGRI AnVIL + Kids First DRC + NHLBI BioData Catalyst** **need training material**  4
#7 opened on Jun 29 by NoopDog  Use Case Complete
- 🕒 UC 1a. NHLBI BioData Catalyst + Kids First DRC** **inactive**
#2 opened on Jun 29 by NoopDog  On Hold

Questions and Suggestions are welcomed!

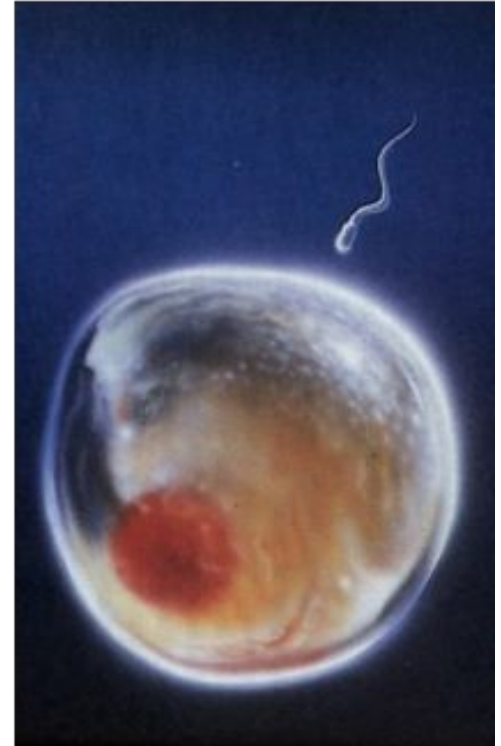
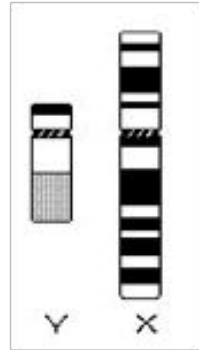
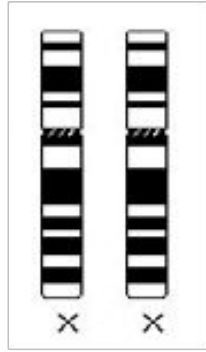
Team: Asiyah Lin, Dave Rogers, Jack DiGiovanna, Ken Wiley, Valerie Cotton, Valentina Di Francesco

Genetic Sex as a Biological Variable and X-inactivation

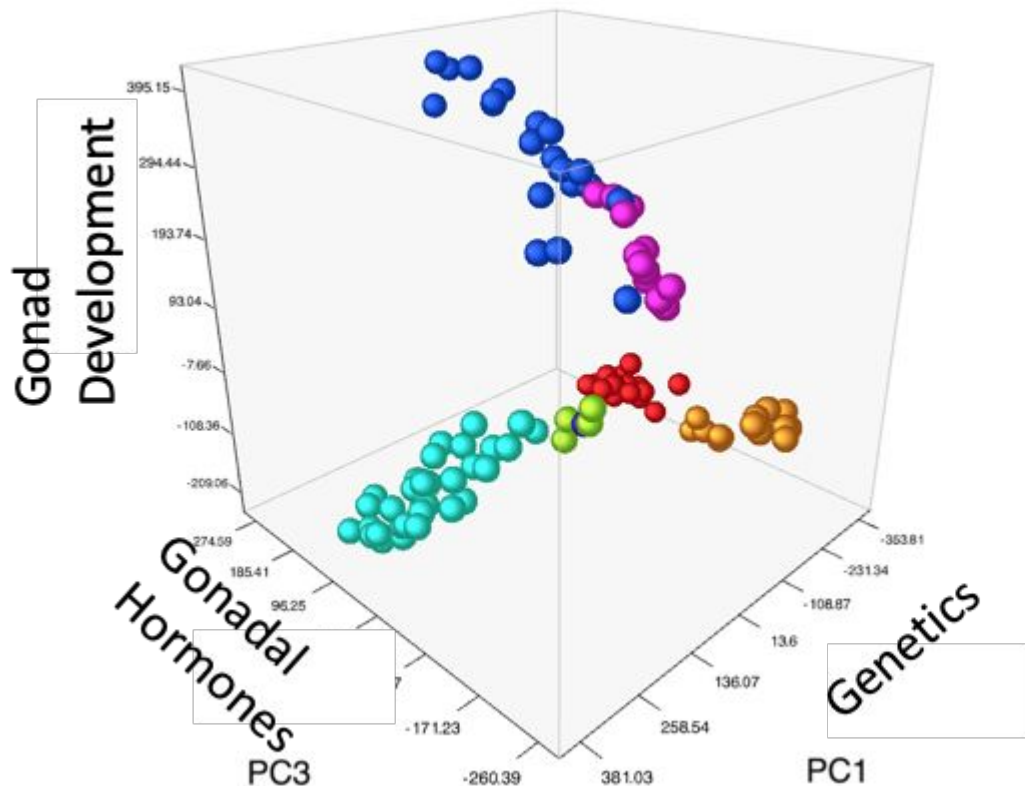
Melissa Wilson, PhD
Arizona State University



- **Genetics**
- **Gonads
(& gonadal hormones)**
- **Gender**



Sex differences are multidimensional



More than bimodal!

PERSPECTIVE | HUMAN GENOMICS

Searching for sex differences

Melissa A. Wilson

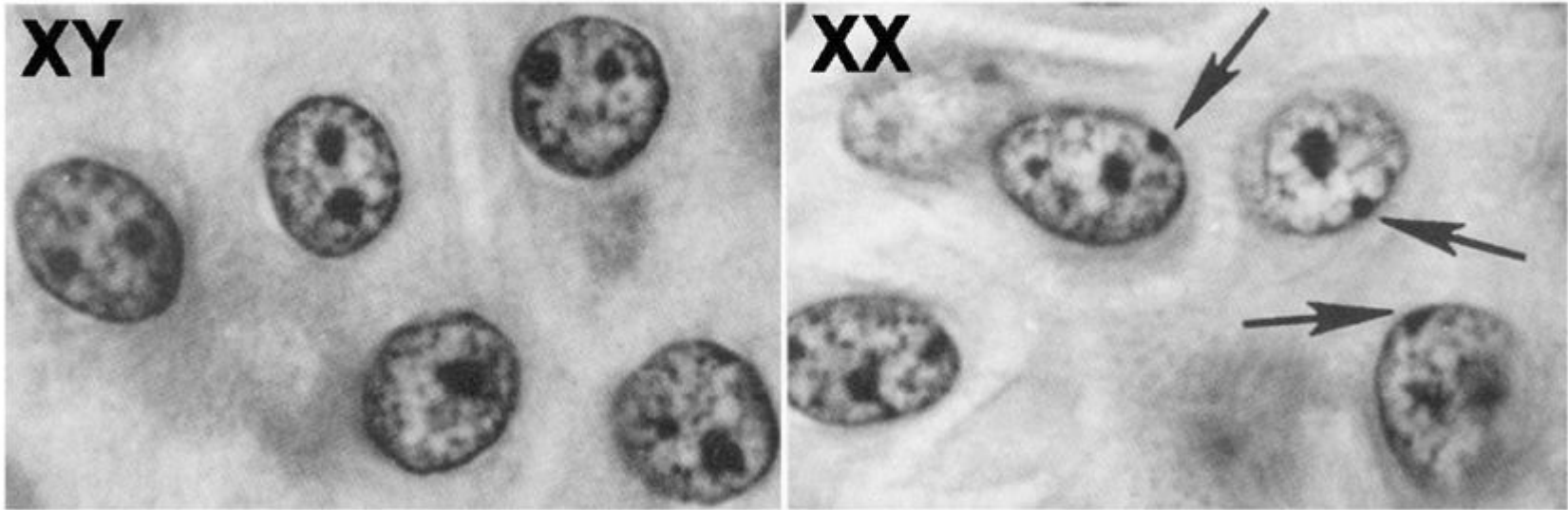
+ See all authors and affiliations

Science 11 Sep 2020:
Vol. 369, Issue 6509, pp. 1298-1299
DOI: 10.1126/science.abd8340

X-inactivation



Barr body as seen under the microscope

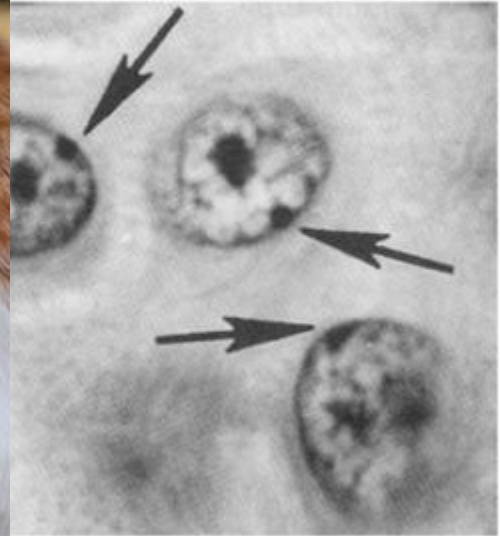




X-inactivation

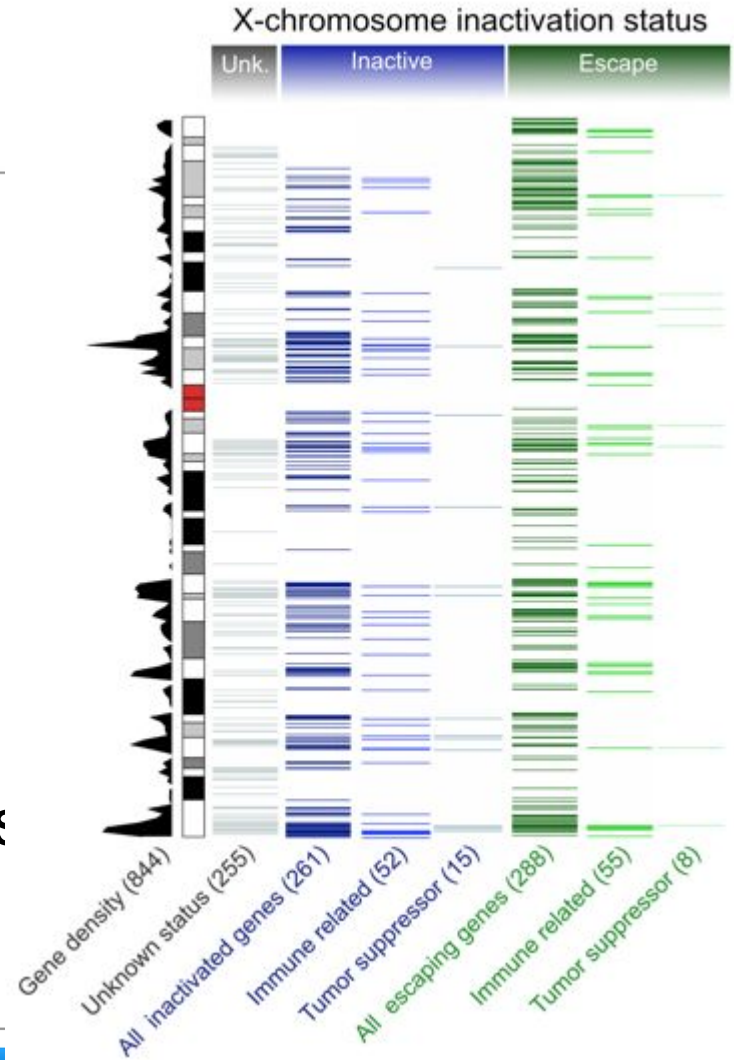


Barr body as seen un



Inactivation varies

- Approximately 1/3 of X-linked genes are **inactivated** in all individuals and tissues assayed thus far
- Approximately 1/3 of X-linked genes are **not inactivated** (**escape**) in at least some tissues and individuals



X-inactivation in the human placenta



Tanya Phung



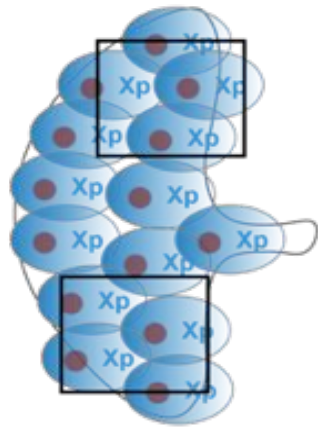
Kimberly Olney

(Phung et al, submitted)

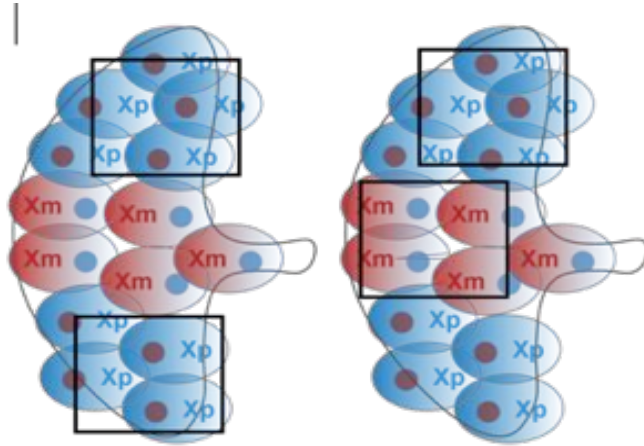


The placenta is
the genotype of
the offspring

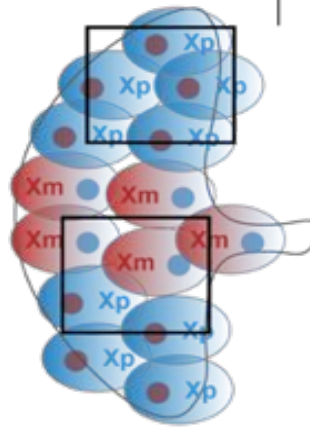
X-inactivation in the placenta



Skewing

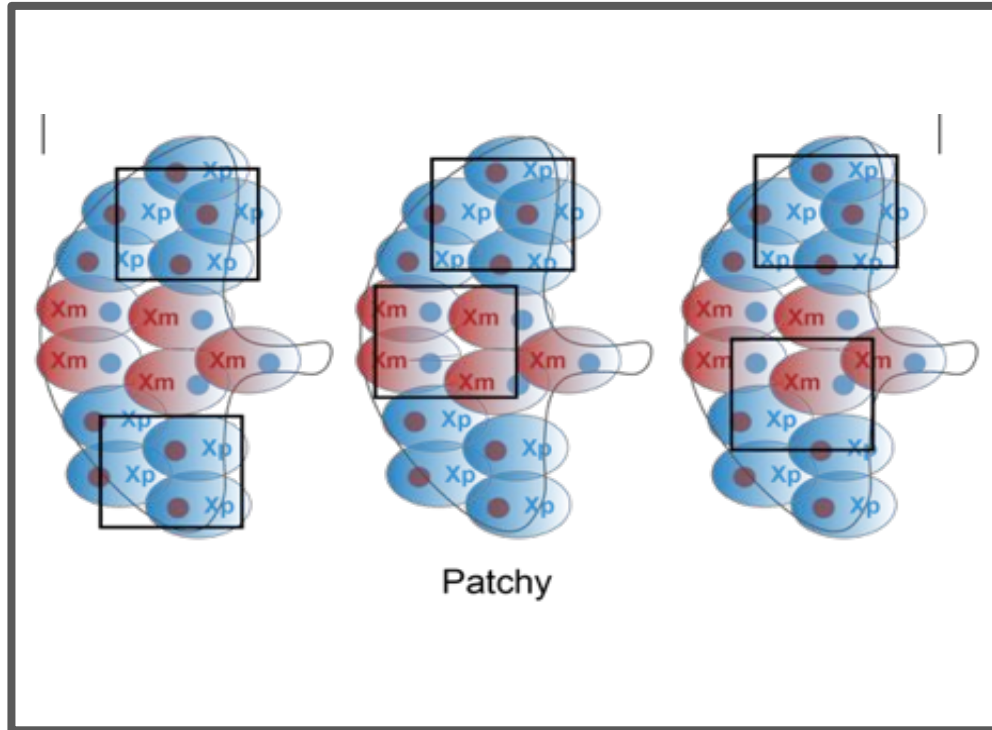


Patchy

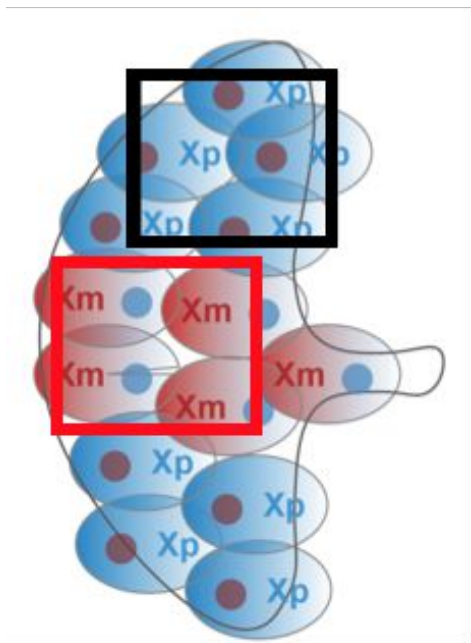


Mosaic

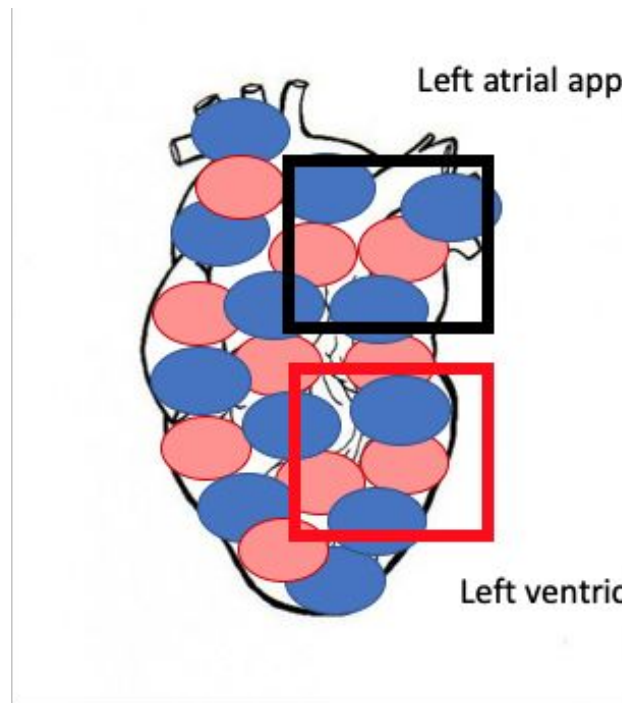
Patchy X-inactivation in the placenta



Placenta distinct from adult tissues



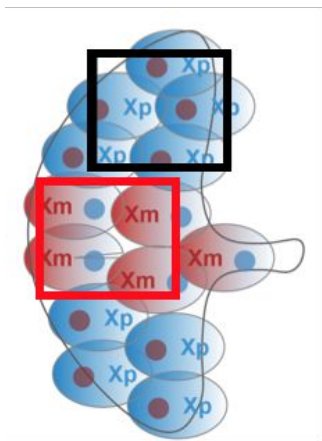
\neq



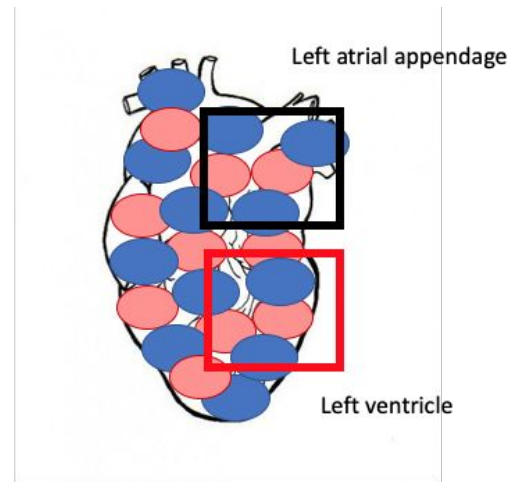
Heart data from GTEx consortium

X-inactivation across samples?

- Which genes escape
- Are these genes unique to a tissue, or to a condition
- Some genes escape only in T-cells and B-cells
- What is XCI across cancers? Different in pediatric or adult?



\neq



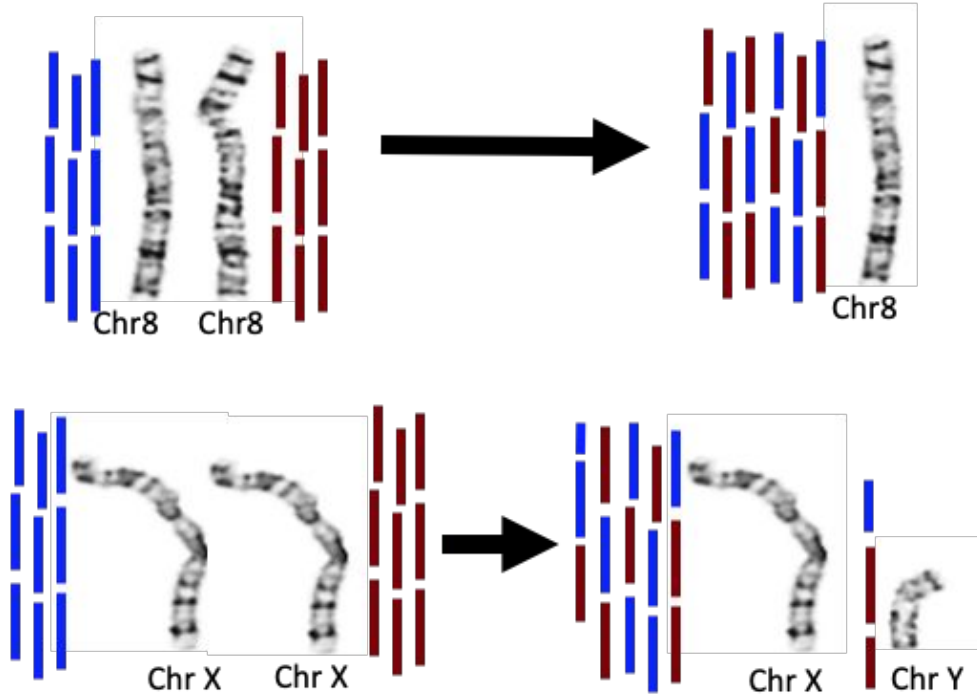
Heart data from GTEx consortium

Sex chromosomes are unique



Sex Chromosomes mis-mapping

46, X X Standard



Sex chr complement reference



github.com/SexChrLab/XYalign

Infer sex chromosome complement

Output in user-defined windows (all chr):

- Quality
- Depth
- Allele-balance

Realign with appropriate sex chr masks

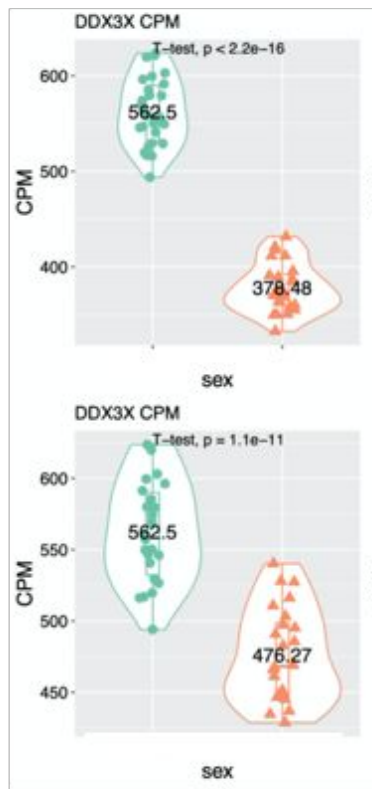


Timothy Webster

Mapping matters

Standard
Sex Chr Compl

DE in both

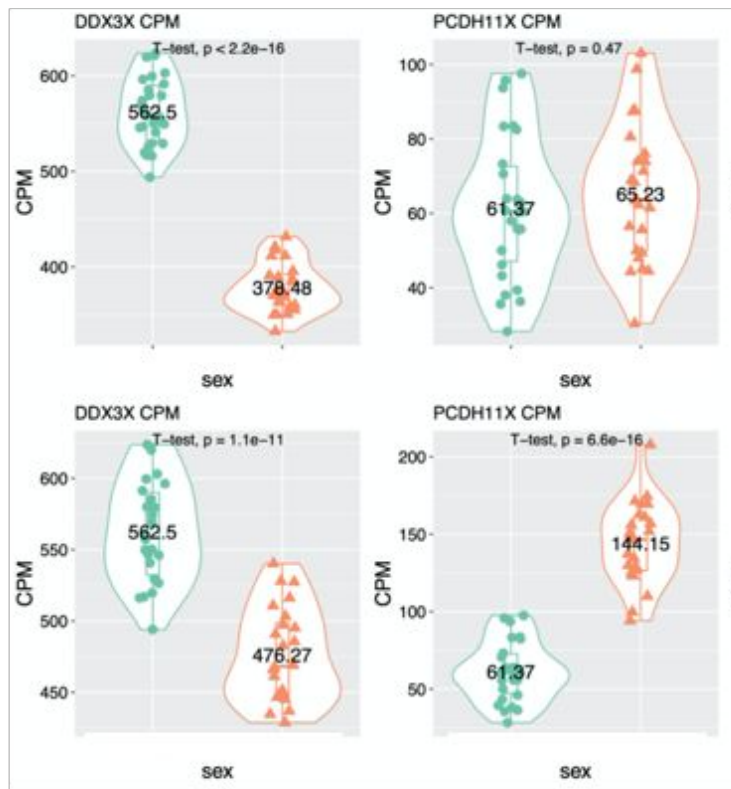


Mapping matters

Sex Chr Compl Standard

DE in both

No sex diff to DE



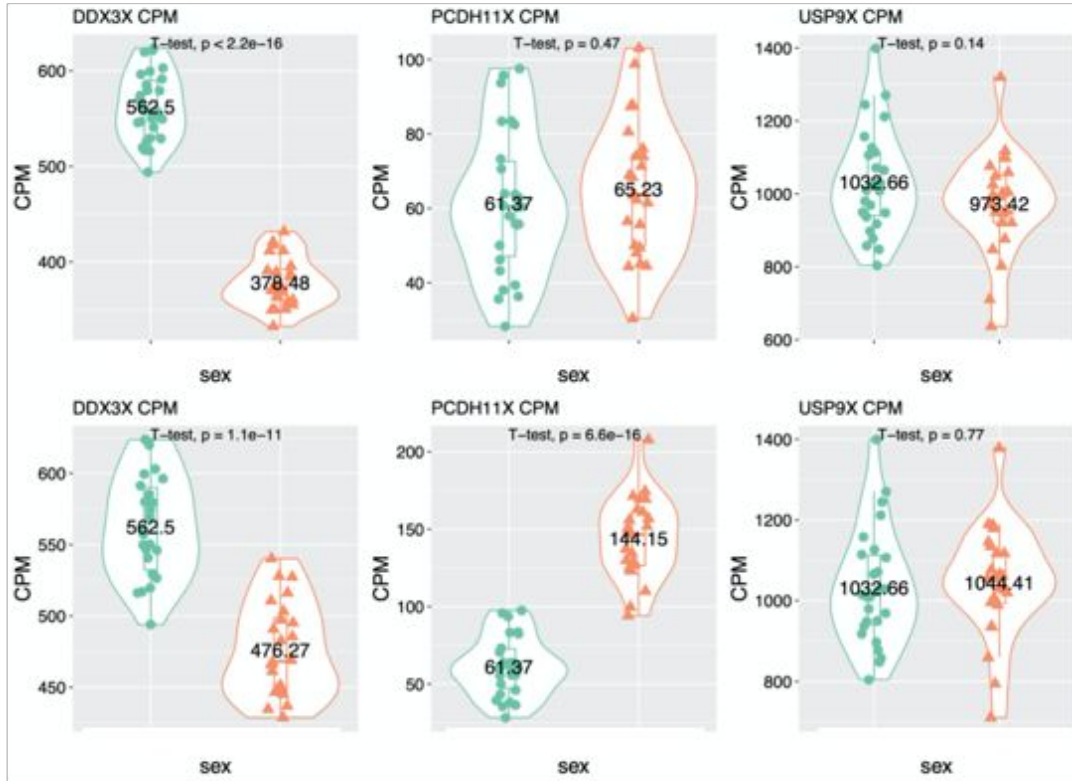
Mapping matters

Sex Chr Compl Standard

DE in both

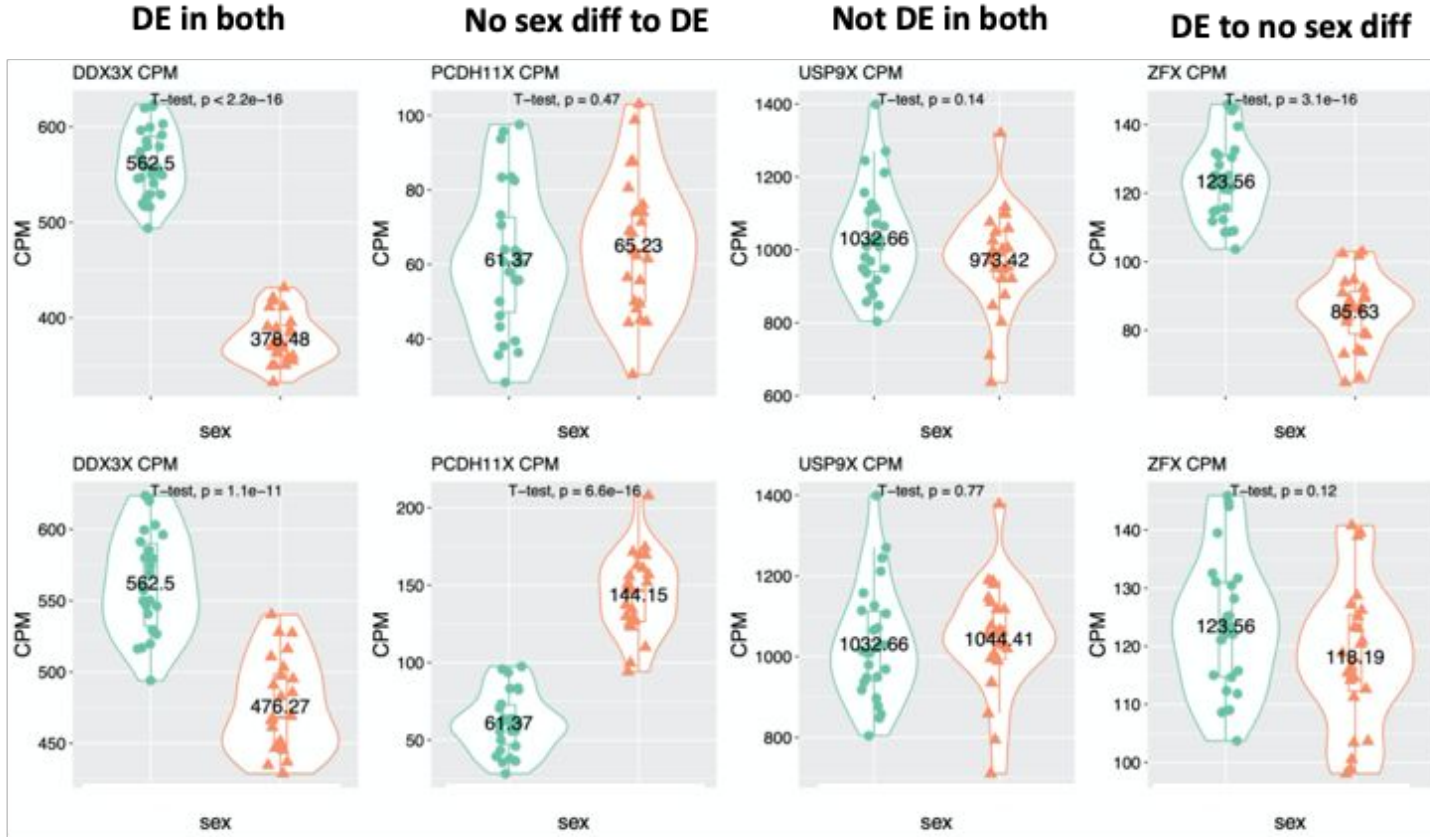
No sex diff to DE

Not DE in both



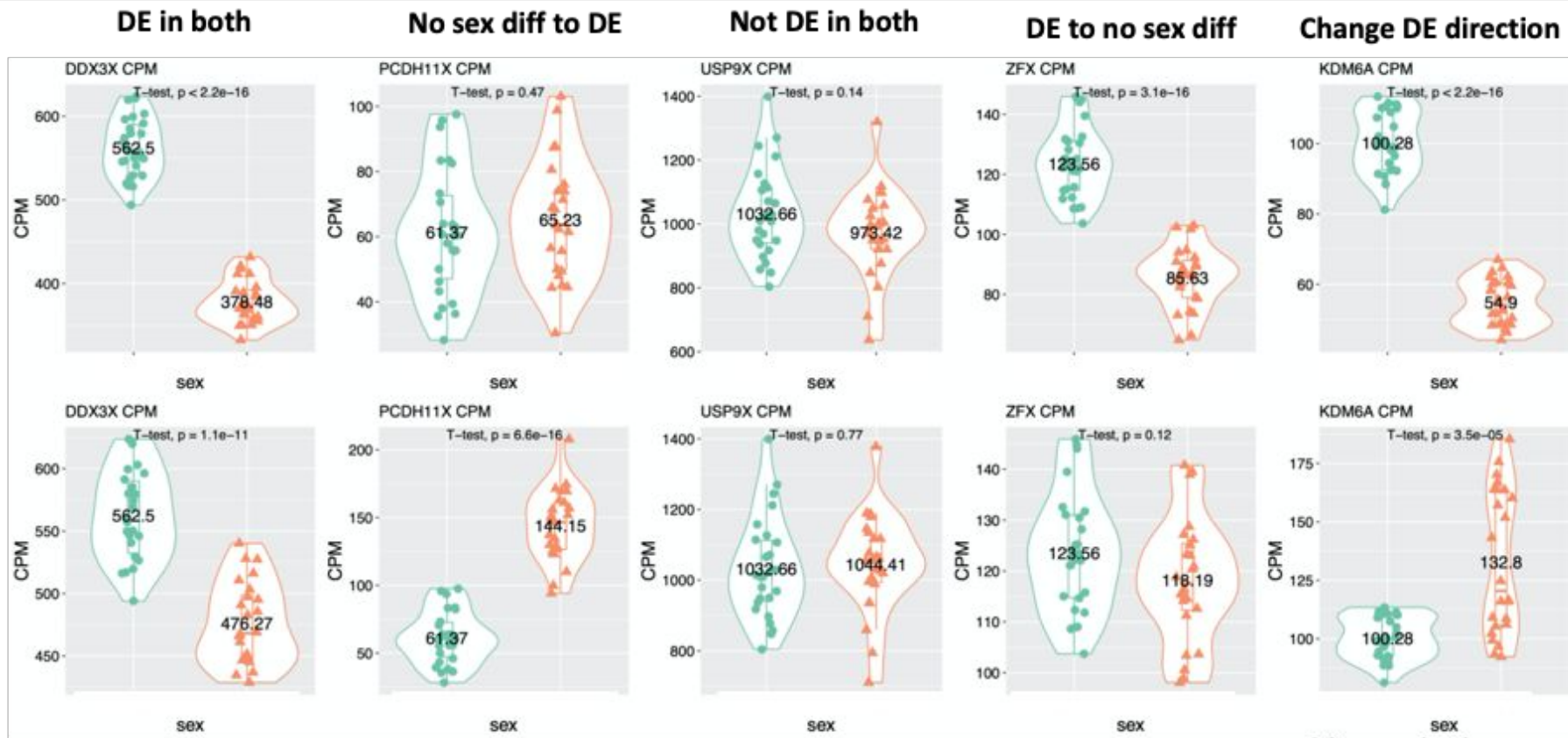
Mapping matters

Sex Chr Compl Standard

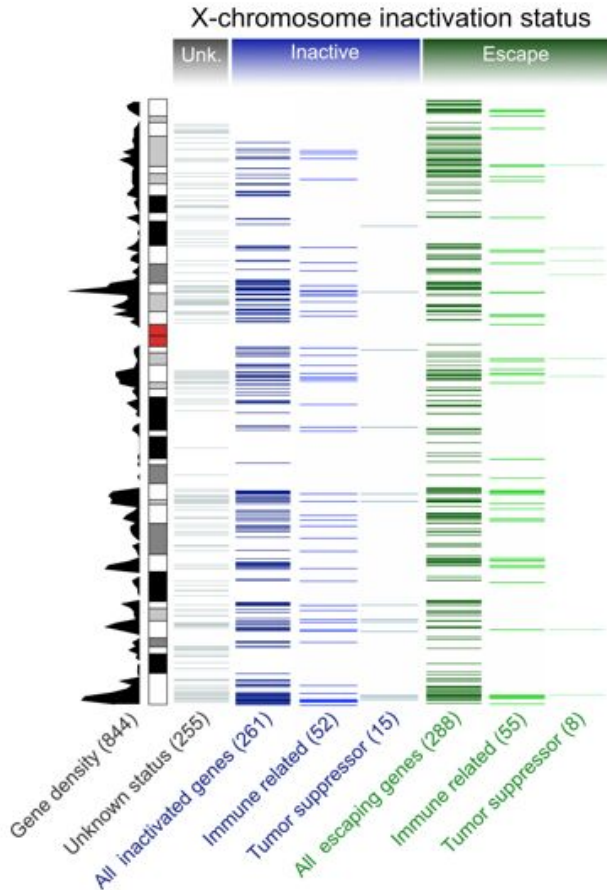


Mapping matters

Sex Chr Compl Standard

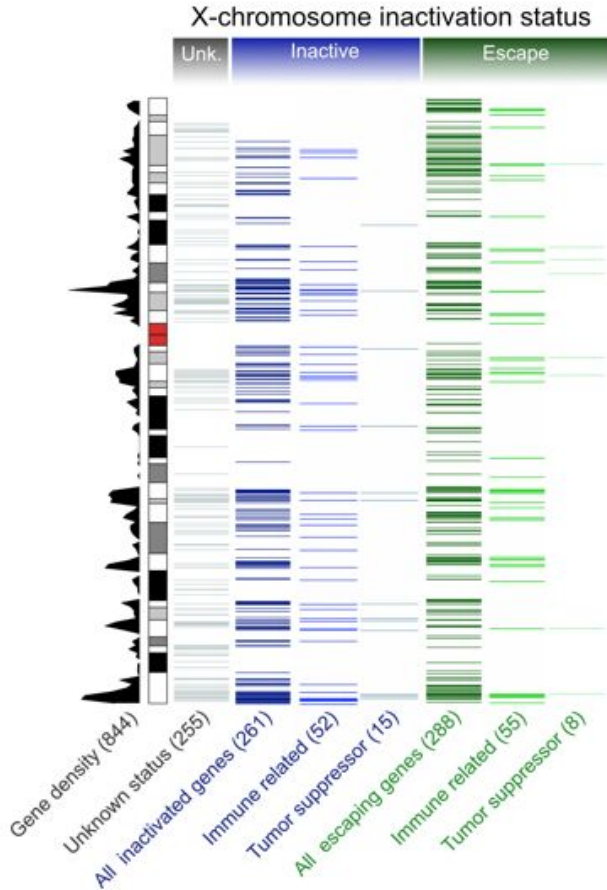


X-inactivation & X-linked expression



- Male bias
 - Adult cancers
 - Pediatric cancers
 - Heart disease
 - Susceptibility to COVID-19 (ACE2 receptor is X-linked)

X-inactivation & X-linked expression



- Male bias
 - Adult cancers
 - Pediatric cancers
 - Heart disease
 - Susceptibility to COVID-19 (ACE2 receptor is X-linked)
- Female bias
 - Heart disease after menopause
 - Autoimmune disease
 - Adverse reactions to COVID-19 vaccines



Acknowledgements



Tanya Phung



Kimberly Olney



Tim Webster

R35-MIRA





Acknowledgements



James Taylor
1979-2020



Brain O'Connor
@bconnor



Becky Boyles
@becky_boyles

*Good ideas don't have owners - they belong to everyone
-James Taylor*

Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA

Valerie Cotton & Allison Heath

Overall Goals

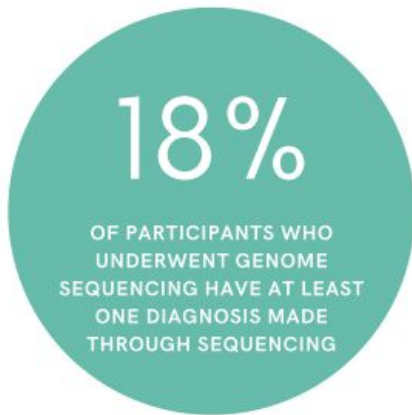
Use Case: *Enable researchers to easily co-analyze data from Kids First & the Undiagnosed Disease Network in the cloud to leverage large-scale pediatric cohorts from Kids First to resolve variants of unknown significance in UDN cases.*

Kids First: The goal of Kids First is to help researchers uncover new insights into the biology of childhood cancer and structural birth defects.



UDN: The Undiagnosed Diseases Network (UDN) is an initiative to facilitate the diagnosis of conditions that have eluded diagnosis through the coordinated action of leading clinical and research centers.





GENOME SEQUENCING

1,142 participants (716 children and 426 adults) have undergone genome sequencing. Many of these participants had non-diagnostic exome sequencing prior to enrollment in the UDN. The most common symptom category for participants undergoing genome sequencing is neurology (51%), followed by multiple congenital anomalies (9%).

- **Data access provided by:** [dbGaP Authorized Access](#)
- **Release Date:** September 27, 2021
- **Embargo Release Date:** September 27, 2021
- [Data Use Certification Requirements \(DUC\)](#)
- **Public Posting of [Genomic Summary Results](#):** Allowed
- **Use Restrictions**

Consent group	Is IRB required?	Data Access Committee	Number of participants
General Research Use 	No	National Human Genome Research Institute (nhgridac@mail.nih.gov)	4239

Scientific Narrative (specific use case)

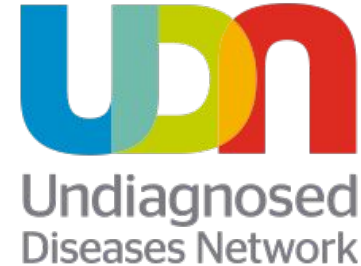
...To address the challenge of VUS's, we have developed a pipeline to assess variants found on clinical sequencing using biobank cohorts with linked phenotyped data.

Our pipeline creates a **phenotype risk score (PheRS)** of the proband based on their clinical presentation described in human phenotype ontology terms (HPO). We then apply the PheRS to the biobank cohort, such that individuals with many overlapping features have a high PheRS, and those with no or few overlapping features have a low score. We then identify variant matched individuals present in the biobank cohort, and test if the variant matched individuals have unexpectedly elevated phenotype risk scores.

We have been using this pipeline to analyze **Undiagnosed Disease Network (UDN)** patients, using a biobank cohort called BioVU... We believe that expanding our search for variant matched individuals to a large cohort like **Kids First** would enable us better interpret candidate variants for unsolved UDN cases.....



Lisa Bastarache



Overview of Standards Used



Undiagnosed
Diseases Network



4,000+ genomes

Up to 24,000 genomes



CAVATICA



Solution Matrix


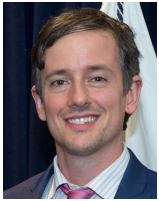












	Kids First Data Resource	NLM/NCBI	Analysis Tools
Genomic data	CAVATICA already integrated with the Kids First/Gen3 DRS server. RAS Milestone 3 is underway.	Connect CAVATICA to dbGaP DRS server, using RAS v1.1 Passports <ul style="list-style-type: none">- Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identify variants of unknown significance (VUSs) underlying undiagnosed conditions and “matched” cases in Kids First
Phenotypic data	CAVATICA is building a FHIR client to ingest from the Kids First FHIR-based data service	dbGaP on FHIR is in development. FHIR & RAS integration will be needed for controlled-access phenotypes	PheRS to compare phenotypes of individuals with the same/similar VUSs



Collaboration Matrix

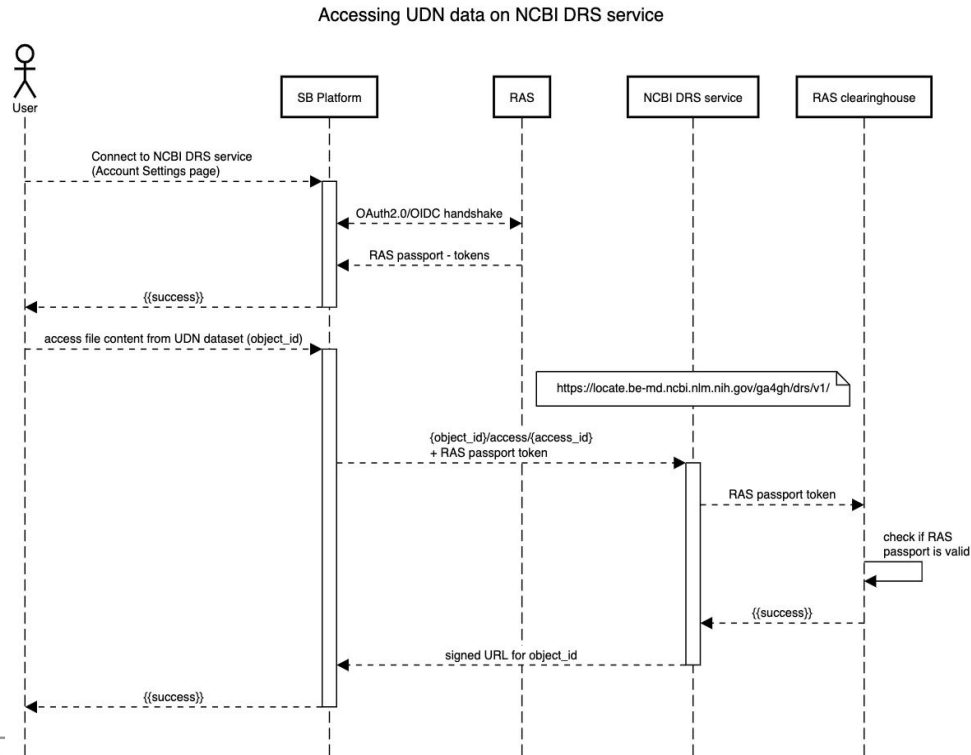


	Kids First Data Resource	NLM/NCBI	Tester/User
Genomic data	<p>Michele Mattioni & Jack DiGiovanna & Adam Resnick</p>   	<p>Kurt Rodarmer & Yuriy Skripchenko</p>  	<p>Yuankun Zhu & Anne Deslattes Mays</p>  
Phenotypic data	<p>Allison Heath & Robert Carroll</p>  	<p>Liz Amos & Mike Feolo</p>  	<p>Lisa Bastarache</p> 

Genomic Data Interoperability



Goal: Enable a user to Access the UDN genomic data via DRS, using RAS Passport





CAVATICA: RAS Connection



NHLBI BioData Catalyst Powered by Seven Bridges

Connect your [BioData Catalyst](#) account to import files via the BioData Catalyst DRS server. [Learn more.](#)

DRS Endpoint	Account	Expires	
drs://ga4gh-api.sb.biodatacatalyst.nih.gov	mmattioni	Oct. 23, 2021 14:04	Reconnect ...

Cancer Genomics Cloud Powered by Seven Bridges -- Import via DRS

Connect your [Cancer Genomics Cloud](#) account to import files via the Cancer Genomics Cloud DRS server. [Learn more.](#)

DRS Endpoint	Account	Expires	
drs://cgc-ga4gh-api.sbgenomics.com	mmattioni	Oct. 23, 2021 14:05	Reconnect ...

Connect with the NCBI DRS Server

DRS EndPoint

<https://locate.be-md.ncbi.nlm.nih.gov/ga4gh/drs/v1/>

[Connect](#)

- Seven Bridges identified solution to add a **new “card”** in the Account DataSets configuration tab



DRS links



1. Use [NCBI Run Selector](#) to obtain a manifest which contains SRA Runs
2. Use the IDX service to obtain the DRS links connected with the SRA Runs
 - Note: The DRS Links are offered in bundles, which Seven Bridges needs to build support for
 - At the moment Seven Bridges extract the bundles, and then obtains the DRS pointer to the file
3. Import the DRS File into Cavatica

Found 4,566 Items

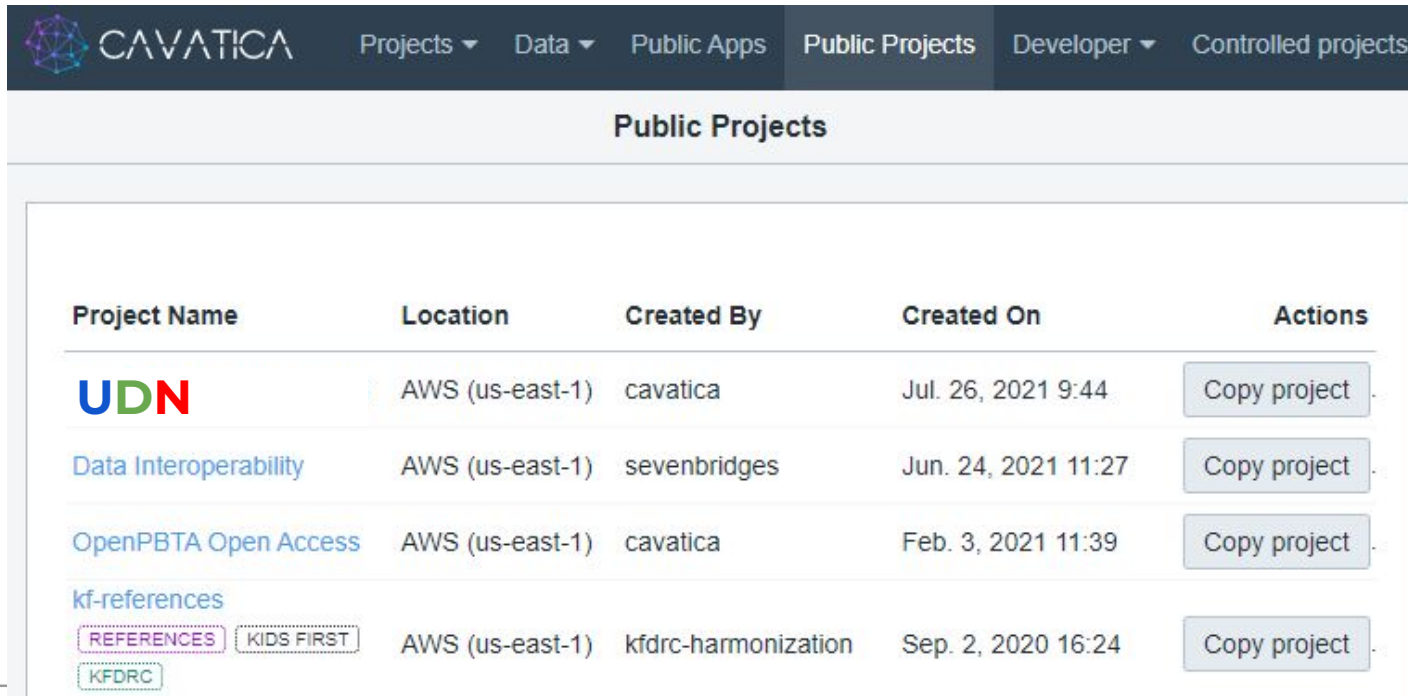
Search within results

1 1 92

<input checked="" type="checkbox"/>	Run	BioSample	alignment_software	analyte_type	Assay Type	biospecimen_repository_sample_id	body_site	Bytes	Center Name	
<input type="checkbox"/>	1	SRR5031422	AMN05980034	BWA-mem v0.7.12	DNA	WGS	8657f8fb-432b-4473-a31b-060384c4b79f	Blood	55.83 Gb	NHGRI-PHS001232
<input type="checkbox"/>	2	SRR5031424	AMN05980042	BWA-mem v0.7.12	DNA	WGS	57b49db5-2778-4557-9fbb-9ff454cf4212	Blood	61.43 Gb	NHGRI-PHS001232
<input type="checkbox"/>	3	SRR5031427	AMN05980030	BWA-mem v0.7.12	DNA	WGS	a2529ebc-4e29-4d60-93a8-fa07ed9f84a4	Blood	78.20 Gb	NHGRI-PHS001232
<input type="checkbox"/>	4	SRR5031429	AMN05980037	BWA-mem v0.7.12	DNA	WGS	33ad6df6-f122-4e14-b75f-82733c39a220	Blood	82.83 Gb	NHGRI-PHS001232
<input type="checkbox"/>	5	SRR5031431	AMN05980040	BWA-mem v0.7.12	DNA	WGS	6f94ba61-73d7-4551-80b3-6591001c437a	Blood	74.80 Gb	NHGRI-PHS001232
<input type="checkbox"/>	6	SRR5031434	AMN05980032	BWA-mem v0.7.12	DNA	WGS	e8bb68df-e276-4604-94cf-05b57902f337	Blood	72.77 Gb	NHGRI-PHS001232
<input type="checkbox"/>	7	SRR8257099	AMN10087985	BWA-mem v0.7.12	DNA	WGS	c6c974cc-86e8-42d8-92ba-ab10f1b37557	Blood	17.33 Gb	HMS-CC
<input type="checkbox"/>	8	SRR8060841	AMN10087770	BWA-mem v0.7.12	DNA	WGS	31e6d861-ccb8-41c2-9ebc-c4e05251e690	Blood	51.81 Gb	HMS-CC
<input type="checkbox"/>	9	SRR8060849	AMN10087459	BWA-mem v0.7.12	DNA	WGS	1725b288-f786-4148-86f2-0afe61d77ec2f	Blood	17.78 Gb	HMS-CC

Draft Approach for UDN Data Findability

The dataset will be findable/searchable as a CAVATICA Public Project (dbGaP approval still required). The DRS file would be built into the Project.



The screenshot shows the CAVATICA web interface. At the top, there is a navigation bar with the CAVATICA logo and several menu items: Projects, Data, Public Apps, Public Projects (which is highlighted), Developer, and Controlled projects. Below the navigation bar, the page title is "Public Projects". The main content area displays a table of public projects. The table has five columns: Project Name, Location, Created By, Created On, and Actions. The first row is for the "UDN" project, which is highlighted in blue. The other rows are for "Data Interoperability", "OpenPBTA Open Access", and "kf-references". The "kf-references" row has several sub-rows for "REFERENCES", "KIDS FIRST", and "KFDRC". Each row has a "Copy project" button in the Actions column.

Project Name	Location	Created By	Created On	Actions
UDN	AWS (us-east-1)	cavatica	Jul. 26, 2021 9:44	Copy project
Data Interoperability	AWS (us-east-1)	sevenbridges	Jun. 24, 2021 11:27	Copy project
OpenPBTA Open Access	AWS (us-east-1)	cavatica	Feb. 3, 2021 11:39	Copy project
kf-references				
REFERENCES	AWS (us-east-1)	kfdrc-harmonization	Sep. 2, 2020 16:24	Copy project
KIDS FIRST				
KFDRC				

Variant Identification

- For functional equivalence, call UDN variants using [Kids First workflows](#)
- Use [Kids First Portal variant search](#) to identify datasets of interest → Apply for those datasets in dbGaP
- Use Kids First VCFs to identify variant matched individuals
- Run PheRS

Variant	Type	dbSnp	Consequences	CLINVAR	Studies	Participants
chrX:g.48792004del	deletion	--	● frameshift_variant GATA1 G126X	--	<u>1</u>	1 / 4843
chrX:g.48794116del	deletion	--	● frameshift_variant GATA1 G397X	--	<u>1</u>	1 / 4843
chrX:g.48791978C>A	SNV	--	● missense_variant GATA1 Q119K	--	<u>1</u>	1 / 4843
chrX:g.48792194C>T	SNV	rs140561920	● missense_variant GATA1 R191C	Benign	<u>1</u>	4 / 4843

Solution Matrix



	Kids First Data Resource	NLM/NCBI	Analysis Tools
Genomic data	CAVATICA already integrated with the Kids First/Gen3 DRS server. RAS Milestone 3 is underway.	Connect CAVATICA to dbGaP DRS server, using RAS v1.1 Passports <ul style="list-style-type: none">- Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identify variants of unknown significance (VUSs) underlying undiagnosed conditions and “matched” cases in Kids First
Phenotypic data	CAVATICA is building a FHIR client to ingest from the Kids First FHIR-based data service	dbGaP on FHIR is in development. FHIR & RAS integration will be needed for controlled-access phenotypes	PheRS to compare phenotypes of individuals with the same/similar VUSs



PheRS pipeline



- R-based tool creates a phenotype risk score (PheRS) of the proband based on their clinical presentation described in human phenotype ontology terms (HPO).
 - ✓ **Kids First already maps phenotypes to HPO**
- Apply PheRS to the cohort, such that individuals with many overlapping features have a high PheRS, and those with no or few overlapping features have a low score.
- Identify variant matched individuals and test if they have unexpectedly elevated phenotype risk scores
- Make available to the community and path for utilization/comparison with other work like [LIRICAL](#)

Proband phenotype

Clinical symptoms and physical findings

GROWTH PARAMETERS

Failure to thrive

CARDIOVASCULAR

Patent ductus arteriosus

GASTROINTESTINAL

Elevated hepatic transaminase

Gastroesophageal reflux

GENITOURINARY

Hydrocele testis

BEHAVIOR, COGNITION AND DEVELOPMENT

Global developmental delay

Delayed speech and language development

DIGESTIVE SYSTEM

Hepatomegaly

METABOLISM/HOMEOSTASIS

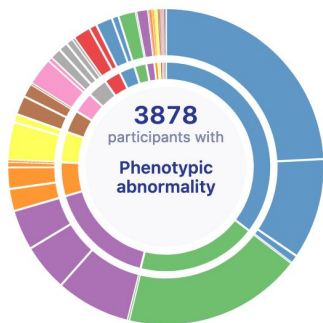
Recurrent hypoglycemia

Neonatal hypoglycemia

Candidate variants

Heterozygous Variants						
Gene	Chr Position rs#	Change	Effect	Proband	Mother (Unaff)	Father (Unaff)
COL9A1 NM_001851.4	chr6	A → T	splice donor 10.9>2.7	●○	○○	●○
	70991091	c.876+2T>A				
	rs149830493					
ELN NM_000501	chr7	G → A	missense	●○	○○	●○
	73470684	c.1234G>A				
	rs375116795	p.Gly412Arg				
PIGN NM_012327	chr18	T → C	missense	●○	○○	●○
	59757754	c.2238A>G				
	rs200658159	p.Ile746Met				
POLG NM_002693.2	chr15	G → C	missense	●○	○○	●○
	89872002	c.1084C>G				
	rs763248358	p.Leu362Val				
RFT1 NM_052859.3	chr3	C → T	missense	●○	●○	○○
	53140879	c.782G>A				
	rs374781452	p.Arg261Gln				

Observed Phenotypes



- Phenotypic abnormality (HP:0000118)
- Abnormality of head or neck (HP:0000152)
- Abnormality of the musculoskeletal system (HP:0033127)
- Abnormality of the cardiovascular system (HP:0001626)
- Abnormality of the nervous system (HP:0000707)
- Abnormality of the eye (HP:0000478)
- Abnormality of the genitourinary system (HP:0000119)
- Abnormality of the digestive system (HP:0025031)
- Abnormality of the respiratory system (HP:0002086)
- Neoplasm (HP:0002664)
- Abnormality of the ear (HP:0000598)
- Abnormality of the integument (HP:0001574)
- Abnormality of limbs (HP:0040064)
- Growth abnormality (HP:0001507)
- Abnormality of the immune system (HP:0002715)
- Abnormality of the endocrine system (HP:0000818)
- Abnormality of prenatal development or birth (HP:0001197)
- Abnormality of blood and blood-forming tissues (HP:0001871)
- Abnormality of the breast (HP:0000769)
- Abnormal cellular phenotype (HP:0025354)
- Abnormality of metabolism/homeostasis (HP:0001939)

27	3878
0	1480
0	1328
41	957
15	431
7	355
25	341
80	327
0	303
0	278
8	266
1	196
53	190
0	90
0	59
0	41
0	26
0	24
0	18
0	10
0	9

chr18:g.62090521T>C Germline

[Summary](#) [Frequencies](#) [Clinical Associations](#)

Chr	18	4 Studies	18 Participants	3.72e-3 Frequency
Start	62090521			
Alt. Allele	C			
Ref. Allele	T			
Type	SNV	Ref Genome	ClinVar	dbSNP
		GRCh38	539565	rs200658159

Gene Consequences

Gene PIGN

AA	Consequence	Coding Dna	Strand	VEP	Impact	Conservation	Transcript
1746M	missense_variant	2238T>C	-	Moderate	Sift: 0.13045 Polyphen2: Benign - 0.13045 More	0.05595	ENST00000640252

[Show Transcripts \(28\)](#)



Driving Tool / Service Layers: General

AnVIL Services

FHIR

KF Services

FHIR

Platform Services

FHIR

NCPI Phenotype
Translation Tool

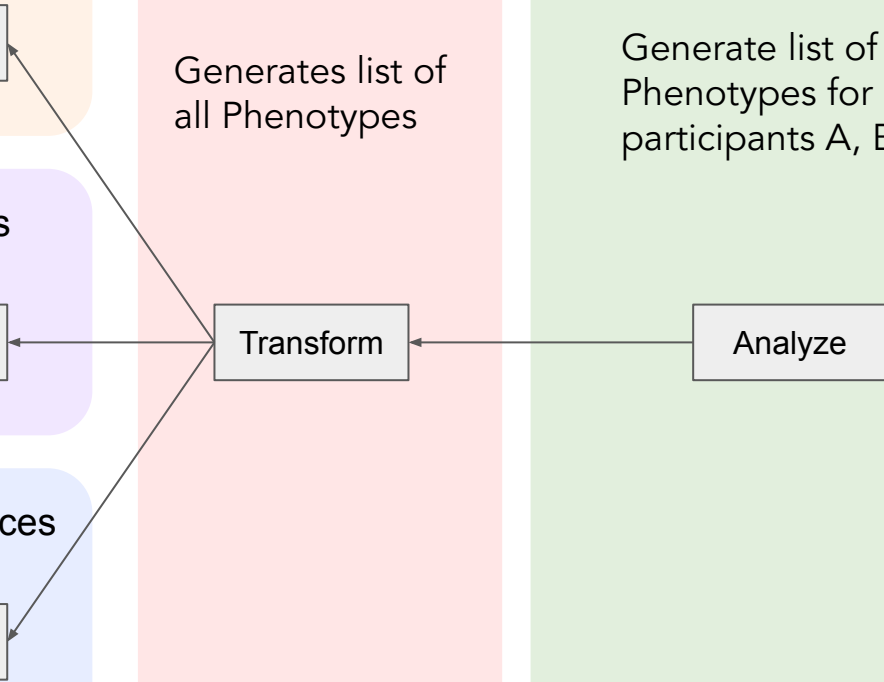
Generates list of
all Phenotypes

Transform

Research User

Generate list of all
Phenotypes for
participants A, B, C

Analyze



Driving Tool / Service Layers: Use Case

NCPI Phenotype Translation Tool

Generates list or table of available HPO terms from KF and UDN

Phenotype Risk Score Pipeline

Generate PheRS for all participants A, B, C

Combining Genomic and Phenotypic Pipeline Results

Do “variant matched individuals have unexpectedly elevated phenotype risk scores” ?

KF Services

FHIR

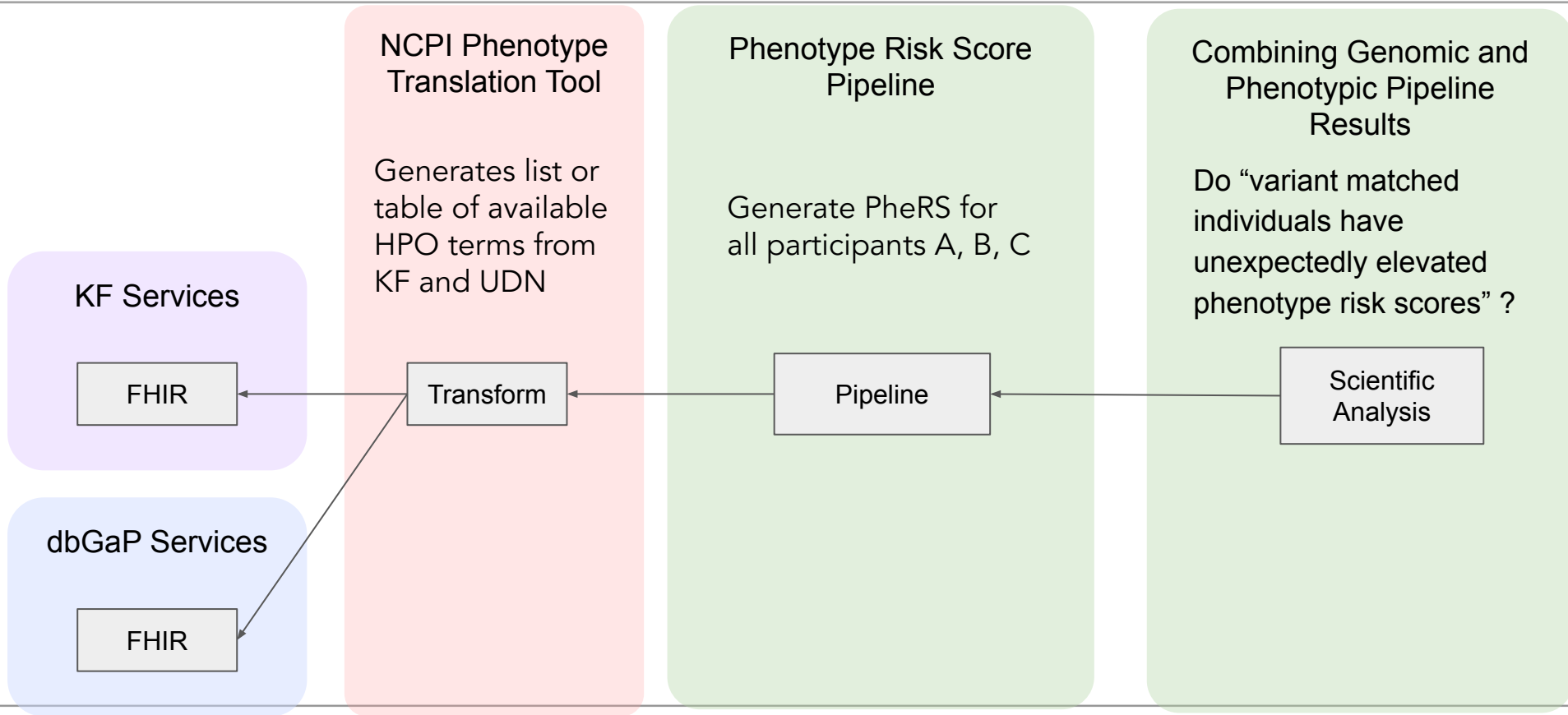
Transform

Pipeline

Scientific Analysis

dbGaP Services

FHIR

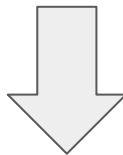


Concrete Progress on Each Step

What protocol is the most practical/useful?

- PFB?
- ndjson?
- FHAvro?

RAS-based access to FHIR data



KF Services

FHIR

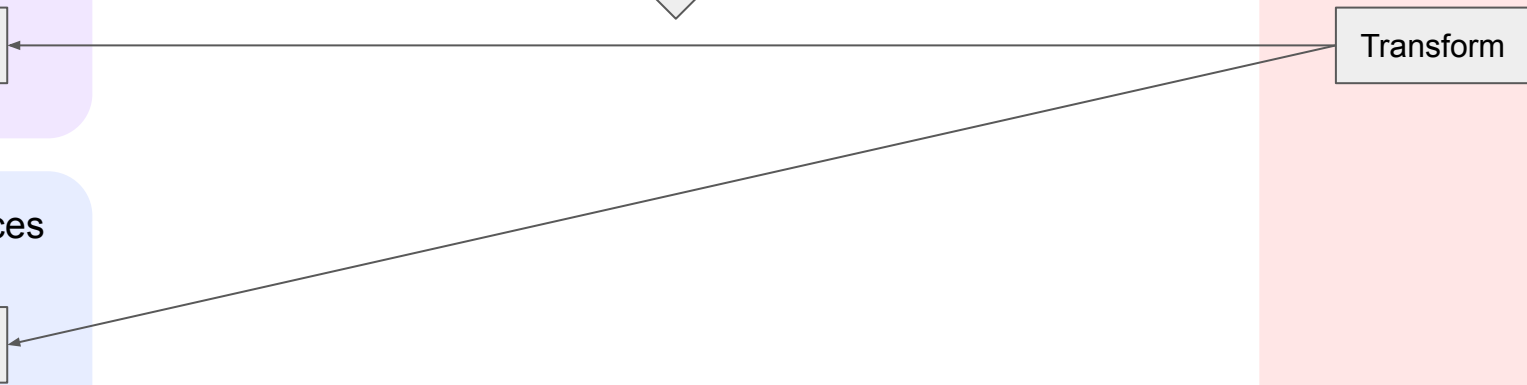
dbGaP Services

FHIR

NCPI Phenotype Translation Tool

Generates list or table of available HPO terms from KF and UDN

Transform



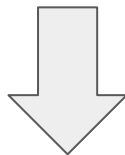
Concrete Progress on Each Step

NCPI Phenotype Translation Tool

Generates list or table of available HPO terms from KF and UDN

Transform

What current cloud workspace tooling fits best here? Do we need to be able to support additional capabilities?



Phenotype-based Pipeline

Generate PheRS for all participants A, B, C

Automated Pipeline



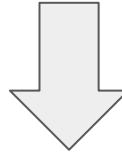
Concrete Progress on Each Step

Phenotype-based Pipeline

Generate PheRS for all participants A, B, C

Automated Pipeline

May be the most well-defined? Happens in a R Studio or Jupyter notebook environment?



Combining Genomic and Phenotypic Pipeline Results

Do “variant matched individuals have unexpectedly elevated phenotype risk scores” ?

Scientific Analysis

Doors to New Capabilities

I found something interesting - is there more data/information about this patient?

Combining Genomic and Phenotypic Pipeline Results

Do “variant matched individuals have unexpectedly elevated phenotype risk scores” ?

EHR Systems /
Research
Warehouses

FHIR

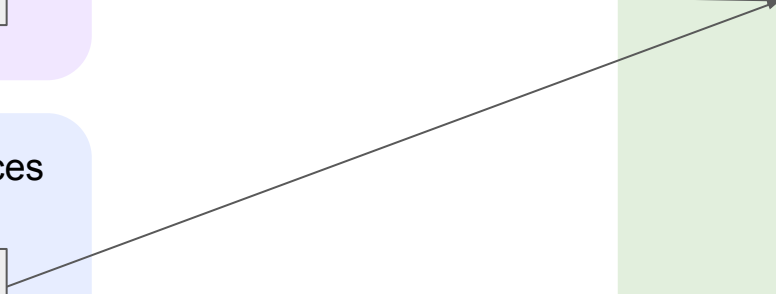
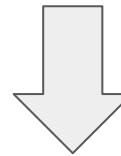
KF Services

FHIR

dbGaP Services

FHIR

Scientific
Analysis



Leveraging Functionally Equivalent Pipelines for Long-Read Data on Different Systems

Owen Hirschi
Dr. Sharon Plon's Lab
Baylor College of Medicine

The Plon lab utilizes multiple platforms to store and analyze sequencing data from pediatric cancer cohorts



Germline WGS

BASIC³

BCM Advancing Sequencing
Into Childhood Cancer Care

Follow-up study
Germline, Tumor Exome
Transcriptome

Germline Exome
Tumor Exome
Transcriptome



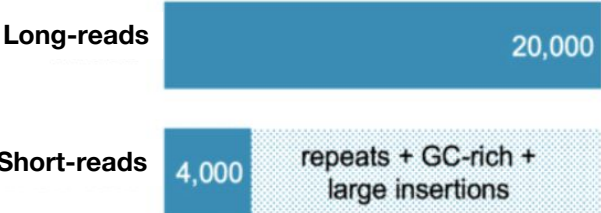
Anvil



BASIC3 is undergoing Pacific Biosciences HiFi CCS long-read sequencing

Long-read sequencing allows for greater detection of SV

Structural Variants Observed



Allows for the comparison of long-read and short-read structural variant calling

Algorithms being utilized:

minimap2
v2.17

minimap2 v2.17

Minimap2 is a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference dat...

ALIGNMENT | GENOMICS | LONG READS

CWL1.0

Copy Run

Sniffles CWL1.1
1.0.12b

Sniffles 1.0.12b

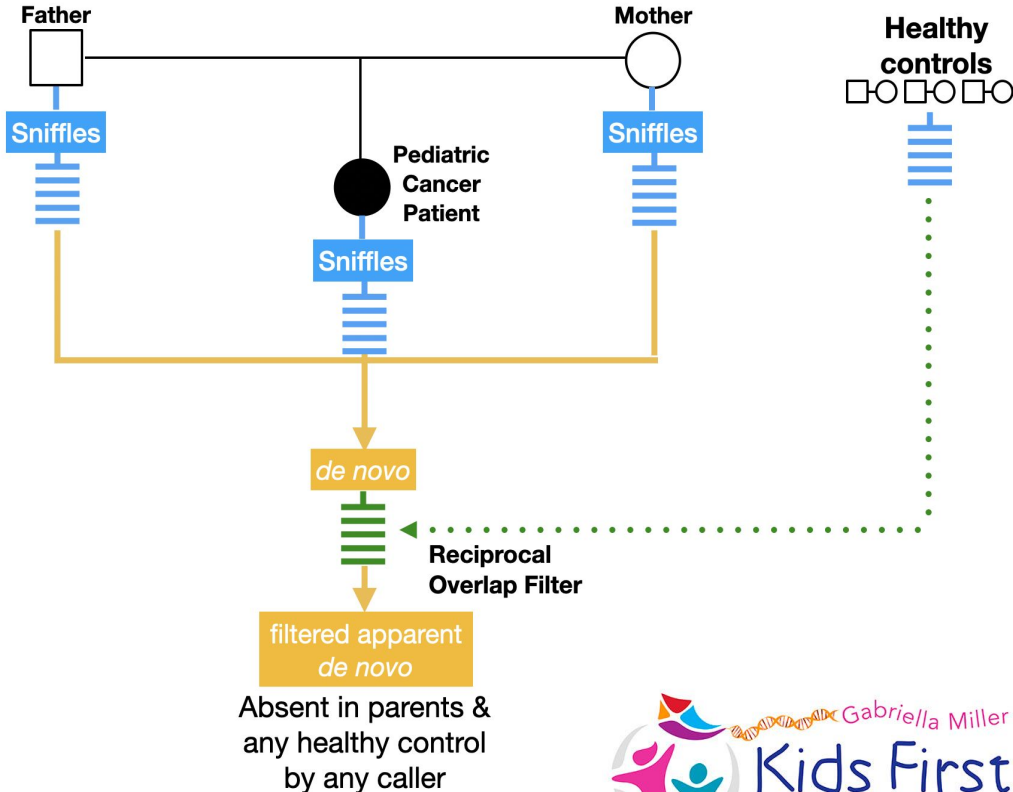
Sniffles is a structural variation caller for PacBio or Oxford Nanopore data [1,2].

*A list of **all inputs and ...

VARIANT CALLING | CWL1.1 | GENOMICS

LONG READS

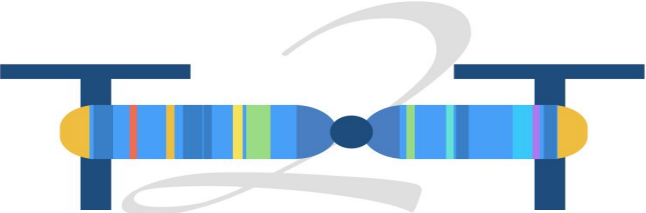
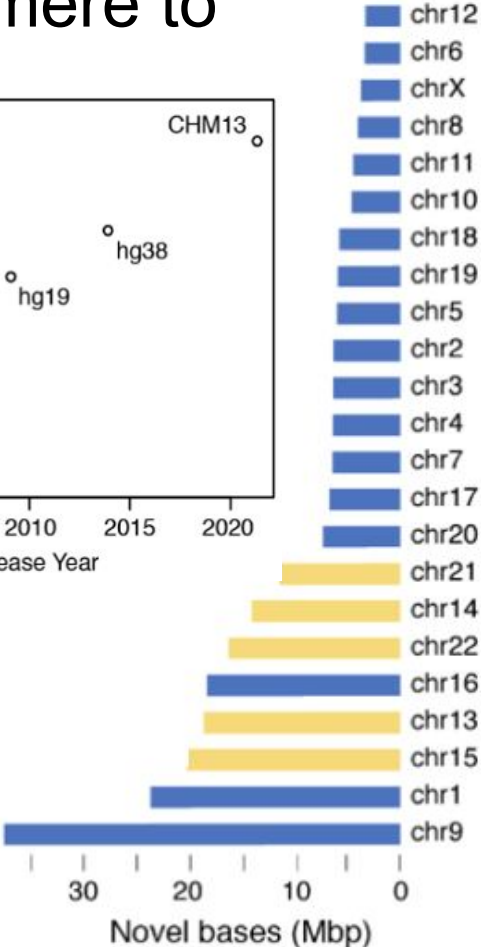
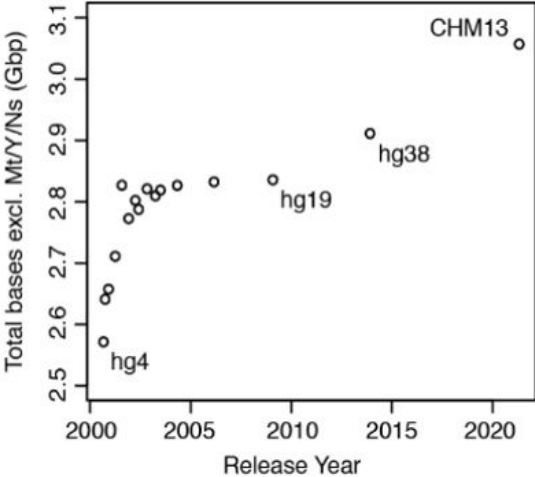
Copy Run



Novel CHM13 reference genome by the Telomere to Telomere (T2T) consortium

The complete sequence of a human genome

Sergey Nurk^{1,*}, Sergey Koren^{1,*}, Arang Rhie^{1,*}, Mikko Rautiainen^{1,*}, Andrey V. Bzikadze², Alla Mikheenko³, Mitchell R. Vollger⁴, Nicolas Altemose⁵, Lev Uralsky^{6,7}, Ariel Gershman⁸, Sergey Aganezov⁹, Savannah J. Hoyt¹⁰, Mark Diekhans¹¹, Glennis A. Logsdon⁴, Michael Alonge⁹, Stylianos E. Antonarakis¹², Matthew Borchers¹³, Gerard G. Bouffard¹⁴, Shelise Y. Brooks¹⁴, Gina V. Caldas¹⁵, Haoyu Cheng^{16,17}, Chen-Shan Chin¹⁸, William Chow¹⁹, Leonardo G. de Lima¹³, Philip C. Dishuck⁴, Richard Durbin²¹, Tatiana Dvorkina³, Ian T. Fiddes²², Giulio Formenti^{23,24}, Robert S. Fulton²⁵, Arkarachai Functamman¹⁸, Erik Garrison^{11,26}, Patrick G.S. Grady¹⁰, Tina A. Graves-Lindsay²⁷, Ira M. Hall²⁸, Nancy F. Hansen²⁹, Gabrielle A. Hartley¹⁰, Marina Haukness¹¹, Kerstin Howe¹⁹, Michael W. Hunkapiller³⁰, Chirag Jain^{1,31}, Miten Jain¹¹, Erich D. Jarvis^{23,24}, Peter Kerpedjiev³², Melanie Kirsche⁹, Mikhail Kolmogorov³³, Jonas Korlach³⁰, Milinn Kremitzki²⁷, Heng Li^{16,17}, Valerie V. Maduro³⁴, Tobias Marschall³⁵, Ann M. McCartney¹, Jennifer McDaniel³⁶, Danny E. Miller^{4,37}, James C. Mullikin^{14,29}, Eugene W. Myers³⁸, Nathan D. Olson³⁶, Benedict Paten¹¹, Paul Peluso³⁰, Pavel A. Pevzner³³, David Porubsky⁴, Tamara Potapova¹³, Evgeny I. Rogae^{6,7,39,40}, Jeffrey A. Rosenfeld⁴¹, Steven L. Salzberg^{9,42}, Valerie A. Schneider⁴³, Fritz J. Sedlazeck⁴⁴, Kishwar Shafin¹¹, Colin J. Shew²⁰, Alaina Shumate⁴², Yumi Sims¹⁹, Arian F. A. Smit⁴⁵, Daniela C. Soto²⁰, Ivan Sovic^{30,46}, Jessica M. Storer⁴⁵, Aaron Streets^{5,47}, Beth A. Sullivan⁴⁸, Françoise Thibaud-Nissen⁴³, James Torrance¹⁹, Justin Wagner³⁶, Brian P. Walenz¹, Aaron Wenger³⁰, Jonathan M. D. Wood¹⁹, Chunlin Xiao⁴³, Stephanie M. Yan⁴⁹, Alice C. Young¹⁴, Samantha Zarate⁹, Urvasi Surti⁵⁰, Rajiv C. McCoy⁴⁹, Megan Y. Dennis²⁰, Ivan A. Alexandrov^{3,7,51}, Jennifer L. Gerton¹³, Rachel J. O'Neill¹⁰, Winston Timp^{8,42}, Justin M. Zook³⁶, Michael C. Schatz^{9,49}, Evan E. Eichler^{4,24,†}, Karen H. Miga^{11,†}, Adam M. Phillippy^{1,†}



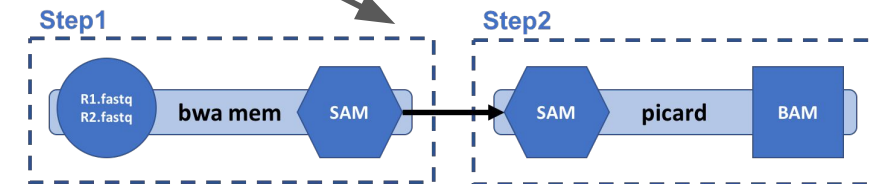
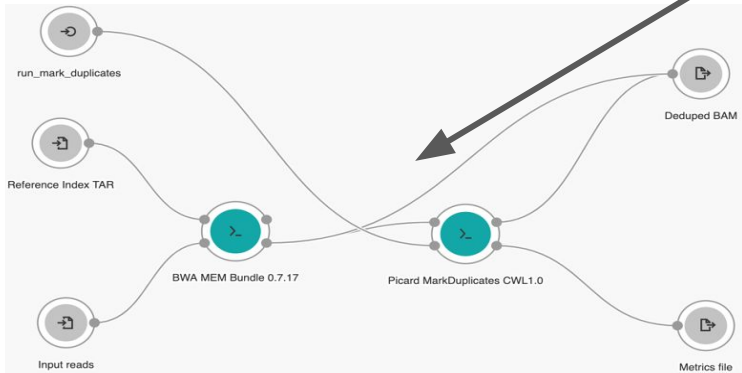
The tools are uploaded in different languages across platforms



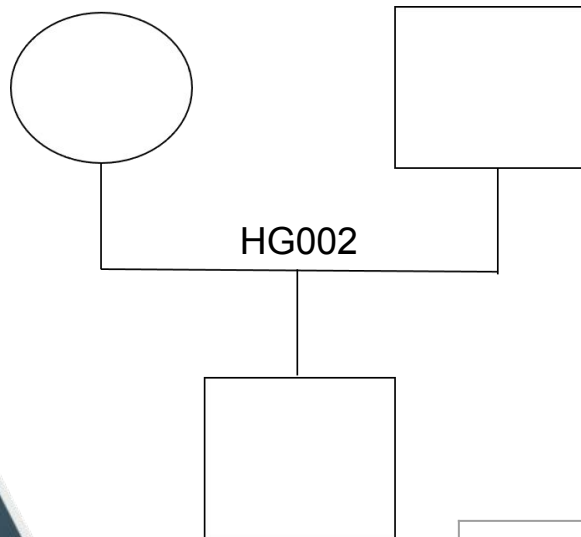
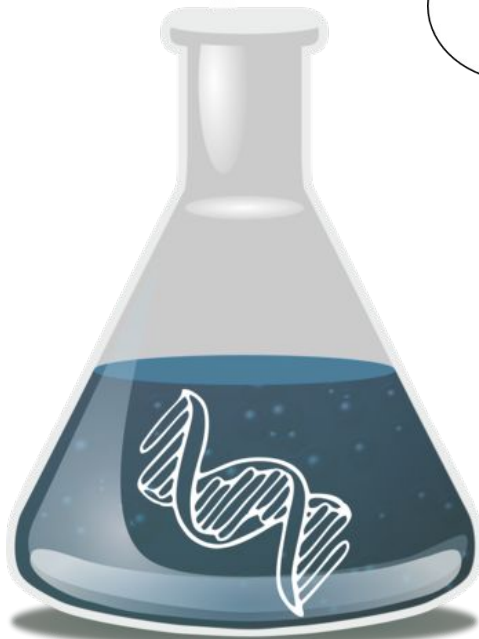
Tools are written in **Common Workflow Language (CWL)**

Tools are written in **Workflow Description Language (WDL)**

BWA MEM & Picard



GIAB Benchmarking Data: HG002 Trio and Benchmarking Pipeline



Long-Read Technology

PacBio Circular Consensus Sequencing (HiFi CCS)

Oxford Nanopore Promethion (ONT)

PacBio Continuous Long Read (CLR)

Tool	Version
Minimap2 (FASTQ Aligner)	2.17
Sniffles (Structural Variant Caller)	1.0.11
SURVIVOR (SV merging)	1.07

Creation of Long-Read SV Calling Pipeline on CAVATICA

CAVATICA Projects Data Public Apps

Public apps

Structural Variants Category

Explore genomics data

Understand complex genomics data with interactive analysis tools.



Data Cruncher

Analyze and explore data using JupyterLab or RStudio

Open

minimap2

minimap2 v2.17

Minimap2 is a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference dat...

ALIGNMENT GENOMICS LONG READS

CWL1.0

Copy Run

Sniffles CWL1.1

Sniffles 1.0.12b

Sniffles is a structural variation caller for PacBio or Oxford Nanopore data [1,2].

*A list of **all inputs and ...

VARIANT CALLING CWL1.1 GENOMICS

LONG READS

Copy Run

File Edit View Run Kernel Tabs Settings Help

Launcher 01ateck.by.sa

```

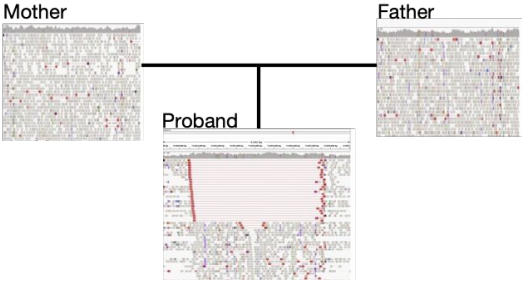
1 #!/bin/bash
2
3
4 #DEL
5 del
6 while [ $n -le 165 ]
7 do
8
9   # Stack all the filtered SV calls per sample,
10  # and perform some additional filtering
11
12  # update: small filter size change from 1kb to 100bp
13  # update: use new exclude region
14  # update: specify caller in script
15  # update: directly decide reciprocal overlap using new script
16  # update: remove variants
17  # update: remove sample name and role in output
18  # update: intersect with healthy control stack data to identify denovo variant
19  # update: need output that stack all healthy calls
20
21 event=DEL
22 caller=(cnvnator delly lumpy manta breakdancer)
23

```

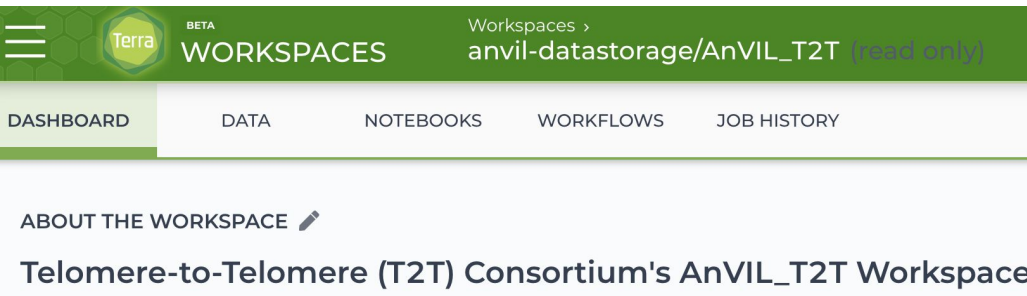
SURVIVOR

SURVIVOR is a tool set for simulating/evaluating SVs, merging and comparing SVs within and among samples, and includes various methods to reformat or summarize SVs.

IGV images



HG002-Trio processed on Anvil as a part of T2T studies

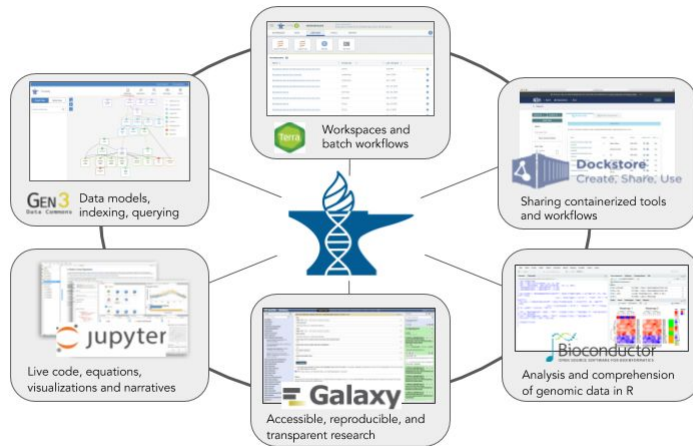


BETA
Terra
WORKSPACES
Workspaces >
anvil-datastorage/AnVIL_T2T (read only)

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

ABOUT THE WORKSPACE ✎

Telomere-to-Telomere (T2T) Consortium's AnVIL_T2T Workspace



A complete reference genome improves analysis of human genetic variation

Sergey Aganezov, Stephanie M. Yan, Daniela C. Soto,  Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J. Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, Justin Wagner, Jennifer McDaniel, Nathan D. Olson, Michael E.G. Sauria,  Mitchell R. Vollger,  Arang Rhie, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren,  Jeffrey A. Rosenfeld, Benedict Paten,  Ryan Layer, Chen-Shan Chin, Fritz J. Sedlazeck, Nancy F. Hansen, Danny E. Miller, Adam M. Phillippy, Karen H. Miga,  Rajiv C. McCoy,  Megan Y. Dennis,  Justin M. Zook,  Michael C. Schatz

doi: <https://doi.org/10.1101/2021.07.12.452063>

This article is a preprint and has not been certified by peer review [what does this mean?].

Jasmine: Population-scale structural variant comparison and analysis

 Melanie Kirsche,  Gautam Prabhu,  Rachel Sherman,  Bohan Ni,  Sergey Aganezov,  Michael C. Schatz

doi: <https://doi.org/10.1101/2021.05.27.445886>

This article is a preprint and has not been certified by peer review [what does this mean?].

doi.org/10.1101/2021.07.12.452063
doi.org/10.1101/2021.05.27.445886

Preliminary Results:

Post Minimap2 alignment:

Sample	Coverage (Terra.Bio)	Coverage (CAVATICA)
HG002	35.25	35.03
HG003	33.68	33.47
HG004	33.18	32.99

Post Sniffles variant calling:

Sample	Raw Structural Variant Count (Terra.Bio)	Raw Structural Variant Count (CAVATICA)	Difference
HG002	92,350	96,977	+4,627
HG003	90,357	94,361	+4,004
HG004	88,803	93,159	+4,356

Discordant variant calls using SURVIVOR:

Sample	Variant Count (Terra.Bio)	Variant Count (CAVATICA)	Difference
Only in HG002	3,934	4,307	+373
Only in HG003	9,478	10,255	+777
Only in HG004	9,468	10,486	+1,018

De novo variants examples:

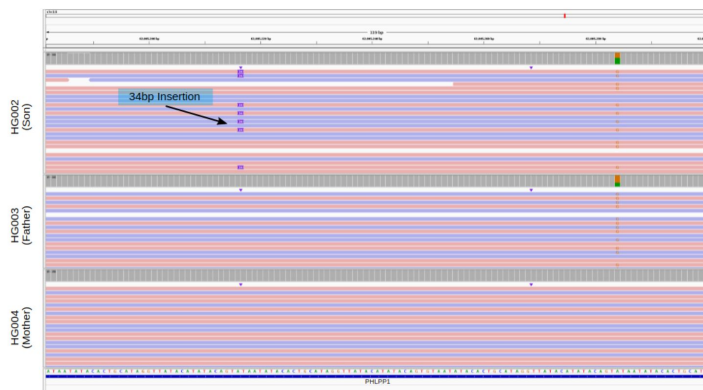
Deletion identified on the Terra platform



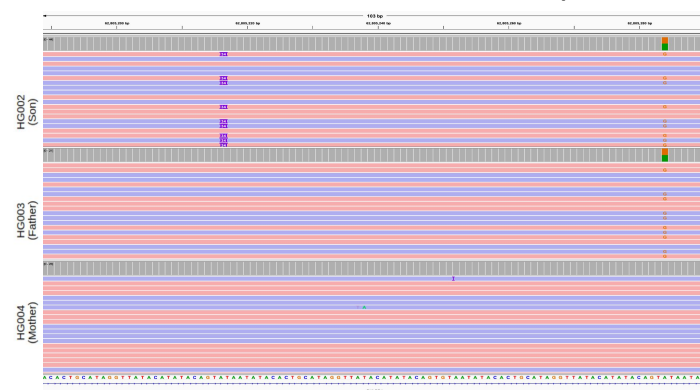
Deletion identified on the CAVATICA platform



Insertion identified on the Terra platform



Insertion identified on the CAVATICA platform



Summary

- Long-read sequence analysis tools uploaded on these platforms exist in different coding languages
- We have set up a functional long-read sequencing analysis pipeline on the CAVATICA platform
- We have been able to identify *de novo* variants previously found via pipelines on the Terra platform
- We have also identified a 5 to 10% difference in raw and merged structural variants across the two platforms

Ongoing Work

- Understand differences in called *de novo* events and aligned sequence files in HG002 trio on both platforms
- Determine if there is an larger data set we can process on CAVATICA and Terra respectively to test full functional equivalence
- Perform long-read sequence analysis on BASIC3 cohort using the pipeline on CAVATICA to identify novel *de novo* structural variant

Acknowledgments

Baylor
College of
Medicine



Plon Lab members:

Sharon Plon, MD, PhD

Saumya Sisoudiya

P. Adam Weinstein

Deborah Ritter, PhD

Xi Luo, PhD

Ryan Zabriskie

BASIC3 Co-PI:

William Parsons, MD, PhD

Schatz Lab:

Michael Schatz, PhD

Melanie Krishce

Seven Bridges:

Jack DiGiovanna, PhD

Jelena Randjelovic, PhD

Funding:

SevenBridges



CULLEN
FOUNDATION

BASIC3: NHGRI/NCI 1U01HG006485

KF BASIC3 : 1 X01 HL136998-01

CTR-CAQ T32: 1T32GM136554-01

F31: 5F31CA265163-02

Conducting reproducible science in PIC-SURE interoperating with Seven Bridges/Terra

Simran Makwana and Paul Avillach

Overview

- PIC-SURE Overview
- Use Case 1: PIC-SURE and Seven Bridges ORCHID study reproducibility
- Use Case 2: PIC-SURE and Terra HCT for SCD
- PIC-SURE as a search tool across NCPI platforms





PIC-SURE

**Patient-centered Information Commons:
Standardized Unification of Research Elements**

<https://picsure.biodatacatalyst.nhlbi.nih.gov/>

	User Interface (UI)	Application Programming Interface (API)
Advantages	Point-and-click interface to explore variables and aggregate counts	Use code to extract data directly into workspace
Access point	PIC-SURE website	Didactic Jupyter notebooks in R, python, R Markdown files
Building queries	Query Builder tool	Python and R functions
Extracting data	Data can be downloaded or exported to an analysis workspace	Run query in python or R to export data to workspace
Data	<p>Integrates clinical and genomic datasets across BioData Catalyst, including:</p> <ul style="list-style-type: none"> ○ TOPMed and TOPMed-related studies ○ COVID-19 studies ○ BioLINCC <p>Patient-level curation and ingestion of each phenotypic variable and genomic variant Variable, table and study metadata ingested and indexed for search. Decoded variables from all studies made available to the user for cohort filtering and export</p>	

PIC-SURE Open and Authorized Access

Authorized Access

Explore Now

29 Studies
234,781 Participants



dbGap Approval Required



Authorized Phenotypic and Genomic Datasets



Aggregate Counts



Patient Level Data



Download Authorized Datasets



R and Python API Access

Open Access

Explore Now

56 Studies
279,145 Participants



No Authorization Required



All Phenotypic Datasets Available in
PIC-SURE



Aggregate Counts Only

Anyone with an eRA Commons ID can access!

<https://picsure.biodatacatalyst.nhlbi.nih.gov/>

Below is a listing of FHS S-IARE datasets. Datasets are grouped according to four categories:

1. Clinic Exam Questionnaire - (Interview and Physical Exam) - Data collected during FHS clinic exam or ancillary study
2. Validated through medical records review and/or derived and/or scored and/or abstracted from other datasets for ease of use
3. Tests - Non-invasive tests
4. Laboratory - blood or urine

Some datasets may appear in more than one category depending upon the nature of the variables they contain.

Clinical data in dbGAP is stored in hundreds of files For EACH consent group Framingham heart study

Clinic Questionnaire (Interview and Physical Exam)

Clinic Exam Questionnaire

MD Interview, Physical Exam, Examiner's Opinion, and Clinical Diagnostic Impression; Non-MD / Non-medical Interview / Self-report and Physical Exam / Anthropometrics / Observed Performance

[ex0_7s](#) [ex0_8s](#) [ex0_9s](#) [ex0_10s](#) [ex0_11s](#) [ex0_12s](#) [ex0_13s](#) [ex0_14s](#) [ex0_15s](#) [ex0_16s](#) [ex0_17s](#) [ex0_18s](#) [ex0_19s](#) [ex0_20s](#) [ex0_21s](#) [ex0_22s](#) [ex0_23s](#) [ex0_24s](#) [ex0_25s](#) [ex0_26s](#) [ex0_27s](#) [ex0_28s](#) [ex1_1s](#) [ex1_2s](#) [ex1_3s](#) [ex1_4s](#) [ex1_5s](#) [ex1_6s](#) [ex1_7s](#) [ex1_8s](#)
[ex3_1s](#) [e_exam_2011_m_0017s](#) [e_exam_ex01_7_0020s](#) [e_exam_ex02_7_0003s](#) [e_exam_ex03_7_0426s](#) [e_exam_ex29_0_0210s](#) [e_exam_ex30_0_0274s](#) [e_exam_ex09_1b_0844s](#) [e_exam_ex01_2_0813s](#) [e_exam_ex01_72_0652s](#) [e_exam_ex32_0_0939s](#) [e_exam_ex31_0_0738s](#)

MD Interview

[menarche1_7s](#)

Non-MD / Non-medical Interview / Self-report

[act1_5s](#) [act1_6s](#) [dis0_18s](#) [psych1_3s](#) [sf36_1_6s](#) [bwqt1_6s](#) [resp1_6s](#) [ffreq1_3s](#) [ffreq1_5s](#) [ffreq1_6s](#) [ffreq1_7s](#) [ffreq0_20s](#) [ffreq0_21s](#) [ffreq0_22s](#) [menarche1_7s](#) [q_mnshist_2001_1_0650s](#) [q_psvsocalp_ex10_0_0657s](#)

Neuropsychology Questionnaire

[obsperform_2005s](#)

Validated / Reviewed / Scored / Abstracted Data

Foot Study

[vr_foot_2008_m_0511s](#) [vr_foot2_2008_m_0651s](#)

Menopause

[mnp0_14s](#) [meno1_8s](#) [vr_meno_ex02_3_0653s](#) [vr_meno_ex03_7_0916s](#) [vr_meno_ex02_2_0719s](#) [vr_meno_ex02_72_0720s](#)

MMSE

[vr_crdstrex_ex02_3_0821s](#) [vr_ceradstr_ex02_3_0807s](#) [vr_mmse_ex09_1b_0943s](#) [vr_mmse_ex32_0_0945s](#)

Rheumatic Heart Disease

[rhd0_9s](#)

ICD Codes

[icd0_19s](#)

Dementia

[vr_npka_1978_0_0872s](#) [vr_cogstadr_2014_m_0966s](#) [vr_demnp_2014_m_0968s](#) [vr_demne_2014_m_0967s](#)

Atrial Fibrillation

[vr_afafsr_2012_a_0970s](#) [vr_afcum_2016_a_1782s](#)

Exam Dates, Age, Sex

[ex0_7s](#) [ex0_8s](#) [ex0_9s](#) [ex0_10s](#) [ex0_11s](#) [ex0_12s](#) [ex0_13s](#) [ex0_14s](#) [ex0_15s](#) [ex0_16s](#) [ex0_17s](#) [ex0_18s](#) [ex0_19s](#) [ex0_20s](#) [ex0_21s](#) [ex0_22s](#) [ex0_23s](#) [ex0_24s](#) [ex0_25s](#) [ex0_26s](#) [ex0_27s](#) [ex1_1s](#) [ex1_2s](#) [ex1_3s](#) [ex1_4s](#) [ex1_5s](#) [ex1_6s](#) [ex1_7s](#) [ex3_1s](#) [birthyr_all](#)
[vr_ctdates_2011_m_0715s](#) [vr_dates_2014_a_0912s](#) [vr_survf_2014_a_0987s](#)

Foot Frequency with Derived Variables

[vr_ffreq_ex01_3_0987s](#) [vr_ffreq_ex08_1_0615s](#) [vr_ffreq_ex02_3_0713s](#) [ffreq0_20s](#) [ffreq0_21s](#) [ffreq0_22s](#) [ffreq1_5s](#) [ffreq1_6s](#) [ffreq1_7s](#) [vr_dgai2010_ex07_1_1108s](#) [vr_dgai2010_ex08_1_1009s](#) [vr_dgai2010_ex05_1_1013s](#) [vr_dgai2010_ex01_3_1078s](#) [vr_dgai2010_ex02_3_0996s](#)

Cancer

[vr_cancer_2013_a_0018s](#)

Diabetes

[vr_diab_ex02_3b_0388s](#) [vr_diab_ex09_1_1002s](#) [vr_diab_ex28_0_0601s](#)

Cardiovascular Procedures

[cbg_2007s](#) [vr_cvdroc_2016_a_1028s](#)

Survival

[vr_survcvd_2014_a_1023s](#) [vr_survdth_2014_a_1025s](#) [vr_survstk_2014_a_1031s](#) [vr_survstkt_2014_a_1030s](#) [vr_survf_2014_a_0987s](#)

Endpoints: Cardiac/Cerebrovascular/Death

[vr_soepedvt_2012_m_0766s](#) [vr_chfnit_2013_a_0828s](#) [vr_vte_2014_a_0913s](#) [vr_survcvd_2014_a_1023s](#) [vr_survdth_2014_a_1025s](#) [vr_soefsr_2014_a_1027s](#) [vr_survstk_2014_a_1031s](#) [vr_soef_2016_a_1073s](#) [vr_soefch_2016_a_1070s](#)

Stroke Related

[psipl_2003s](#) [pslpr_2003s](#) [vr_survstc_2014_a_1031s](#)

Bone Related

[foapain_200s](#) [vr_fxrev_2011_0_0613s](#) [vr_pase_ex02_3_0642s](#) [vr_fxrev_2012_0_0746s](#) [vr_fxrev_2013_3_0663s](#) [vr_fxrev_2013_1_0847s](#)

Commonly Used Risk Factors (Workthru)

[vr_wkthru_ex02_3b_0464s](#) [vr_wkthru_ex09_1_1001s](#) [vr_wkthru_ex32_0_0997s](#)

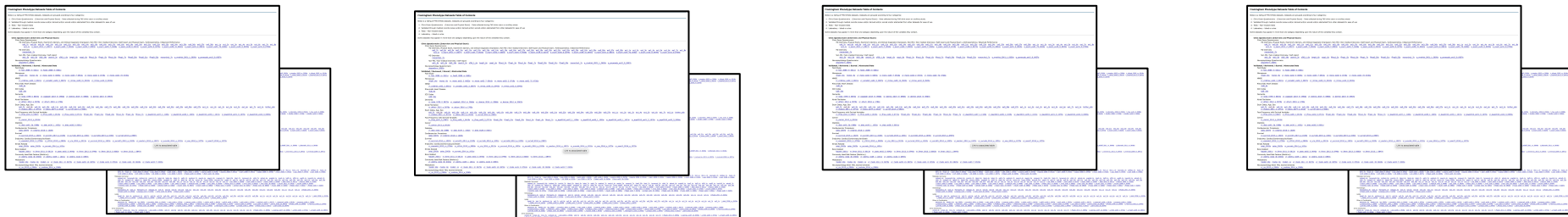
Medications

[meds0_28s](#) [meds1_8s](#) [meds3_1s](#) [vr_meds_2011_m_0675s](#) [vr_meds_ex09_1b_0879s](#) [vr_meds_ex31_0_0763s](#) [vr_meds_ex01_3b_0825s](#) [vr_meds_ex03_7_0535s](#)

Neuropsychology Brain MRI, Scored Variables

[vr_np_2013_a_0960s](#) [vr_npdates_2014_a_0962s](#)

[Link to associated table](#)



Consent group 1

Consent group 2

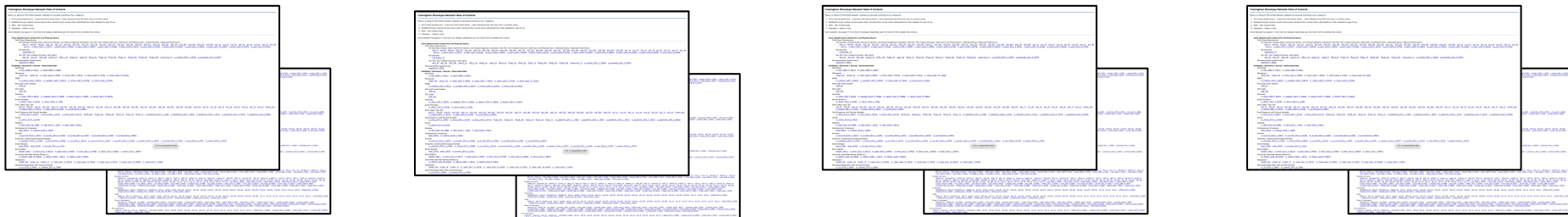
Consent group 3

Consent group 4



Investigator access **FILES** based on study and consent groups per study.
Then he needs to decrypt the files and **COMBINE** them to run any analysis

On a dbGAP authorized project an investigator may have access to consents 1 and 2
and on an other dbGAP project he may have access to consents 2,3 and 4



Consent group 1

Consent group 2

Consent group 3

Consent group 4

PIC-SURE API

R data frame or Python panda data frame



Via PIC-SURE API an Investigator access **VARIABLES** (and not **FILES**) based on study and consent groups per study.
Everything is **ALREADY COMBINED** them to run any analysis
He can SEARCH and RETRIEVE across all data he is authorized

On a dbGAP authorized project an investigator may have access to consents 1 and 2
and on an other dbGAP project he may have access to consents 2,3 and 4

Use Case 1

Using PIC-SURE to reproduce the ORCHID Study on
Seven Bridges

ORCHID Study Example

We have utilized PIC-SURE and Seven Bridges in BioData Catalyst to successfully reproduced the results and analysis of the following paper:

Outcomes Related to COVID-19 Treated with Hydroxychloroquine among In-patients with Symptomatic Disease (ORCHID) Study:



Research

Published online:
November 9, 2020

JAMA | **Original Investigation**

Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19 A Randomized Clinical Trial

Wesley H. Self, MD, MPH; Matthew W. Semler, MD; Lindsay M. Leither, DO; Jonathan D. Casey, MD, MSc; Derek C. Angus, MD, MPH; Roy G. Brower, MD; Steven Y. Chang, MD, PhD; Sean P. Collins, MD; John C. Eppensteiner, MD; Michael R. Filbin, MD; D. Clark Files, MD; Kevin W. Gibbs, MD; Adit A. Ginde, MD, MPH; Michelle N. Gong, MD, MS; Frank E. Harrell Jr, PhD; Douglas L. Hayden, PhD; Catherine L. Hough, MD, MSc; Nicholas J. Johnson, MD; Akram Khan, MD; Christopher J. Lindsell, PhD; Michael A. Matthay, MD; Marc Moss, MD; Pauline K. Park, MD; Todd W. Rice, MD; Bryce R. H. Robinson, MD, MS; David A. Schoenfeld, PhD; Nathan I. Shapiro, MD, MPH; Jay S. Steingrub, MD; Christine A. Ulysse, MS; Alexandra Weissman, MD, MPH; Donald M. Yealy, MD; B. Taylor Thompson, MD; Samuel M. Brown, MD, MS; for the National Heart, Lung, and Blood Institute PETAL Clinical Trials Network



Self WH, Semler MW, Leither LM, et al. **Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19: A Randomized Clinical Trial.** JAMA. 2020;324(21):2165–2176. doi:10.1001/jama.2020.22240

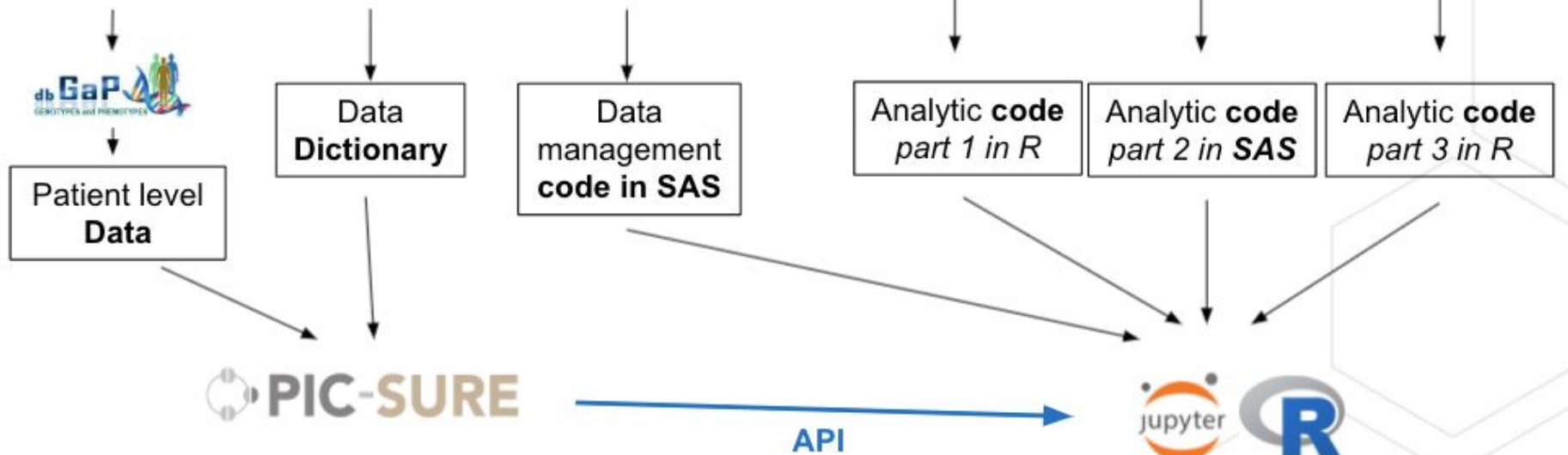
Research

JAMA | Original Investigation

Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19

A Randomized Clinical Trial

Wesley H. Self, MD, MPH; Matthew W. Serites, MD; Lindsay M. Letterer, DO; Jonathan D. Casey, MD, MS; Derek C. Angus, MD, MPH; Roy G. Brower, MD; Steven Y. Chang, MD, PhD; Sean P. Collins, MD; John C. Eppes, MD; Michael R. Filbin, MD; D. Clark Files, MD; Kevin W. Giblin, MD; Adri A. Ginde, MD, MPH; Michelle N. Gong, MD, MS; Frank E. Harrell Jr, PhD; Douglas L. Hayden, PhD; Catherine L. Hough, MD, MS; Nicholas J. Johnson, MD; Akram Khan, MD; Christopher J. Lindell, PhD; Michael A. Matthay, MD; Marc Moss, MD; Pauline K. Park, MD; Todd W. Rice, MD; Bryce R. H. Robinson, MD, MS; David A. Schoenfeld, PhD; Nathan I. Shapiro, MD, MPH; Jay S. Shengrub, MD; Christine A. Ulysse, MS; Alexandra Weisman, MD, MPH; Donald M. Yealy, MD; B. Taylor Thompson, MD; Samuel M. Brown, MD, MS, for the National Heart, Lung, and Blood Institute-PETAL Clinical Trials Network



ORCHID Study RStudio Example Available in BioData Catalyst Powered by Seven Bridges

ORCHID_COVID19.Rmd

```
1 |---
2 |title: An R Markdown document converted from "Access-to-Data-using-PIC-SURE-API/NHLBI_BioData_Catalyst/R/ORCHID_COVID19.ipynb"
3 |output: html_document
4 |---
5 |
6 |# ORCHID Clinical Trial: statistical analysis reproduction
7 |
8 |# Version 1.0
9 |
10 |This notebook reproduces the statistical analysis of the ORCHID clinical trial. Results have been published to JAMA, on November 7th 2021: ["Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19"](https://jamanetwork.com/journals/jama/fullarticle/2772922). The statistical analysis plan can be found on [clinicaltrials.gov](https://clinicaltrials.gov/ct2/show/NCT04332991?term=orchid&cond=Covid19&cntry=US&draw=2&rank=1).
11 |
12 |The clinical trial has been conducted between April and July 2020, and stopped before enrollment completion for futility, finding no difference of efficacy between hydroxychloroquine and placebo. This notebook is a reproduction of the clinical trial results based on the clinical trial protocol and the investigators original source code.
13 |
14 |## Requirements
15 |
16 |*This notebook has been tested to work with R version 4.0.0*. Below is the output of the sessionInfo() function:
17 |...
18 |R version 4.0.0 (2020-04-24)
19 |Platform: x86_64-pc-linux-gnu (64-bit)
20 |Running under: Ubuntu 18.04.4 LTS
21 |
22 |Matrix products: default
23 |BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/libopenblas-r0.2.20.so
24 |
25 |locale:
26 |[1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C                LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8    LC_MONETARY=en_US.UTF-8   LC_MESSAGES=C
27 |[7] LC_PAPER=en_US.UTF-8      LC_NAME=C                   LC_ADDRESS=C               LC_TELEPHONE=C           LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
28 |
29 |attached base packages:
30 |[1] stats      graphics  grDevices  utils      datasets  methods   base
31 |
32 |other attached packages:
33 |[1] BiocManager_1.30.10
34 |
35 |loaded via a namespace (and not attached):
1 |1 | An R Markdown document converted from "Access-to-Data-using-PIC-SURE-API/NHLBI_BioData_Catalyst/R/ORCHID_COVID19.ipynb" | R Markdown |
```

Environment | History | Connections

Global Environment

Data	
admission_table	List of 3
baseline_table	List of 3
comorbidity_table	List of 3
coos_df	3353 obs. of 4 variables
coxph_death	List of 21
death_plot	List of 3
demographics_tab_	List of 3
df_quartiles	6 obs. of 6 variables

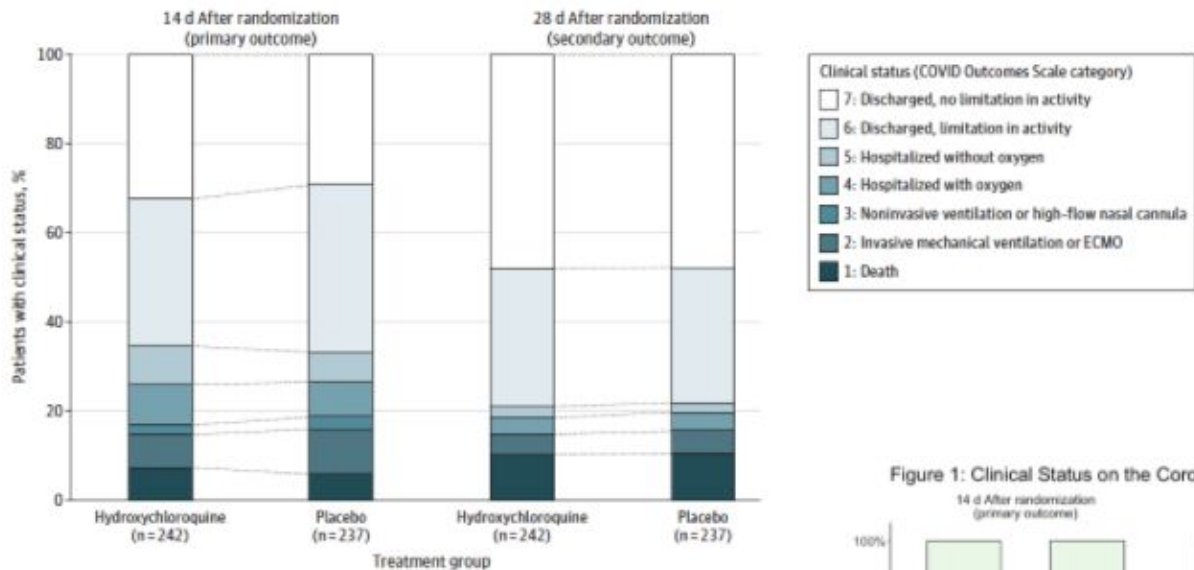
Files | Plots | Packages | Help | Viewer

New Folder | Upload | Delete | Rename | More

sbgenomics > workspace

Name	Size	Modified
..		
.RData	97.4 MB	Oct 1, 2021, 1:06 PM
.Renviron	40 B	Oct 1, 2021, 1:06 PM
.Rhistory	19.9 KB	Oct 1, 2021, 1:07 PM
.Rprofile	48 B	Oct 1, 2021, 1:07 PM
1_PICSURE_API_101.Rmd	17.6 KB	Oct 1, 2021, 1:06 PM
2_HarmonizedVariables_analysis.Rmd	7.9 KB	Oct 1, 2021, 1:06 PM
4_Genomic_Queries.Rmd	14.4 KB	Oct 1, 2021, 1:06 PM
5_LongitudinalData.Rmd	9.5 KB	Oct 1, 2021, 1:06 PM
6_Sickle_Cell.Rmd	13.4 KB	Oct 1, 2021, 1:06 PM
install_packages.R	2 KB	Oct 1, 2021, 1:06 PM
ORCHID_COVID19.Rmd	41.7 KB	Oct 1, 2021, 1:06 PM
PheWAS.Rmd	14.5 KB	Oct 1, 2021, 1:07 PM
R_lib		
Rstudio_lib		
userLibrary		
token.txt	310 B	Oct 1, 2021, 1:14 PM

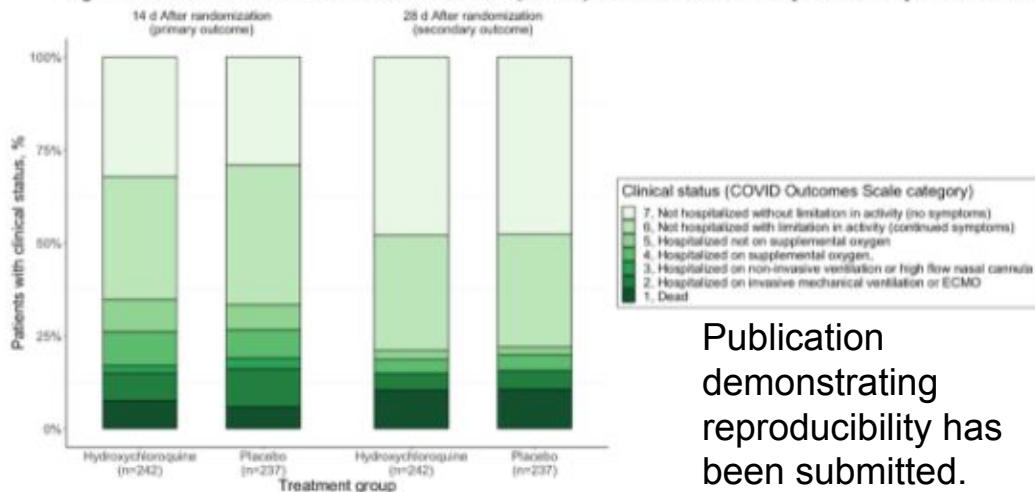
Figure 2. Clinical Status on the Coronavirus Disease (COVID) Outcomes Scale 14 Days and 28 Days After Randomization



JAMA[®]

Published online:
November 9, 2020

Figure 1: Clinical Status on the Coronavirus Disease (COVID) Outcomes Scale 14 Days and 28 Days After Randomization



Publication demonstrating reproducibility has been submitted.

NIH National Heart, Lung, and Blood Institute | BioData CATALYST
Powered by PIC-SURE

December 9, 2020

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

Use Case 2

Using PIC-SURE to reproduce and expand analysis of the HCT for SCD Study on Terra

Hematopoietic Cell Transplant for Sickle Cell Disease Study Use Case

- Collaborated with a Sickle Cell Disease (SCD) researcher to use PIC-SURE and Terra to conduct an analytic research study.
- Introduced researcher to BioData Catalyst to use tools in PIC-SURE and Terra to build upon their existing work
- Created a jupyter notebook using the PIC-SURE API to build a cohort and perform analysis in Terra
 - Extracted the data dictionary
 - Built queries to retrieve data
 - Successfully tested reproducibility and validated findings of original study in BDCatalyst
 - Conducted an additional analysis using the PIC-SURE API and Terra to produce new research findings
- Manuscript in preparation

HCT for SCD Study Example Available in BioData Catalyst Powered by Terra

NIH National Heart, Lung, and Blood Institute | BioData CATALYST Powered by Terra | BETA WORKSPACES | Workspaces > biodata-catalyst/BioData Catalyst PIC-SURE API R Examples > notebooks > 6_Sickle_Cell.ipynb | Cloud Environment Stopped (← \$0.01/hr) | PREVIEW (READ-ONLY) | EDIT | PLAYGROUND MODE | X

PIC-SURE API use-case: quick analysis on Hematopoietic Cell Transplant for Sickle Cell Disease (HCT for SCD) data

This is a tutorial notebook aimed to get the user quickly up and running with the R PIC-SURE API. It covers the main functionalities of the API.

PIC-SURE R API

What is PIC-SURE?

As part of the BioData Catalyst initiative, the Patient Information Commons Standard Unification of Research Elements (PIC-SURE) platform has been integrating clinical and genomic datasets funded by the National Heart Lung and Blood Institute (NHLBI).

Original data exposed through the PIC-SURE API encompasses a large heterogeneity of data organization underneath. PIC-SURE hides this complexity and exposes the different study datasets in a single tabular format. By simplifying the process of data extraction, it allows investigators to focus on the downstream analyses and to facilitate reproducible sciences.

More about PIC-SURE

The API is available in two different programming languages, python and R, enabling investigators to query the databases the same way using either language.

PIC-SURE is a larger project from which the R/python PIC-SURE API is only a brick. Among other things, PIC-SURE also offers a graphical user interface that allows researchers to explore variables across multiple studies, filter patients that match criteria, and create cohorts from this interactive exploration.

The R API is actively developed by the Avillach Lab at Harvard Medical School.

PIC-SURE API GitHub repo:

- <https://github.com/hms-dbmi/pic-sure-r-adapter-hpds>
- <https://github.com/hms-dbmi/pic-sure-r-client>
- <https://github.com/hms-dbmi/pic-sure-biodatacatalyst-r-adapter-hpds>



Genomic
Information
Commons

U01



National Center
for Advancing
Translational Sciences





Central GRIN Access

Please click one of the buttons below to log in.



Discover Portal



Analysis Portal



Investigators

PIC-SURE API



The Genomics Research and Innovation Network.
Genet Med. 2019 Sep 4

5,454,487

Patients



BCH
CCHMC
CHOP

2,886,837
1,188,661
1,378,989

140,218

Biosamples



BCH
CCHMC
CHOP

45,230
93,461
1,527

[More Information](#)

QUERY BUILDER

ACT_Demographics

Sex, Restrict By Value Female

back

delete

edit

AND

Gene_with_variant

Variant Info Column Gene_with_variant: GRIN2A

back

delete

edit

AND

Variant_severity

Variant Info Column Variant_severity: LOW

back

delete

edit

1,242
Patients

BCH	1,142
CCHMC	19
CHOP	81

3,391
Biosamples

BCH	3,215
CCHMC	28
CHOP	148

More Information

PIC-SURE

As a search tool across NCPI platforms

- **Any Clinical data** (EHR, Registries, clinical trials)
- **Any Sequencing data** (WES, WGS)
- Any biosamples
- **Any index files** (Radiology, EEG, etc...)

1) Centralized approach



Investigators

Aggregate Search with Open PIC-SURE

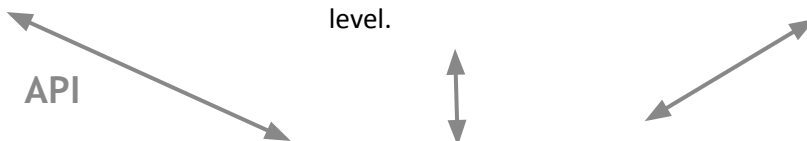
PIC-SURE Search in Authorized Access

Workspaces / Analysis Environments

Anyone can Explore aggregate data without authorization

Explore phenotypic and genomic data to refine cohorts, cohort creation at variable level.

Direct export with PIC-SURE API of selected cohort(s) to chosen analysis environment



PIC-SURE API
Application Programming Interface

PIC-SURE HPDS
High Performance Data Store

Patient level data



2) Federated approach



Investigators

Aggregate Search with Open PIC-SURE

PIC-SURE Search in Authorized Access

Workspaces / Analysis Environments

Anyone can Explore aggregate data without authorization

Explore phenotypic and genomic data to refine cohorts, cohort creation at variable level.

Direct export with PIC-SURE API of selected cohort(s) to chosen analysis environment

Patient level data stays in each platform

PIC-SURE API
Application Programming Interface



API



3) Mixed approach

Patient Data stays local / index is centralized



Investigators

Aggregate Search with Open PIC-SURE

PIC-SURE Search in Authorized Access

Workspaces / Analysis Environments

Anyone can Explore aggregate data without authorization

Explore phenotypic and genomic data to refine cohorts, cohort creation at variable level.

Direct export with PIC-SURE API of selected cohort(s) to chosen analysis environment

Patient level data stays in each platform

Index of all files centralized

PIC-SURE API
Application Programming Interface

PIC-SURE HPDS
High Performance Data Store



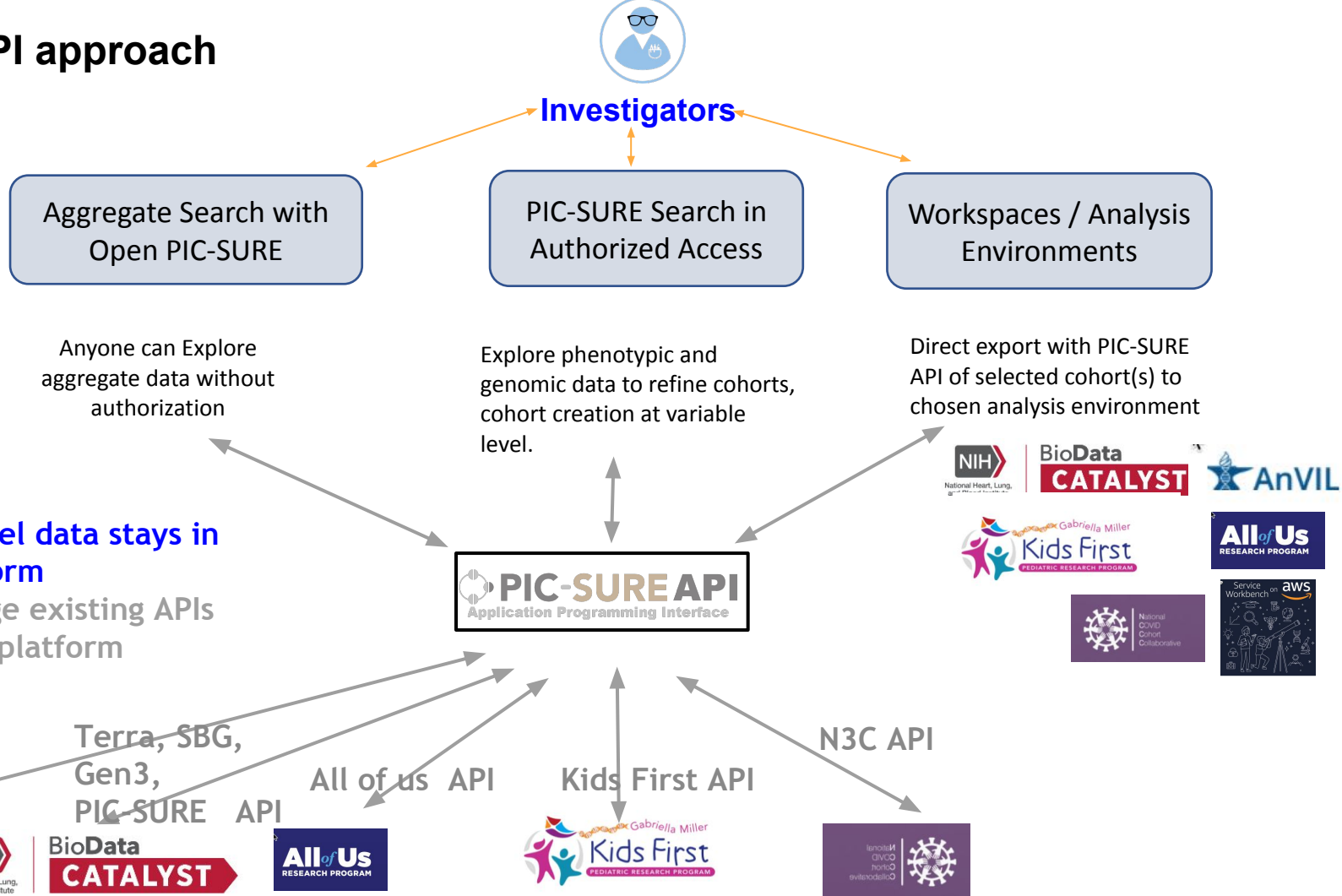
PIC-SURE is a Meta-API

It provides mechanisms to transform, modify, augment **existing APIs**, including itself. The main enabling mechanism is the **Resource** abstraction.

The **Aggregate Data Sharing Resource** acts as a filter to prevent sharing identifiable data.

The **Passthrough Resource** acts as a proxy authentication mechanism so users don't have to be created in two places.

4) Meta-API approach



- Patient level data stays in each platform
- We leverage existing APIs from each platform

Resources

[BioData Catalyst Powered by PIC-SURE User Guide](#)

[Access to PIC-SURE API GitHub repository](#)

[PIC-SURE YouTube channel](#)

[BioData Catalyst GitBook \(pending PIC-SURE updates\)](#)

[PIC-SURE API Documentation](#)



Synthesize Goals and Next Steps

for the next 6 Months, with focus on driving use cases

Stan Ahalt, Jon Kaltman



Goals and Next Steps (Ahalt)



Emerging common motif: importance of user-centered, user-friendly design & functionality

PFB:

1. Identify and document use cases that would result in “PFB-lite” v PFB
2. Differentiate utility of PFB/VDB/etc. vs FHIR
3. Clarify what PFB is/is not ([Glossary](#)). [[Full list here](#)]

FHIR:

1. Align on research study and metadata v1 representation (public data)
2. Identify roadmaps for platforms around services/use cases/limitations
3. Continue work on existing FHIR use cases [[full list here](#)]

RAS: Complete current plan and begin planning next phase:

1. Solve the challenges of milestone 3 (SSO, etc.) & meet the deadline
2. Plan beyond milestone 3: next steps proposal (milestone 4, passport partners expanded outreach) [[full list here](#)]



Goals and Next Steps (Ahalt)



End User Cloud Cost: Help users to adapt to new cloud reality through

1. Create free workspaces for training in the cloud
2. Budget templates & guides
3. End-to-end user stories generation
4. “Database” of cost modeling efforts across NCPI
5. Long term activities (e.g. NCPI codeathon) [[full list here](#)]

Search: Deploy user-centered thinking of Search

1. Form a Working Group that will drive the development of use- case driven Search strategy (e.g. develop personas, guide to existing searches/components, etc.)
2. Create a list of search components and documentation
3. Create a search taxonomy to inform a search roadmap
4. Respond to Search RFI
5. Define and promote semantic maturity in data to enable search [[full list here](#)]



Goals and Next Steps (Ahalt)



Other Interoperability Efforts: Engage users for

1. Testing of current functionality
2. Feedback re: new features
3. Development of users/use cases to drive new interop features,
4. Standardization of Tools/Apps deployment,
5. Development of methods to publish completed use cases (to replicate, train, etc) Development of training on interop methods [[full list here](#)]

GA4GH: Constantly developing new standards. NCPI members can participate by:

1. Getting engaged, and through coordinating our representation and interest in GA4GH across NCPI
2. Document the GA4GH standards in use across NCPI and identify future options
3. Collecting considerations for new standards to propose to GA4GH [[full info here](#)]

FYI: GA4GH Pedigree WG presents a new pedigree ontology (OWL) and a new pedigree model, and their implementations in FHIR on 10/12.



Goals and Next Steps (Ahalt)



Use Cases:

- Structure now in place to help with coordination and transparency, and extend utility!
- Very exciting to see both a) multiple mature use cases yielding fascinating science AND b) new use cases!
- Data can be called by DRS via distributed pipeline to understand sex as a biological variable
- Complementarity of DRS and FHIR
- Comparing algorithms across platforms to compare results. Continue work to ascertain reasons behind difference in results.
- Meta-API approach across NCPI: a) Use case development, b) where development gets done?

General:

- Remember that we are engaged in cultural change as well as technical changes
- Seek NCPI-wide opportunities to leverage program resources for max impact
- A lot of utility is possible now but in many cases we could use an “easy button”



NIH Closing Thoughts (Kaltman)





Meeting Deliverable: NCPI Glossary



- Remember to keep populating the NCPI Glossary with new words or additional definitions
- We hope this Glossary will be a concrete deliverable at the end of the meeting to help us coalesce around common definitions and/or highlight differences.

Thank you for attending!

Please take a moment to complete our [Workshop Evaluation Form](#)

See you in the Spring!