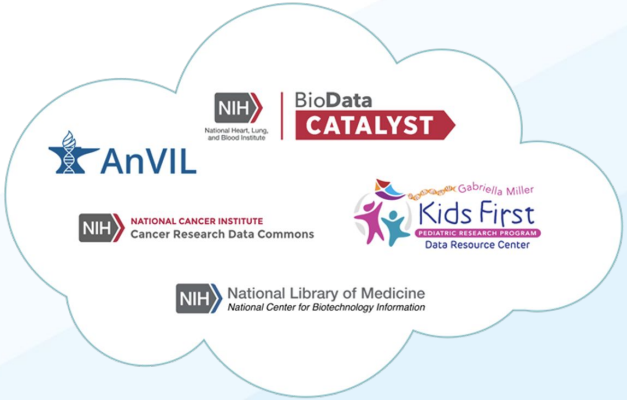
Welcome to Day 2...

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	@Patrick Patton @Paul Kerr @Allie Gartland Gray	@Joe Asare @Tom Madden @John Cheadle
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		
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.

Zoom download…. (Patton)

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.

Community Rules of Engagement (Ahalt)

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 •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

Day 2 Goals

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

PFB - Gaps and/or Key Blocks (Grossman)

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)

FHIR (Carroll)

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

RAS (O'Connor)

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)

RAS (O'Connor)

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

RAS (O'Connor)

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?

End-User Cloud Costs (Schatz)

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

Search (Rogers)

Gaps	Next Steps
Understanding cross-platform personas and use cases. **Finding Studies (priority order)** ● Understanding how the data is consented and how to apply for access. ● Searching over phenotype. ● Searching by experimental metadata. ● Searching by subject demographics. ● Determining if a given genotype is present in a given dataset before having access. **Building Cohorts (priority order)** ● Finding and gaining access to different search portals. ● Portals lacking “send to workspace env’ buttons to easily take search results to analysis platforms. Easy Retro Board	**Form a search working group and ...** ● Conduct UX research to determine personas and use cases for search from actual users. Determine who how to source users e.g. BDC Fellows. ● Create a list of search components and APIs used in the NCPI platforms, demonstrate how to use, and collect feedback. ● Create a search taxonomy to define the different kinds of search used/envisioned to inform an integrated search roadmap. ● Link back to studies in context from the NCPI dataset catalog. ● Generate input for the upcoming search RFI NOT-OD-21-187. ● Explain data consents. ● 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 an 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

Global Alliance for Genomics and Health (GA4GH)

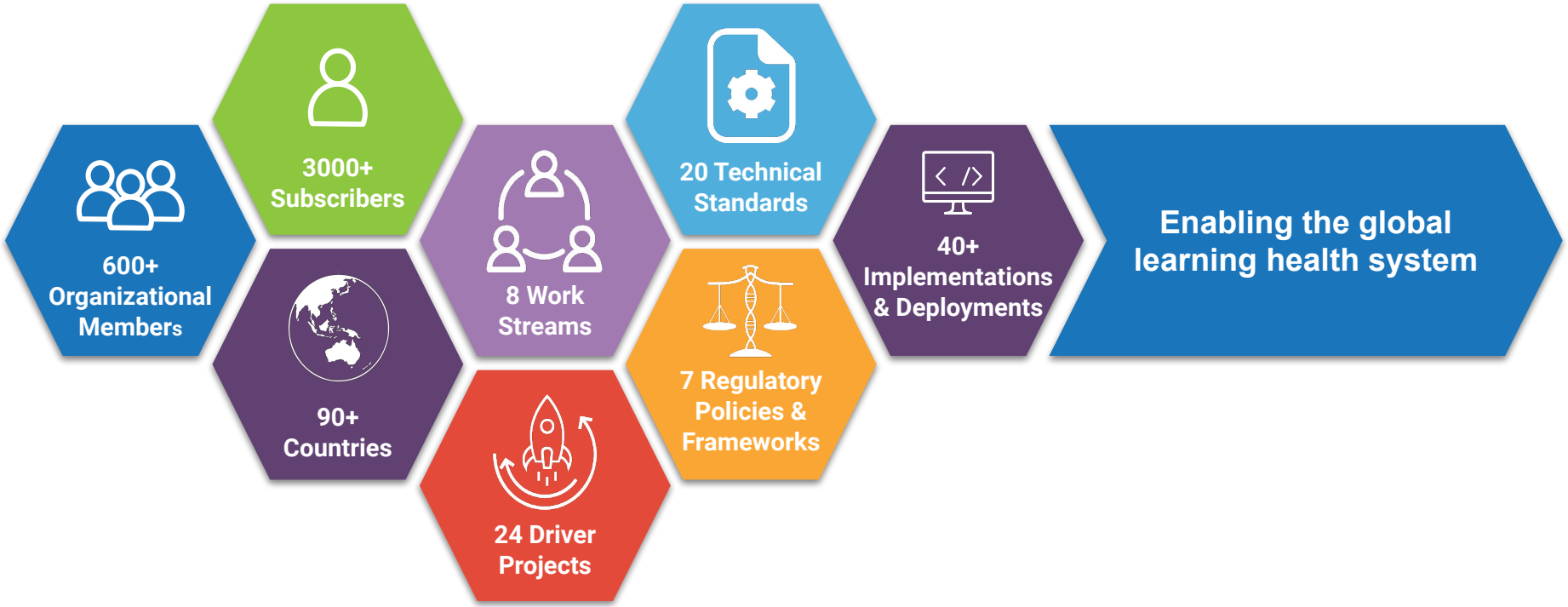
***Mission**: Enable genomic data sharing for the benefit of human health*

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.

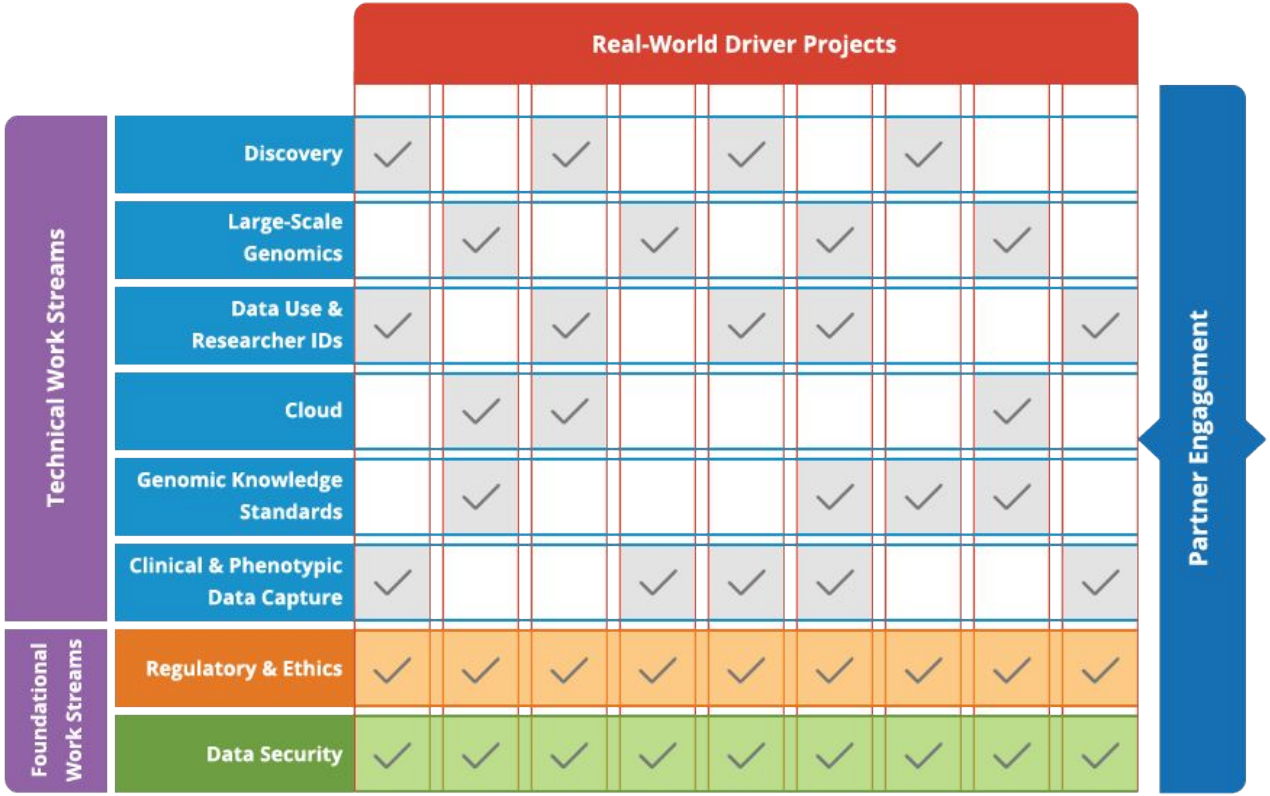
LEARN MORE

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



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.

NIH NATIONAL CANCER INSTITUTE

National Cancer Institute Cancer Research Data Commons (NCI CRDC)

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

National Cancer Institute Genomic Data Commons (NCI GDC)

NIH 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

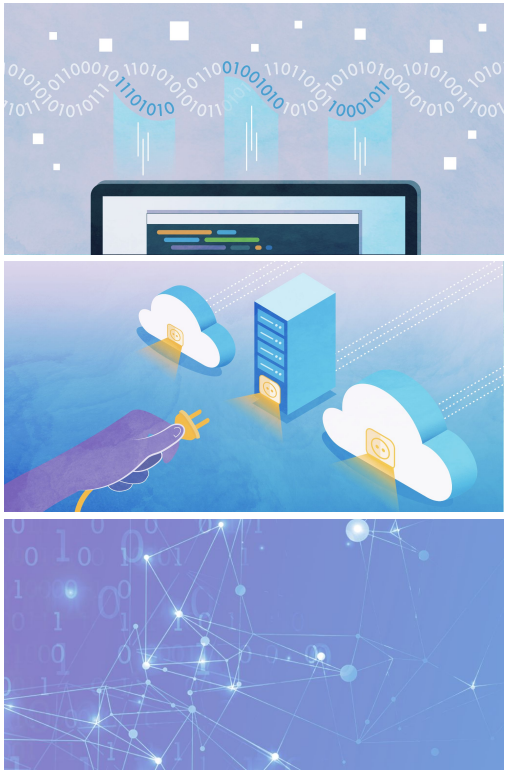
- *Others?*

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

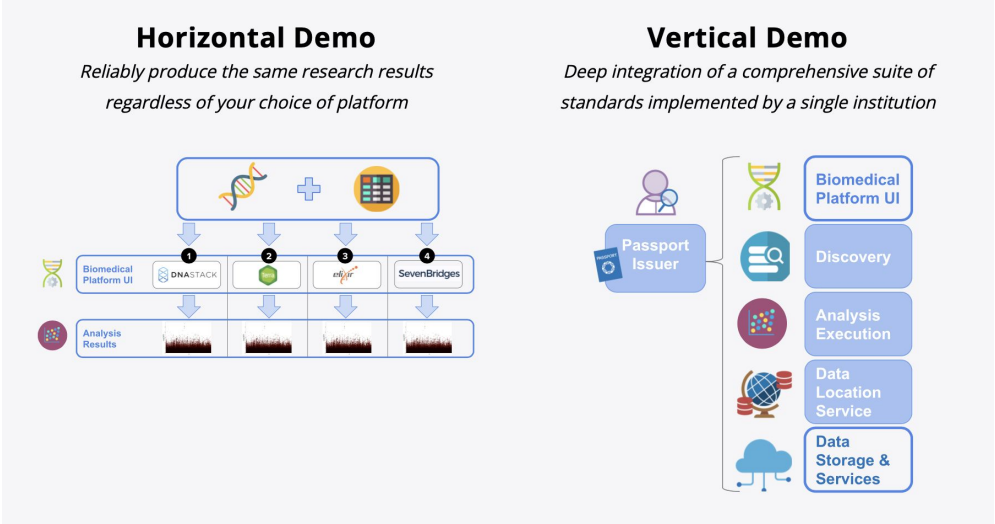
- Reference server implementation suite of GA4GH API specs (DRS & WES right now)
- Simplicity and versatility of setup
  - local laptop, HPC, cloud
- Technical on-ramp for:
  - Individuals new to GA4GH
  - Organizations exploring GA4GH on non-cloud native architectures
- Modular - Run APIs tailored to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



https://bit.ly/starterkit-slides

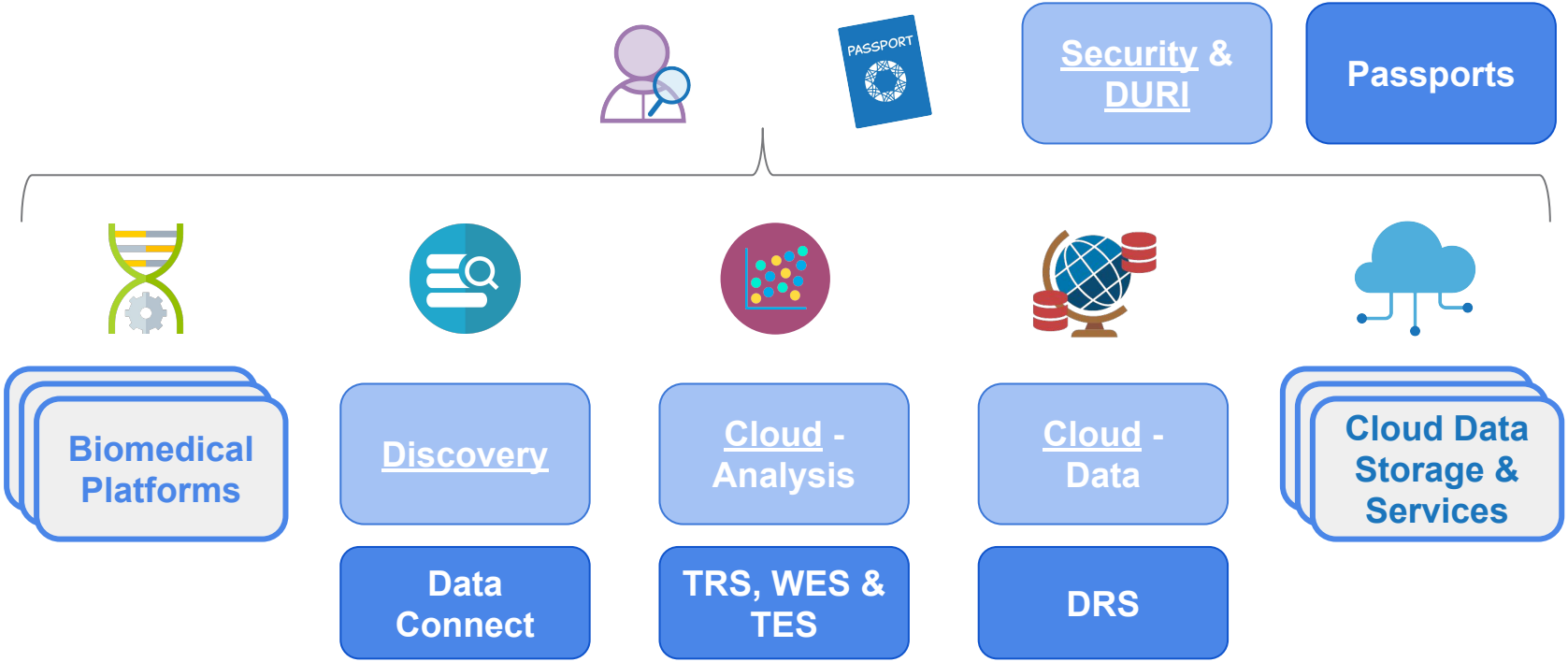
GA4GH FASP

- *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***



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

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

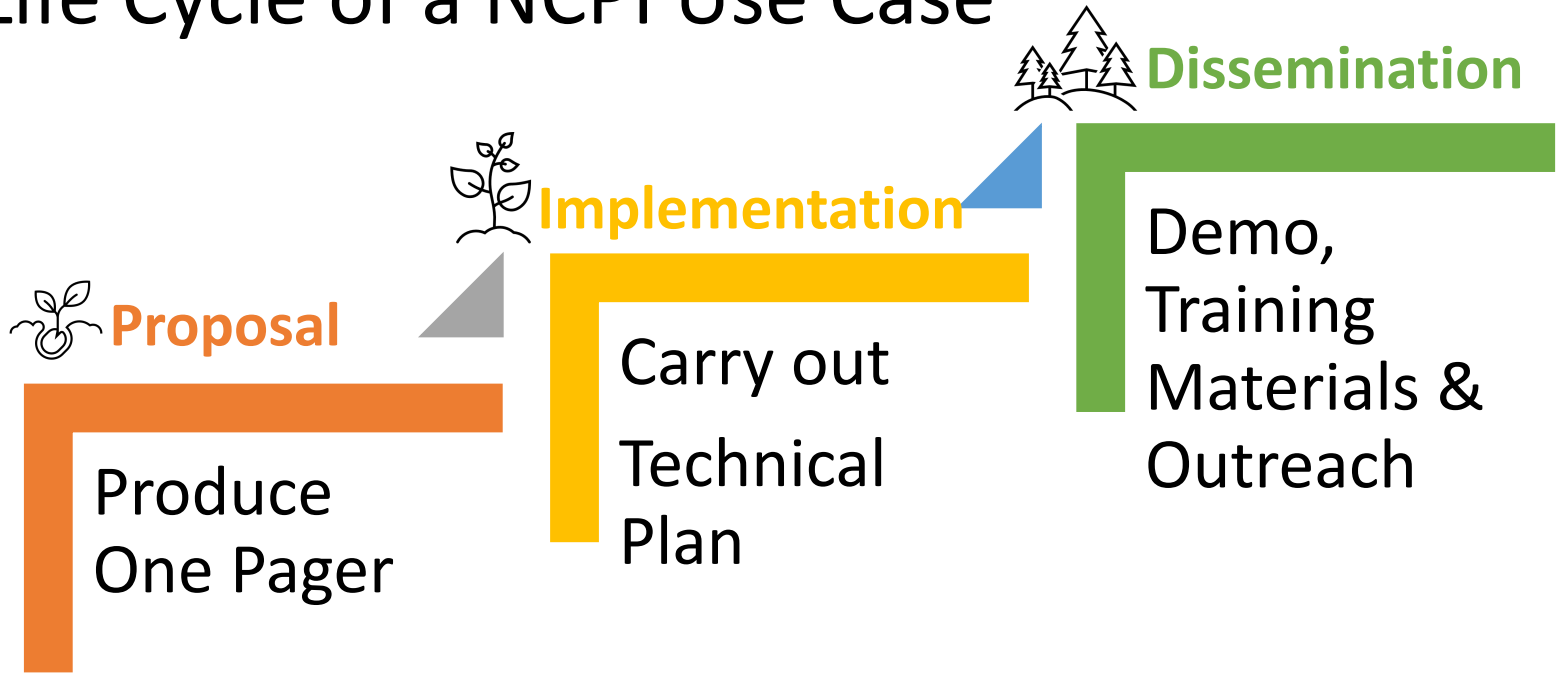
Oct 6, 2021

What is a NCPI Use Case?

- Access and integration of data from NCPI platforms (n>=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 Diseas

NCPI Use Case Details

Status: NCBI actively moving all files (BAMs) to hot AWS/SRA storage. Files become immedi
DRS as they are moved into S3. Next steps: Seven Bridges development work to obtain RAS
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

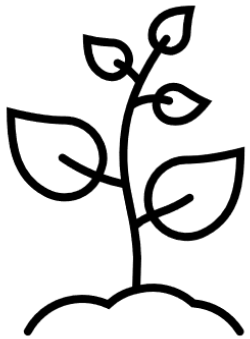
Summary:

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

One pager Example

https://github.com/NIH-NCPI/NCPI_use_case_tracker/issues/18

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

https://github.com/NIH-NCPI/NCPI_use_case_tracker/issues

- 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
  #15 opened on Jul 13 by jackDiGi | Ready to develop — 2
- UC 12 - (Xihong) Whole Genome Sequencing Association Analysis pipeline one pager done
  #12 opened on Jun 29 by NoopDog | On Hold — 3
- UC 11. (Wilson) Sex as a Biological Variable one pager done
  #11 opened on Jun 29 by NoopDog | Ready to develop — 3
- UC 10. SRA & Kids First DRC for Kids First & UDN co-analysis one pager done
  #10 opened on Jun 29 by NoopDog | Ready to develop — 2
- UC 9. Whole slide images need one pager
  #9 opened on Jun 29 by NoopDog | Ready to develop — 2
- UC 8. PIC-SURE API search of clinical and genomic data available from Seven Bridges Platform need one pager
  #8 opened on Jun 29 by NoopDog | Ready to develop — 2
- 7. NHGRI AnVIL + Kids First DRC + NHLBI BioData Catalyst need training material
  #7 opened on Jun 29 by NoopDog | Use Case Complete — 4
- UC 1a. NHLBI BioData Catalyst + Kids First DRC inactive
  #2 opened on Jun 29 by NoopDog | On Hold

Questions and Suggestions are welcomed!

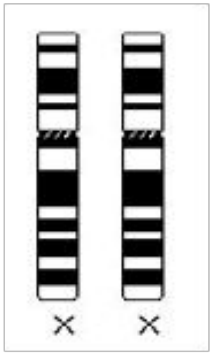
Team: Asiyah Lin, Dave Rogers, Jack DiGiovanna, Ken Wiley, Valerie Cotton, Valentina Di Francesco

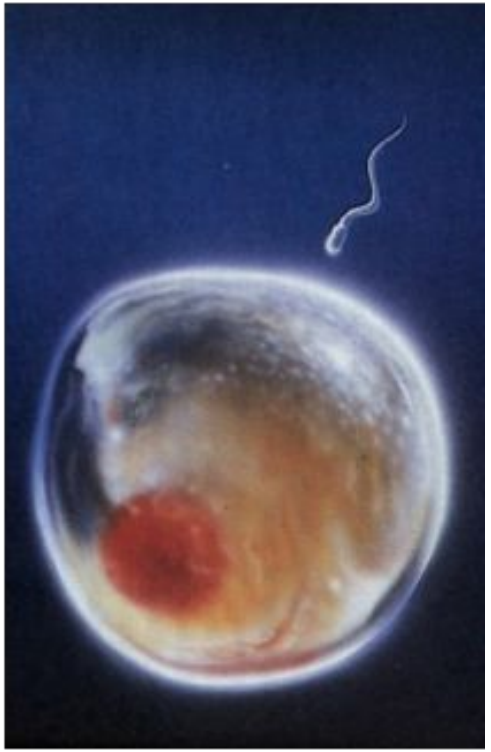
Genetic Sex as a Biological Variable and X-inactivation

Melissa Wilson, PhD
Arizona State University

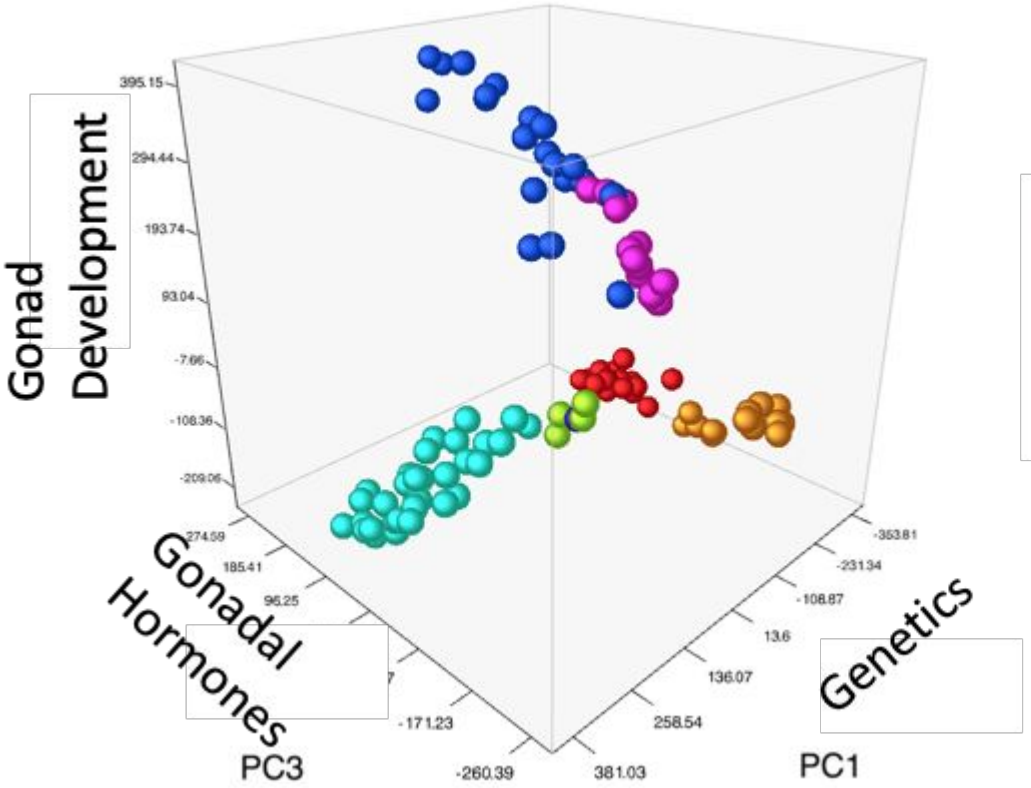
Language

- Genetics
- Gonads
  (& gonadal hormones)
- Gender





Sex differences are multidimensional



More than bimodal!

PERSPECTIVE | HUMAN GENOMICS

Searching for sex differences

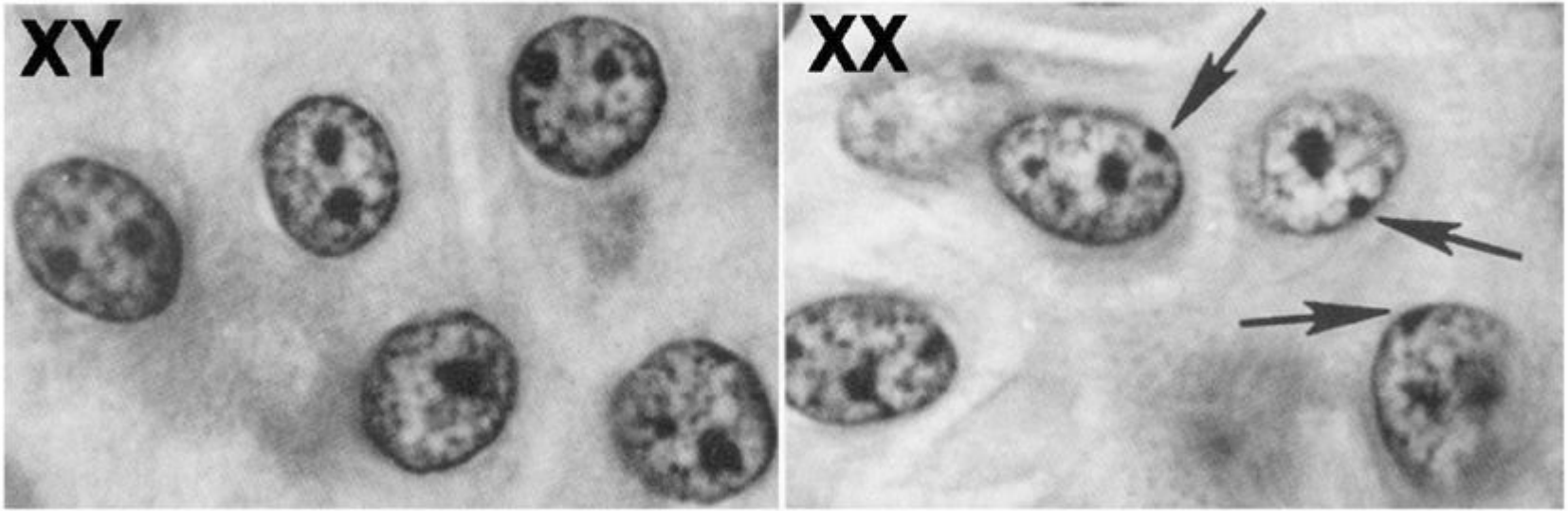
Melissa A. Wilson

+ 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

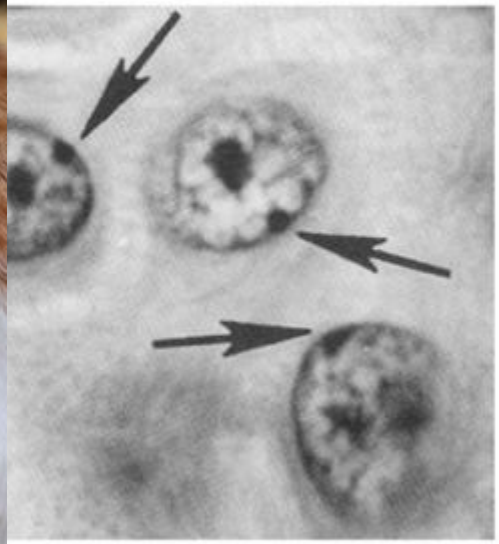


http://www.mun.ca/biology/scarr/Barr_Bodies.jpg

X-inactiva

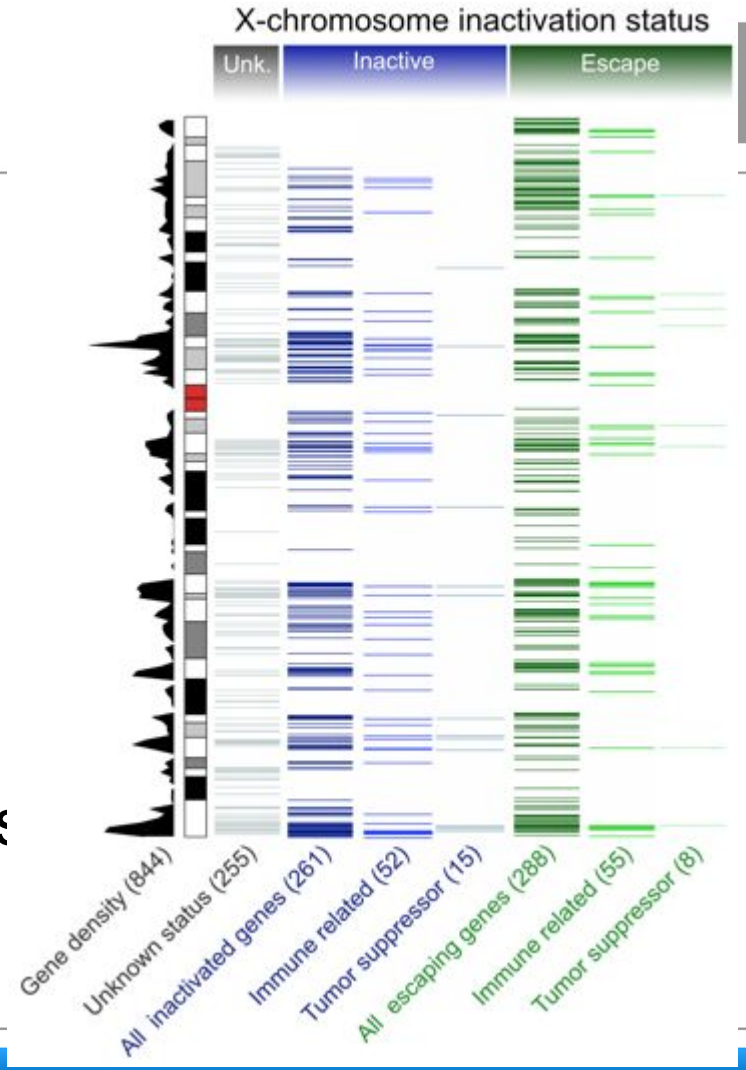
Barr body as seen un

http://www.mun.ca/biology/scarr/Barr_



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

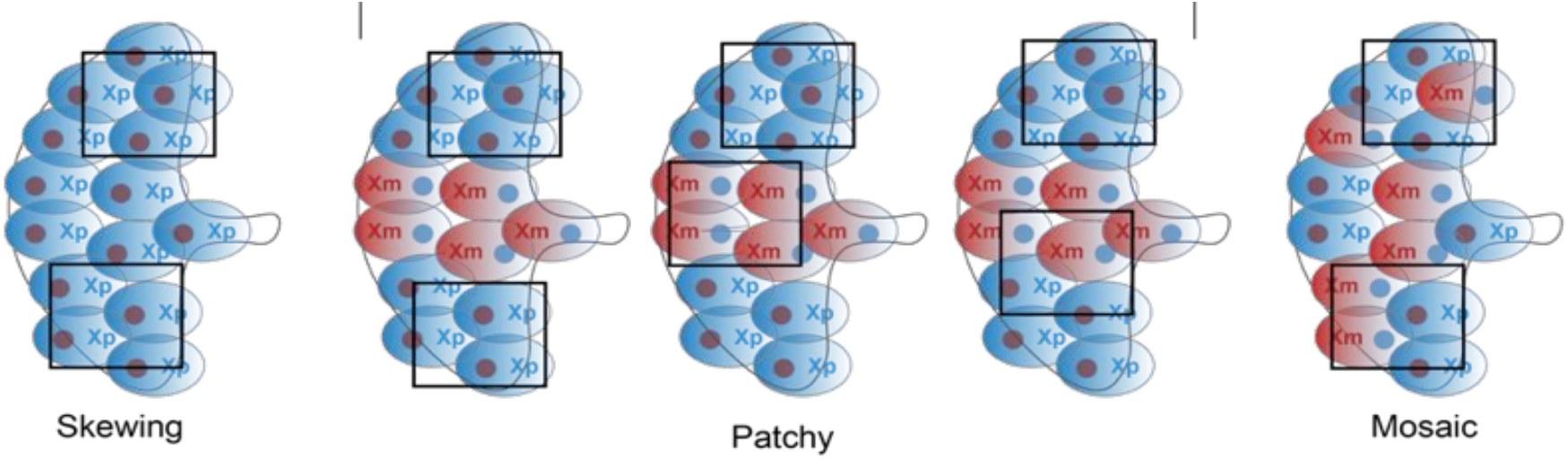
Tanya Phung

Kimberly Olney

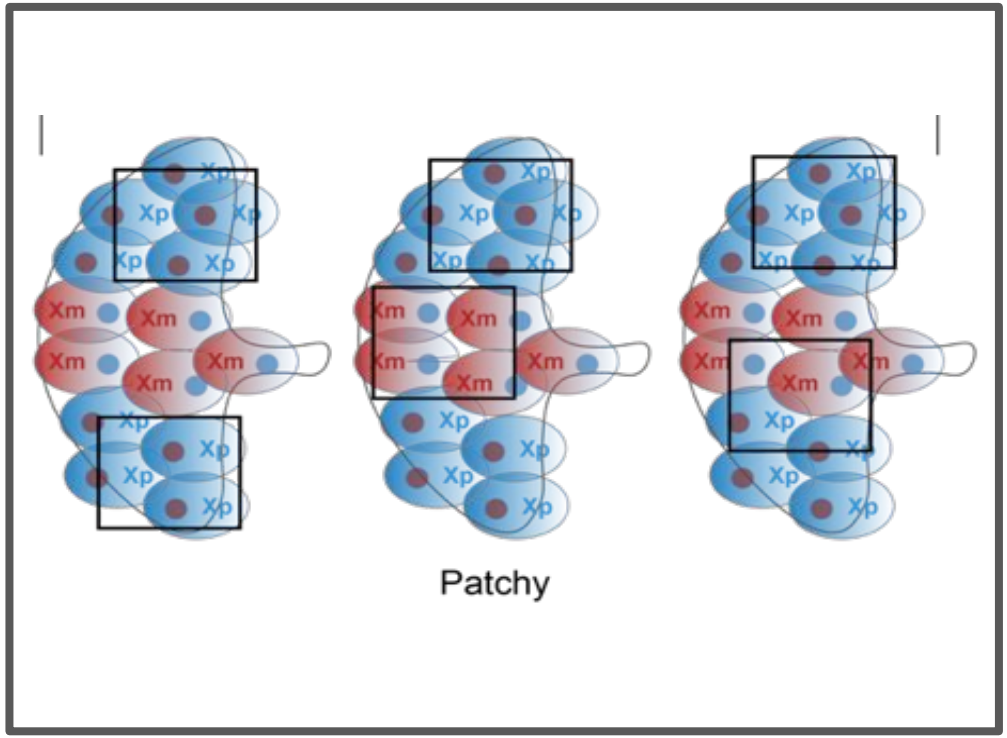
(Phung et al, submitted)

The placenta is the genotype of the offspring

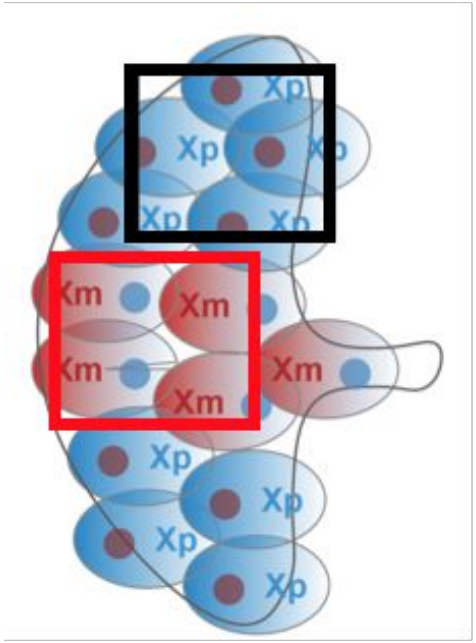
X-inactivation in the placenta



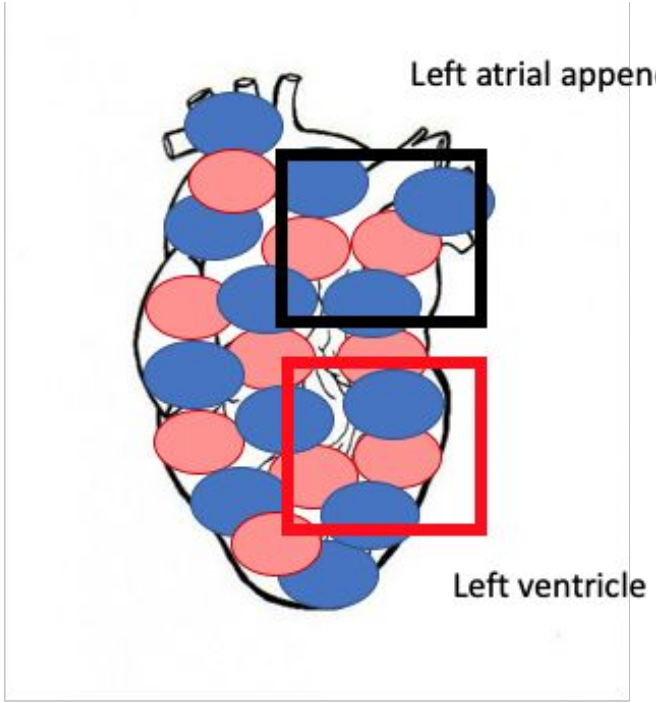
Patchy X-inactivation in the placenta



Placenta distinct from adult tissues



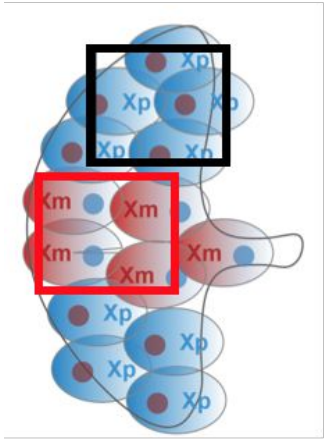
=/=



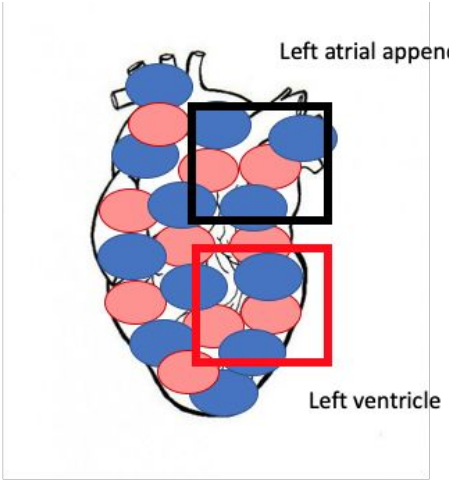
Heart data from GTEx consortium

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?

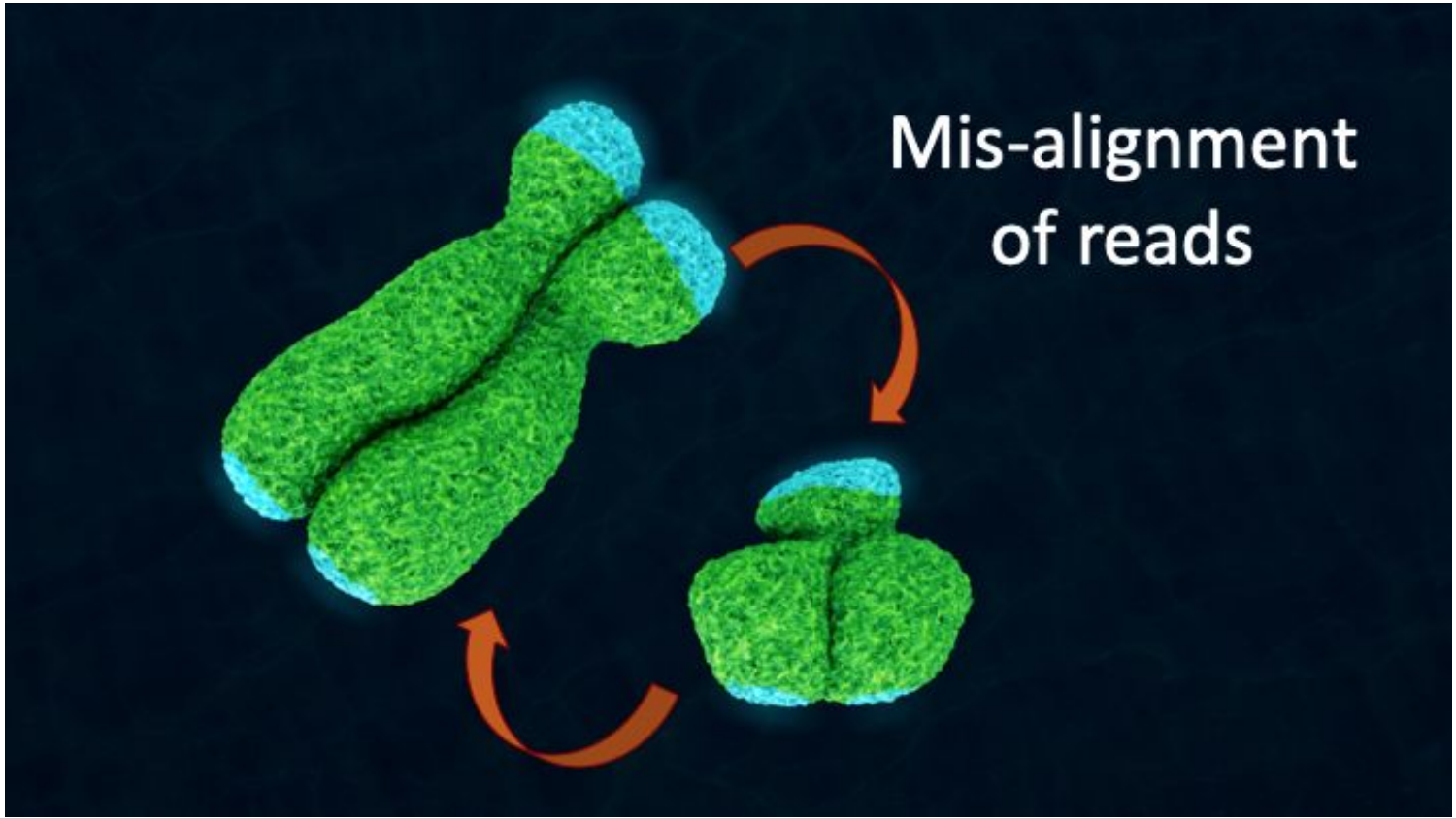


=/=

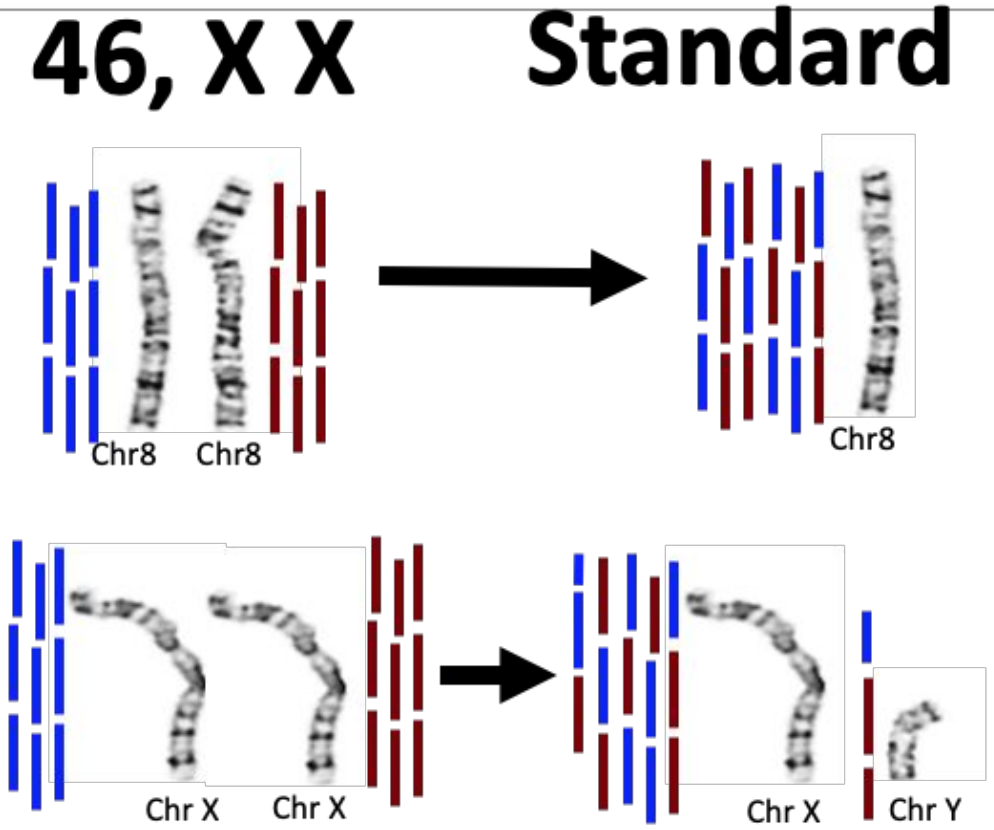


Heart data from GTEx consortium

Sex chromosomes are unique



Sex Chromosomes mis-mapping



Sex chr complement reference

XYalign

github.com/SexChrLab/XYalign

Infer sex chromosome complement

Output in user-defined windows (all chr):

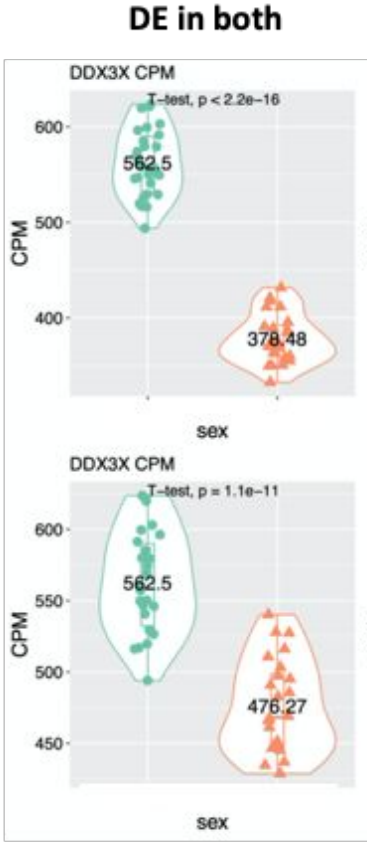
- Quality
- Depth
- Allele-balance

Realign with appropriate sex chr masks

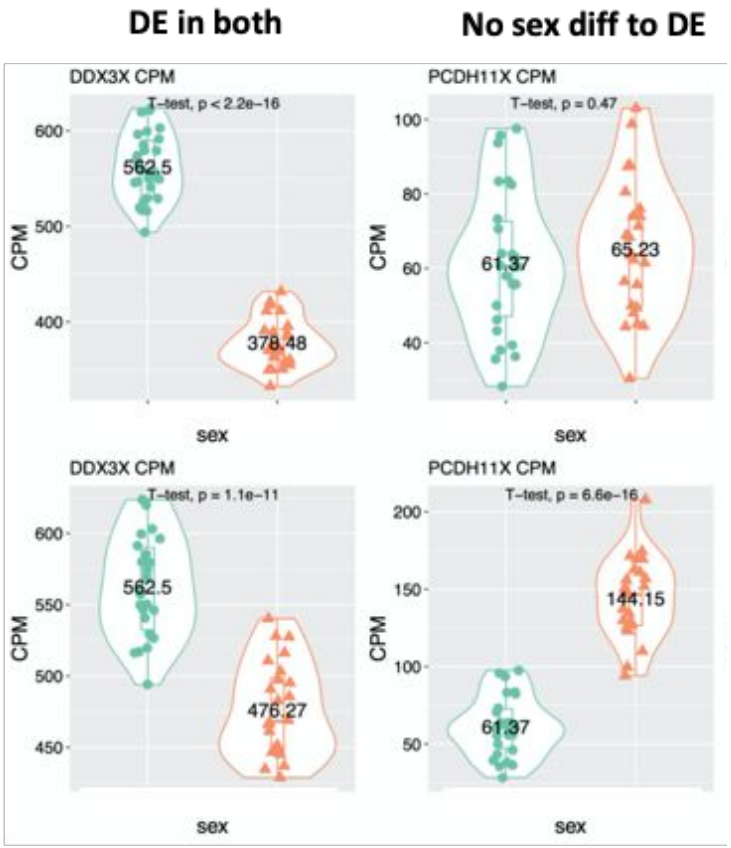
Timothy Webster

(Webster et al, 2019)

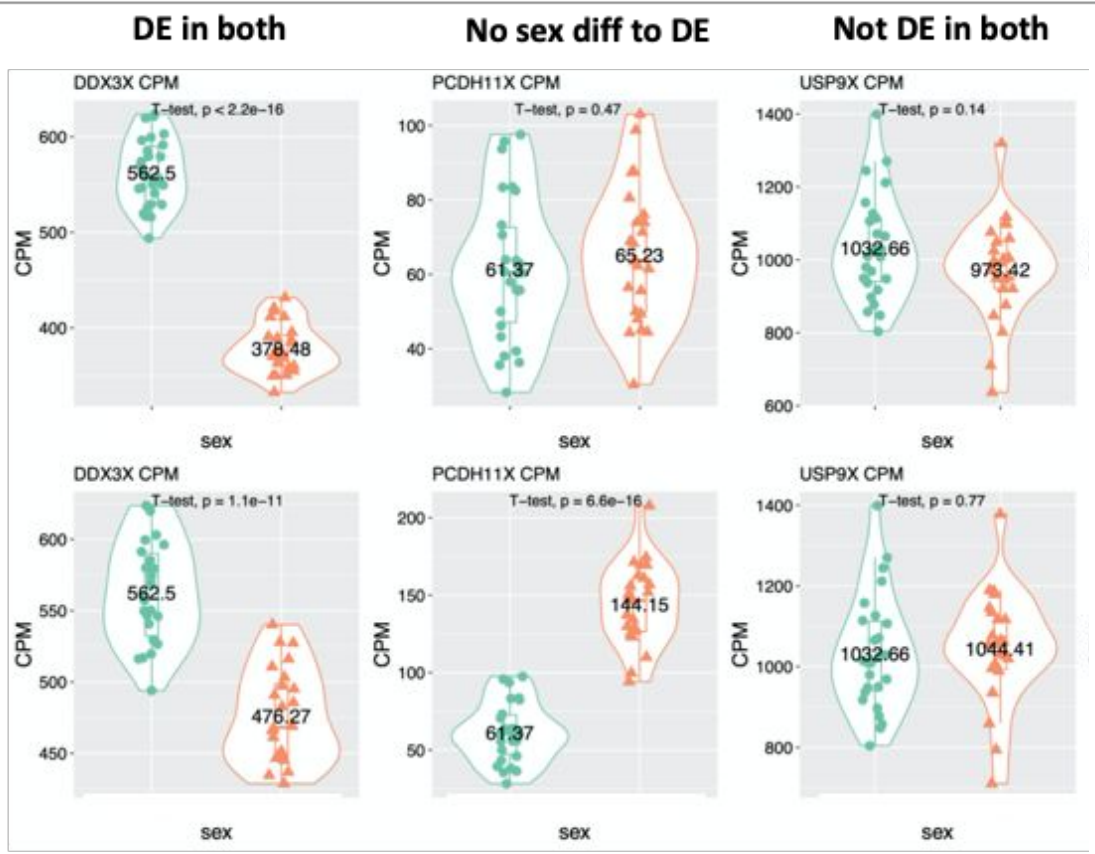
Mapping matters



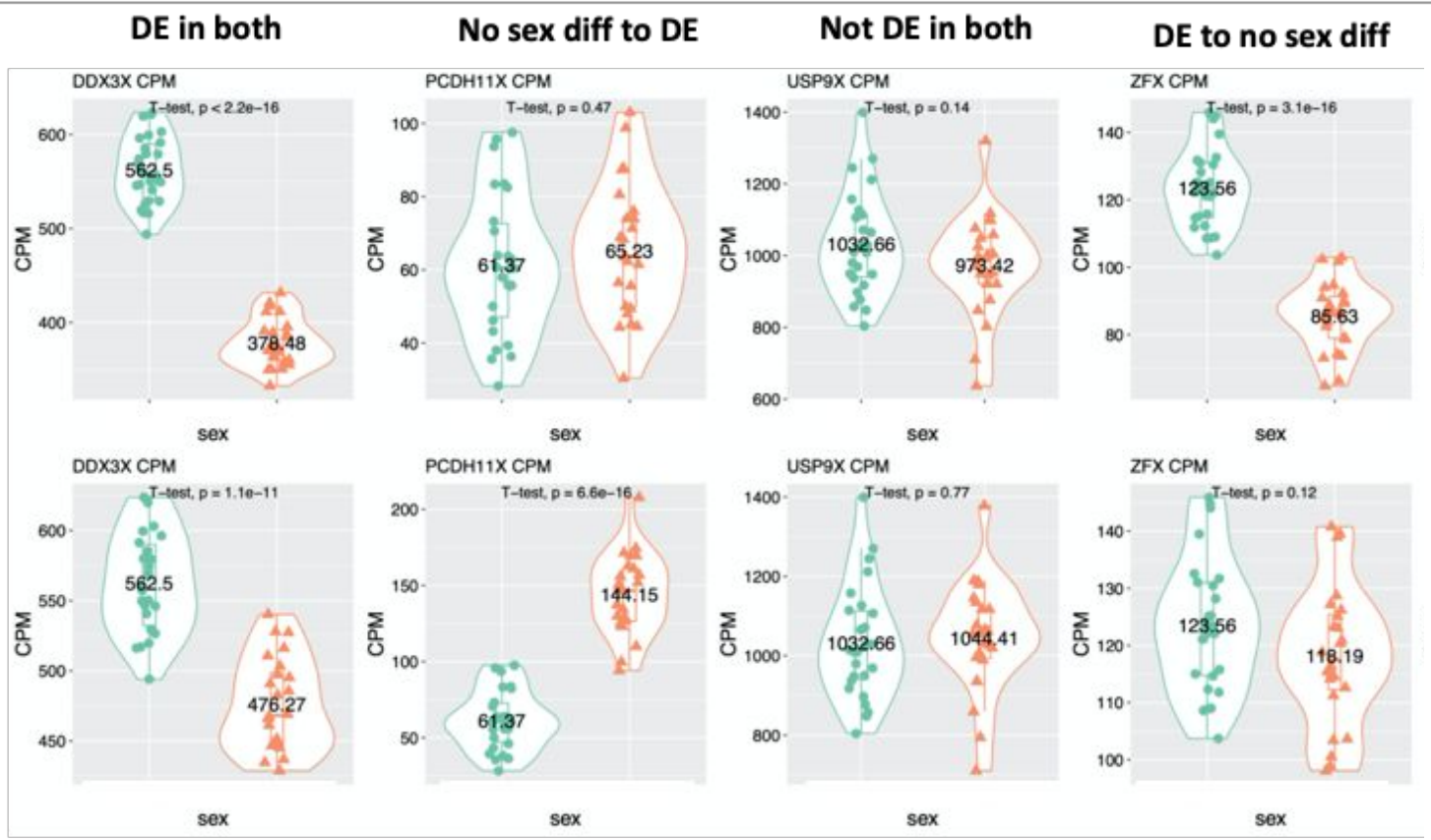
Mapping matters



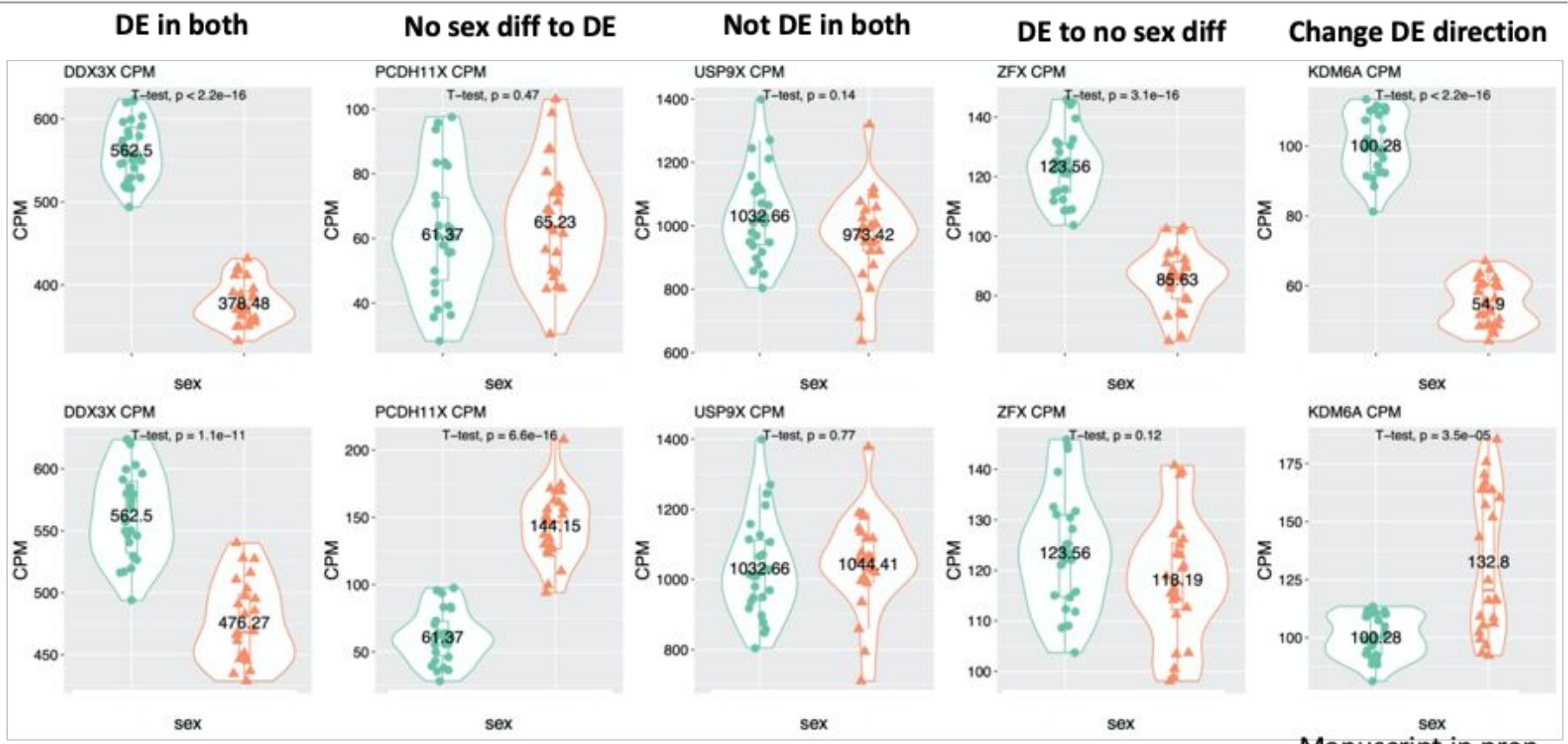
Mapping matters



Mapping matters

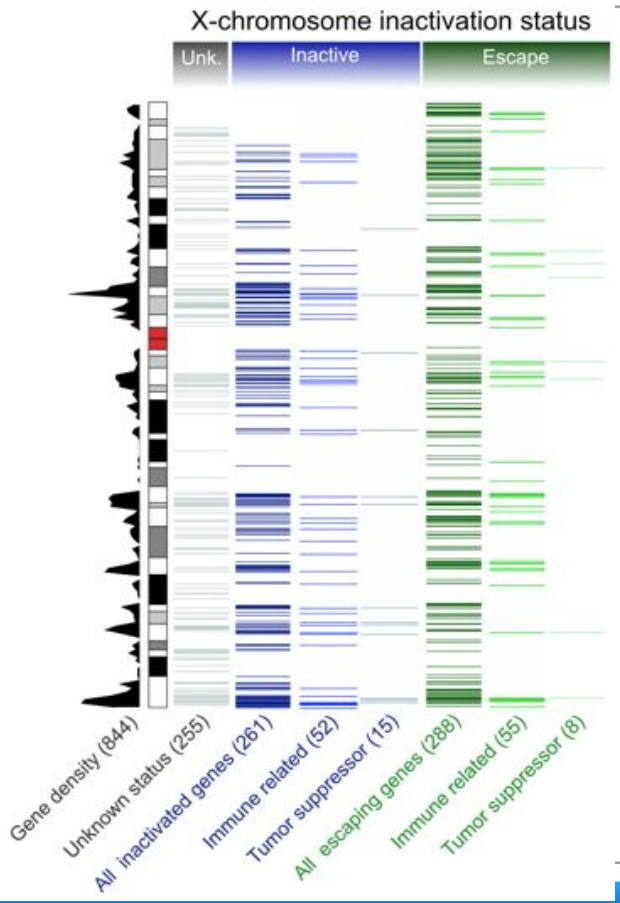


Mapping matters



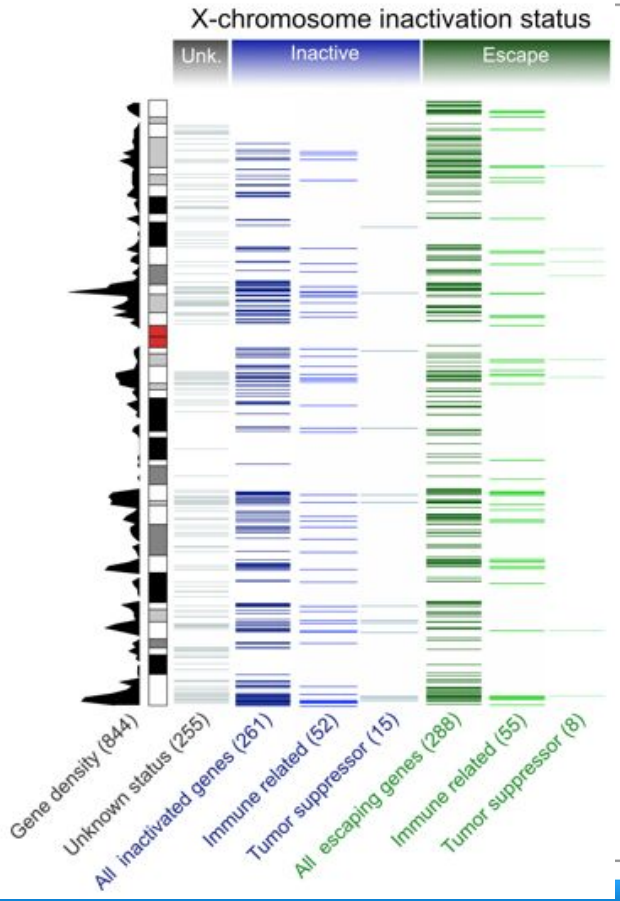
Manuscript in prep

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

NIH NIGMS

ASU SCHOOL OF Life Sciences
ARIZONA STATE UNIVERSITY

ASU Center for Evolution & Medicine
ARIZONA STATE UNIVERSITY

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.

UDN & Pediatric Genomics

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.

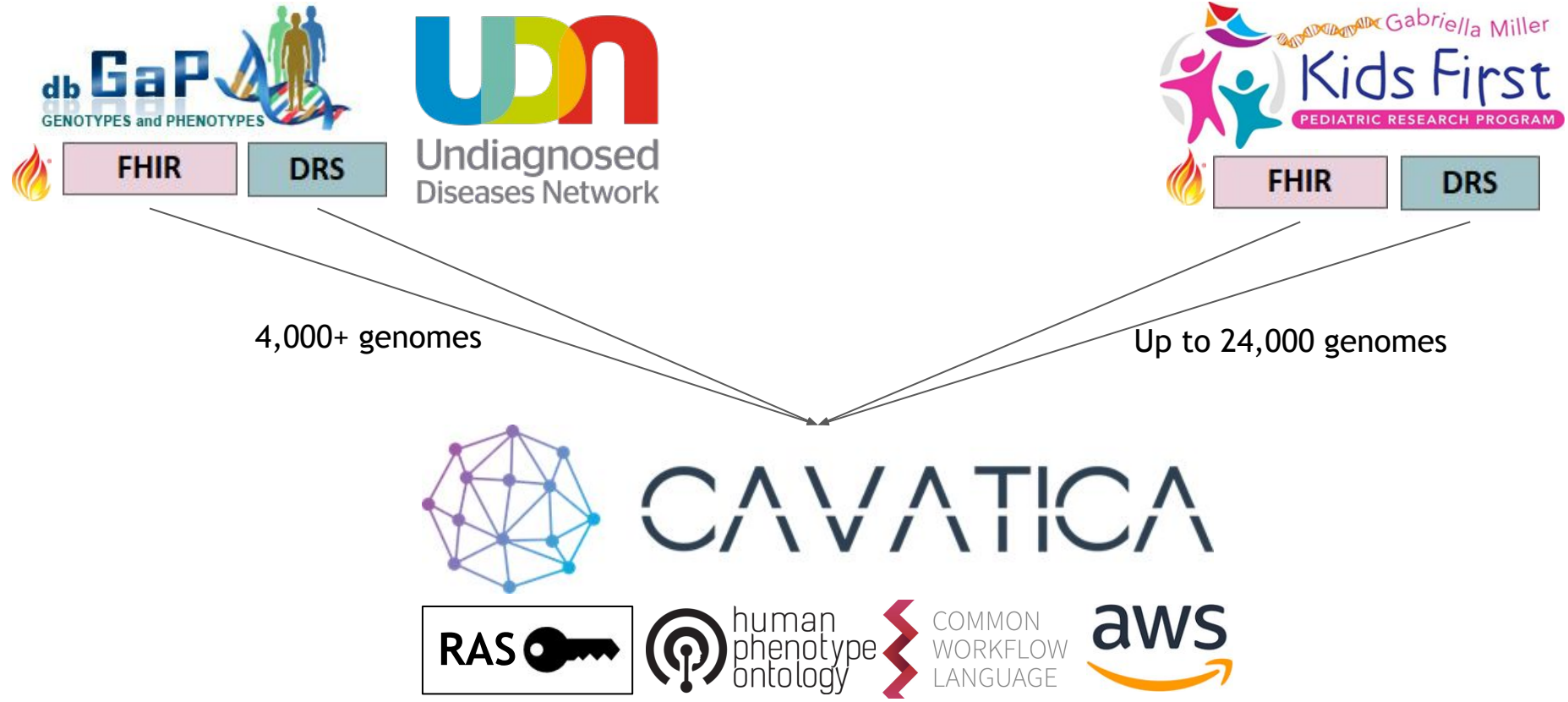
Our pipeline creates a **phenotype risk score (PheRS)** of the proband based on their clinical presentation described in human phenotype ontology terms (HPO). We then apply the PheRS to the biobank cohort, such that individuals with many overlapping features have a high PheRS, and those with no or few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

Undiagnosed Diseases Network

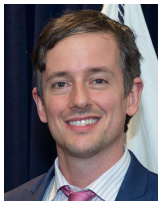
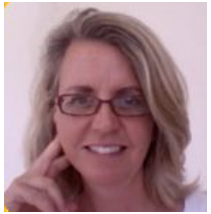
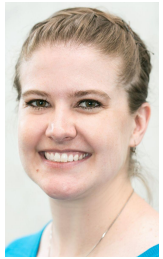
Overview of Standards Used



Solution Matrix

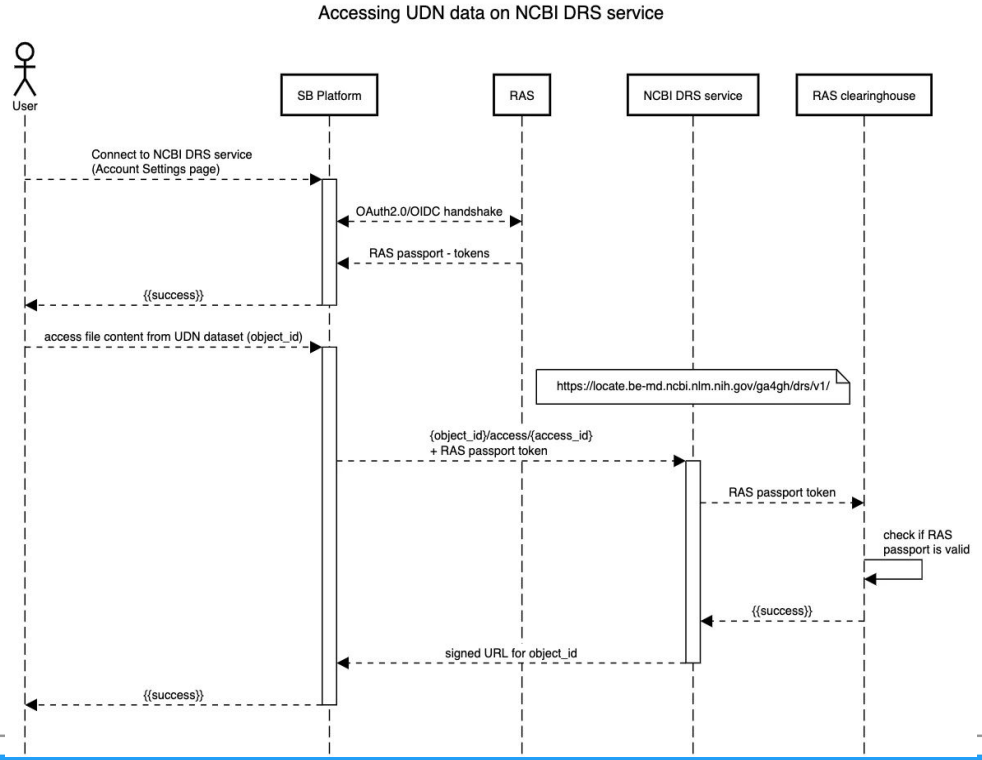
	Kids First Data Resource	NLM/NCBI	Analysis Tools
Genomic data	CAVATICA already integrated with the Kids First/Gen3 **DRS** server. **RAS** Milestone 3 is underway.	Connect CAVATICA to dbGaP **DRS** server, using RAS v1.1 Passports - Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identity 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

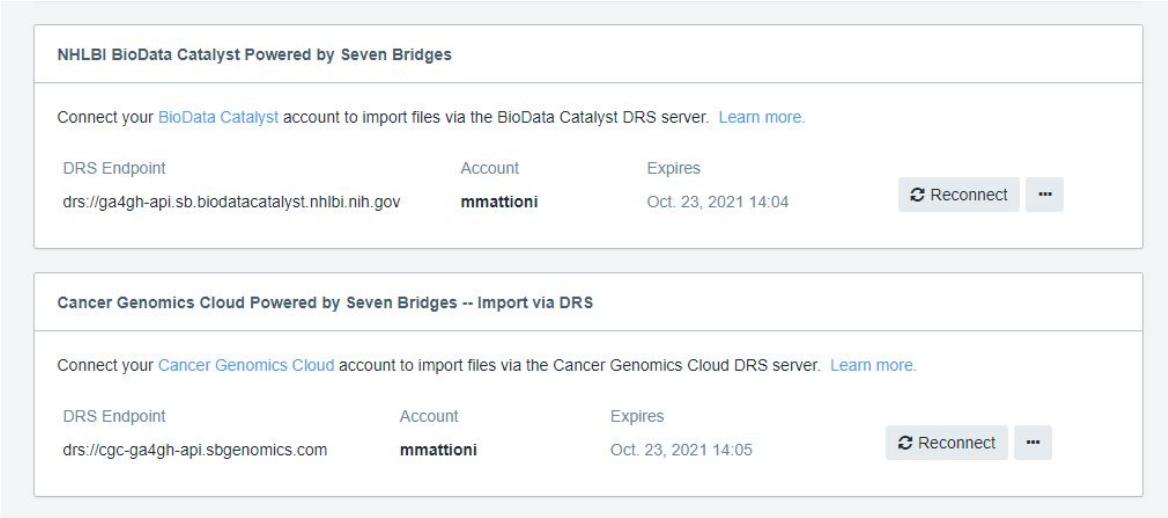
	Kids First Data Resource	NLM/NCBI	Tester/User
Genomic data	Michele Mattioni & Jack DiGiovanna & Adam Resnick   	Kurt Rodarmer & Yuriy Skripchenko  	Yuankun Zhu & Anne Deslattes Mays  
Phenotypic data	Allison Heath & Robert Carroll  	Liz Amos & Mike Feolo  	Lisa Bastarache 

Genomic Data Interoperability

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



CAVATICA: RAS Connection



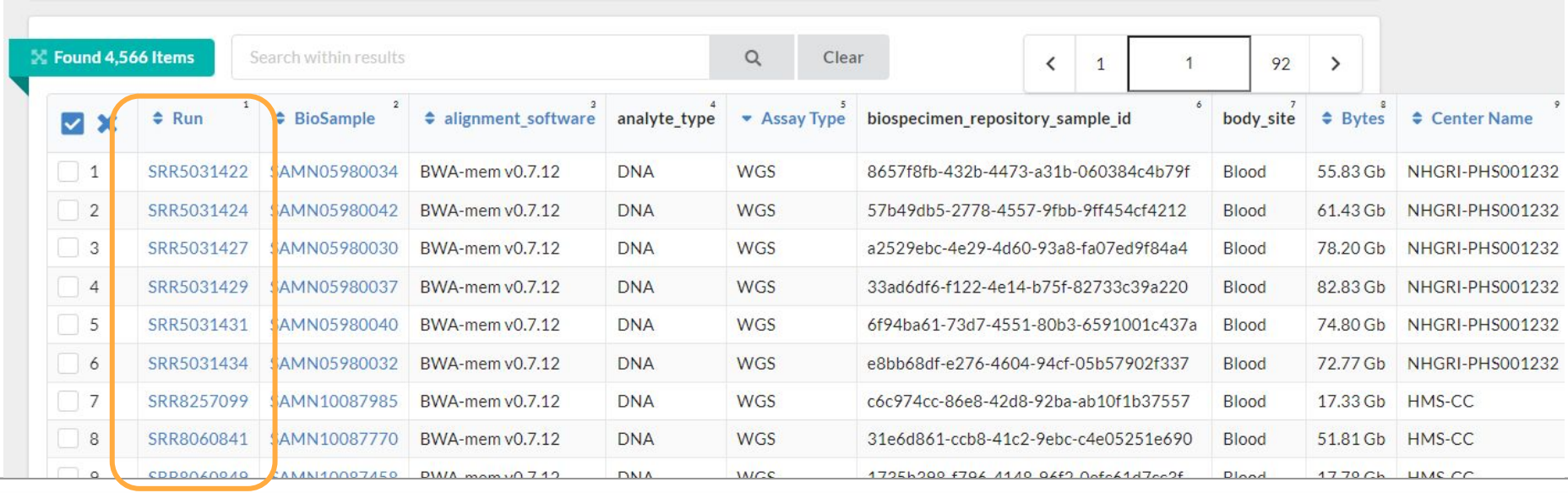
Connect with the NCBI DRS Server

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

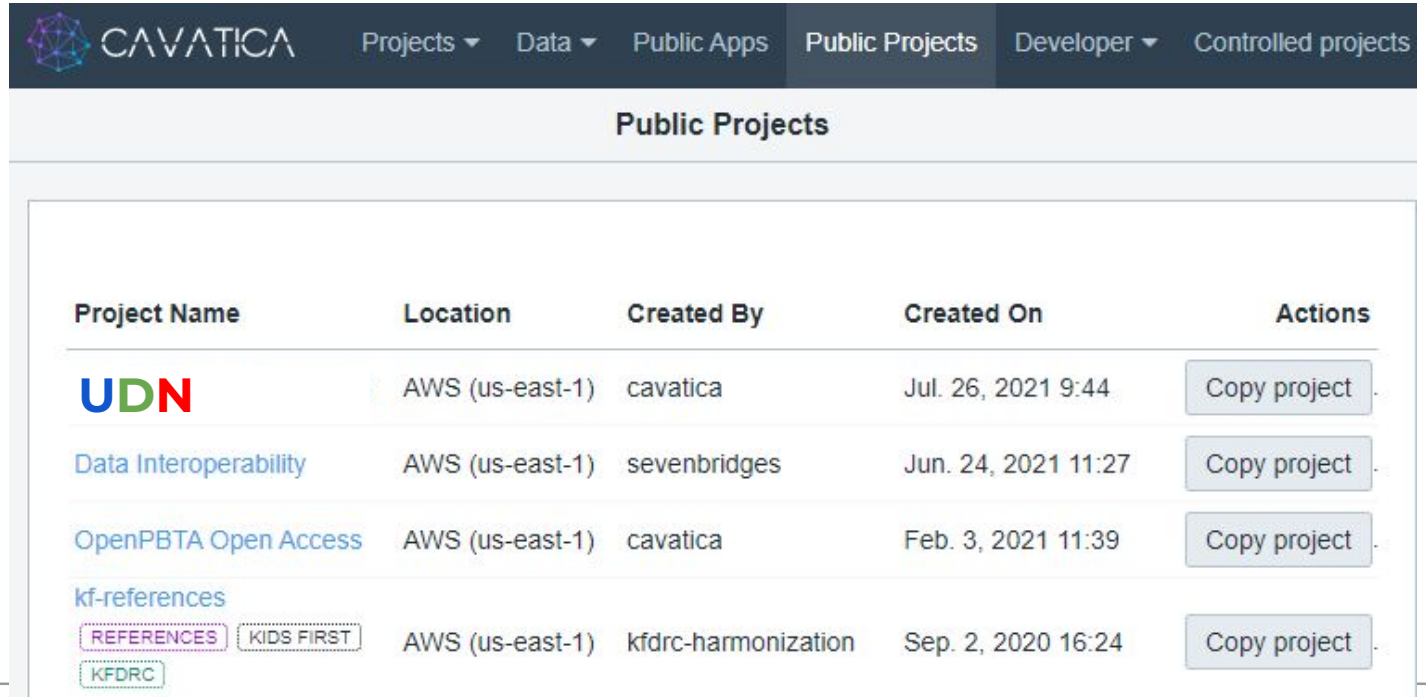


Found 4,566 Items | Search within results | Clear | 1 | 1 | 92

	Run	BioSample	alignment_software	analyte_type	Assay Type	biospecimen_repository_sample_id	body_site	Bytes	Center Name
1	SRR5031422	SAMN05980034	BWA-mem v0.7.12	DNA	WGS	8657f8fb-432b-4473-a31b-060384c4b79f	Blood	55.83 Gb	NHGRI-PHS001232
2	SRR5031424	SAMN05980042	BWA-mem v0.7.12	DNA	WGS	57b49db5-2778-4557-9fbb-9ff454cf4212	Blood	61.43 Gb	NHGRI-PHS001232
3	SRR5031427	SAMN05980030	BWA-mem v0.7.12	DNA	WGS	a2529ebc-4e29-4d60-93a8-fa07ed9f84a4	Blood	78.20 Gb	NHGRI-PHS001232
4	SRR5031429	SAMN05980037	BWA-mem v0.7.12	DNA	WGS	33ad6df6-f122-4e14-b75f-82733c39a220	Blood	82.83 Gb	NHGRI-PHS001232
5	SRR5031431	SAMN05980040	BWA-mem v0.7.12	DNA	WGS	6f94ba61-73d7-4551-80b3-6591001c437a	Blood	74.80 Gb	NHGRI-PHS001232
6	SRR5031434	SAMN05980032	BWA-mem v0.7.12	DNA	WGS	e8bb68df-e276-4604-94cf-05b57902f337	Blood	72.77 Gb	NHGRI-PHS001232
7	SRR8257099	SAMN10087985	BWA-mem v0.7.12	DNA	WGS	c6c974cc-86e8-42d8-92ba-ab10f1b37557	Blood	17.33 Gb	HMS-CC
8	SRR8060841	SAMN10087770	BWA-mem v0.7.12	DNA	WGS	31e6d861-ccb8-41c2-9ebc-c4e05251e690	Blood	51.81 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.



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	--	1	1 / 4843
chrX:g.48794116del	deletion	--	frameshift_variant GATA1 G397X	--	1	1 / 4843
chrX:g.48791978C>A	SNV	--	missense_variant GATA1 Q119K	--	1	1 / 4843
chrX:g.48792194C>T	SNV	rs140561920	missense_variant GATA1 R191C	Benign	1	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 - Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identity 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

Undiagnosed Diseases Network

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

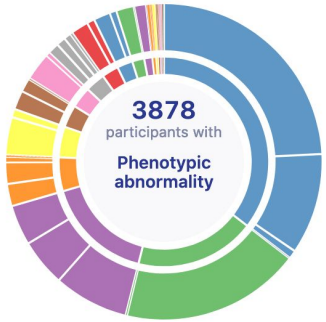
METABOLISM/HOMEOSTASIS
- Recurrent hypoglycemia
- Neonatal hypoglycemia

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)	27	3878
Abnormality of head or neck (HP:0000152)	0	1480
Abnormality of the musculoskeletal system (HP:0033127)	0	1328
Abnormality of the cardiovascular system (HP:0001626)	41	957
Abnormality of the nervous system (HP:0000707)	15	431
Abnormality of the eye (HP:0000478)	7	355
Abnormality of the genitourinary system (HP:0000119)	25	341
Abnormality of the digestive system (HP:0025031)	80	327
Abnormality of the respiratory system (HP:0002086)	0	303
Neoplasm (HP:0002664)	0	278
Abnormality of the ear (HP:0000598)	8	266
Abnormality of the integument (HP:0001574)	1	196
Abnormality of limbs (HP:0040064)	53	190
Growth abnormality (HP:0001507)	0	90
Abnormality of the immune system (HP:0002715)	0	59
Abnormality of the endocrine system (HP:0000818)	0	41
Abnormality of prenatal development or birth (HP:0001197)	0	26
Abnormality of blood and blood-forming tissues (HP:0001871)	0	24
Abnormality of the breast (HP:0000769)	0	18
Abnormal cellular phenotype (HP:0025354)	0	10
Abnormality of metabolism/homeostasis (HP:0001939)	0	9

chr18:g.62090521T>C Germline

Summary | Frequencies | Clinical Associations

Chr	18
Start	62090521
Alt. Allele	C
Ref. Allele	T

4 Studies

18 Participants

3.72e-3 Frequency

Type	Ref Genome	ClinVar	dbSNP
SNV	GRCh38	539565	rs200658159

Gene Consequences

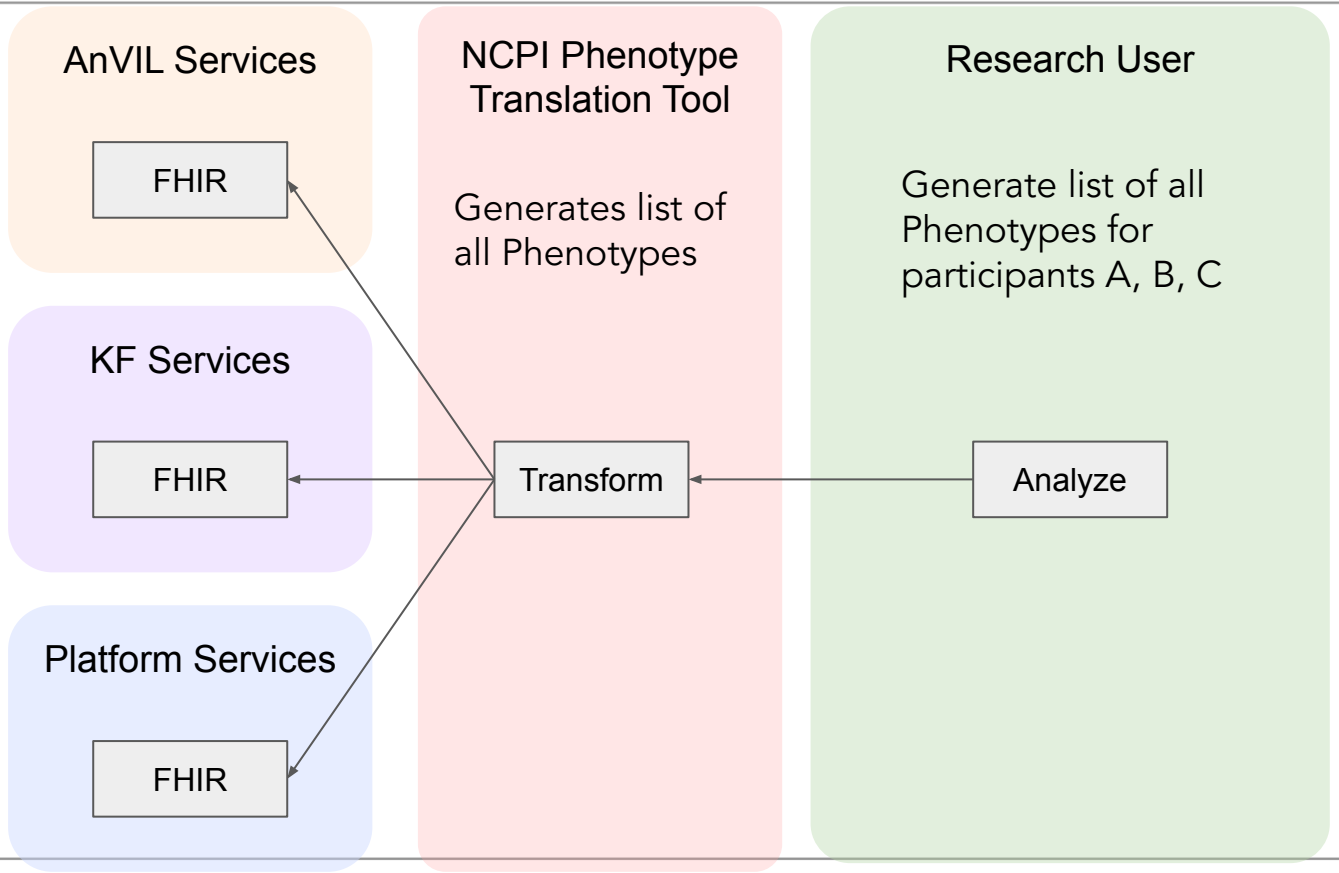
Gene PIGN

AA	Consequence	Coding Dna	Strand	VEP	Impact	Conservation	Transcript
I746M	missense_variant	2238T>C	—	Moderate	Sift: 0.13045 Polyphen2: Benign - 0.13045 More	0.05595	ENST00000640252

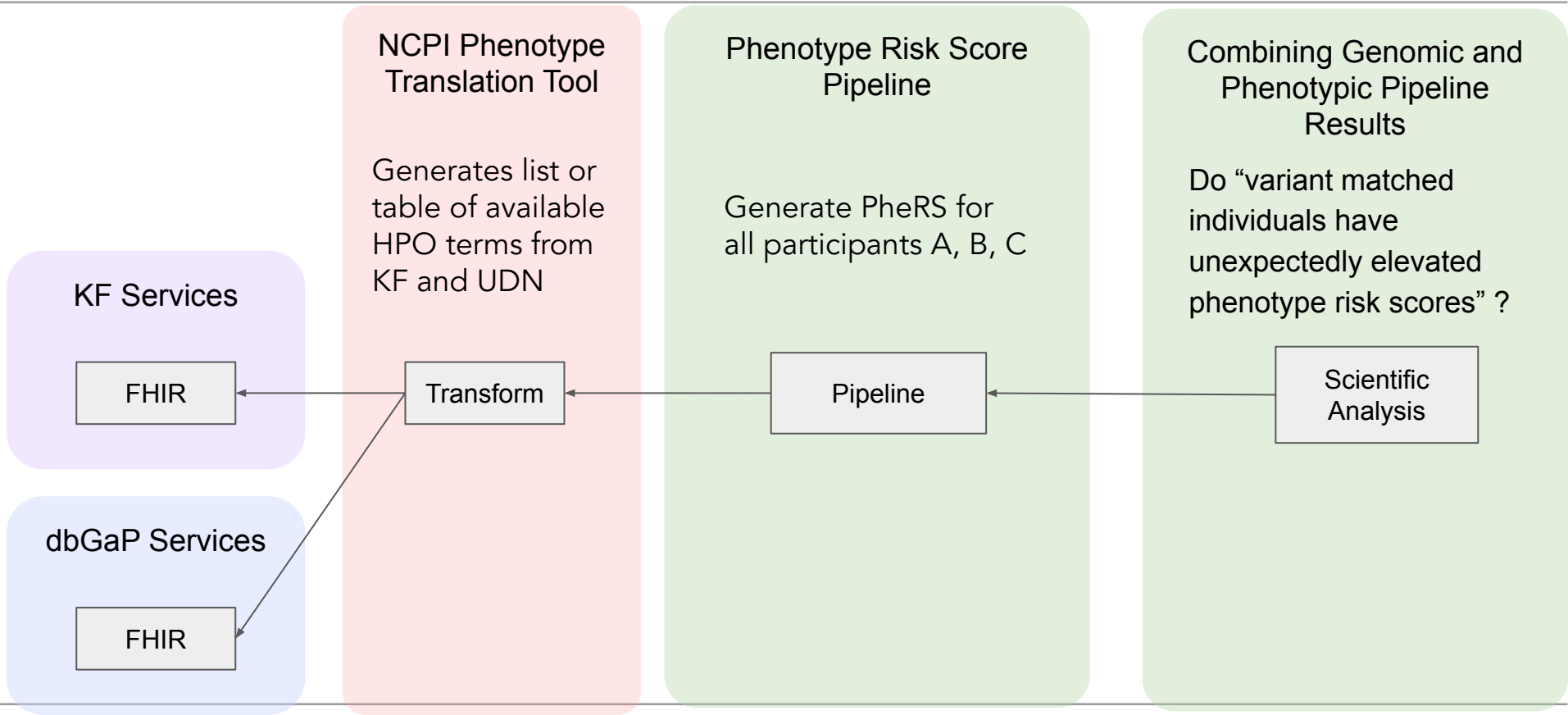
Show Transcripts (28)

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

Driving Tool / Service Layers: General



Driving Tool / Service Layers: Use Case

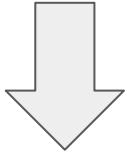


Concrete Progress on Each Step

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

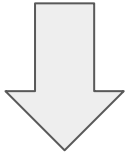
Transform

Concrete Progress on Each Step

NCPI Phenotype Translation Tool

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

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



Transform

Phenotype-based Pipeline

Generate PheRS for all participants A, B, C

Automated Pipeline

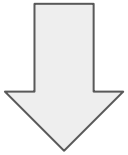
Concrete Progress on Each Step

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?

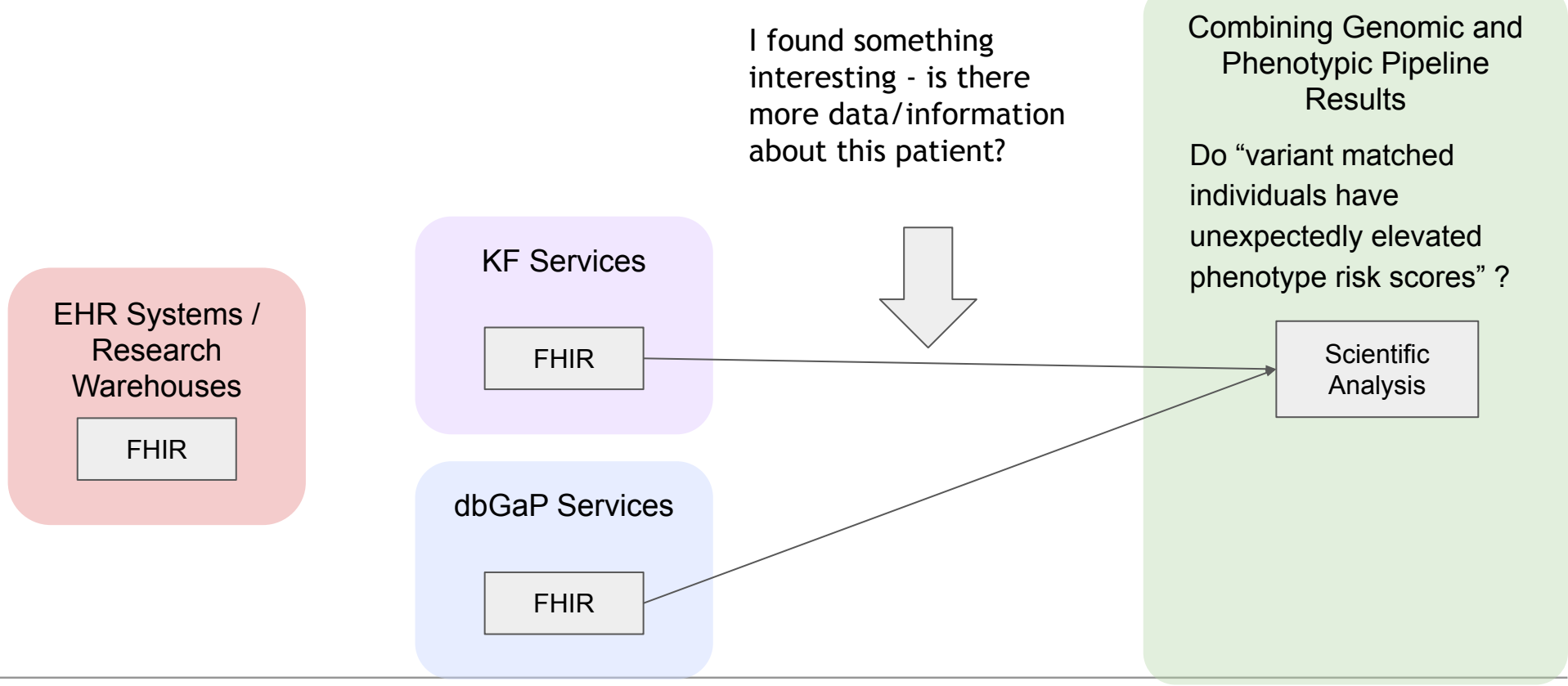


Combining Genomic and Phenotypic Pipeline Results

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

Scientific Analysis

Doors to New Capabilities



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

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

CAVATICA

Germline WGS

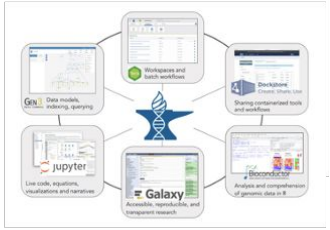
BASIC^3

BCM **A**dvancing **S**equencing
Into **C**hildhood **C**ancer **C**are

Follow-up study

Germline, Tumor Exome
Transcriptome

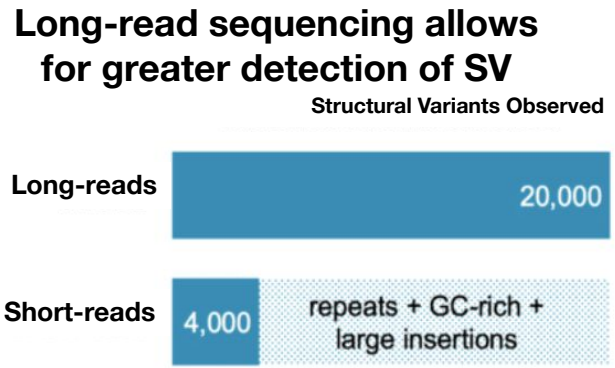
Kids Can Seq



Anvil

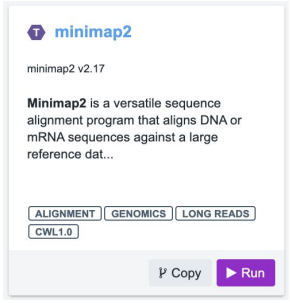
Germline Exome
Tumor Exome
Transcriptome

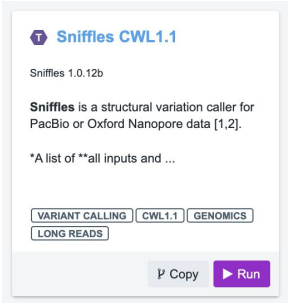
BASIC3 is undergoing Pacific Biosciences HiFi CCS long-read sequencing

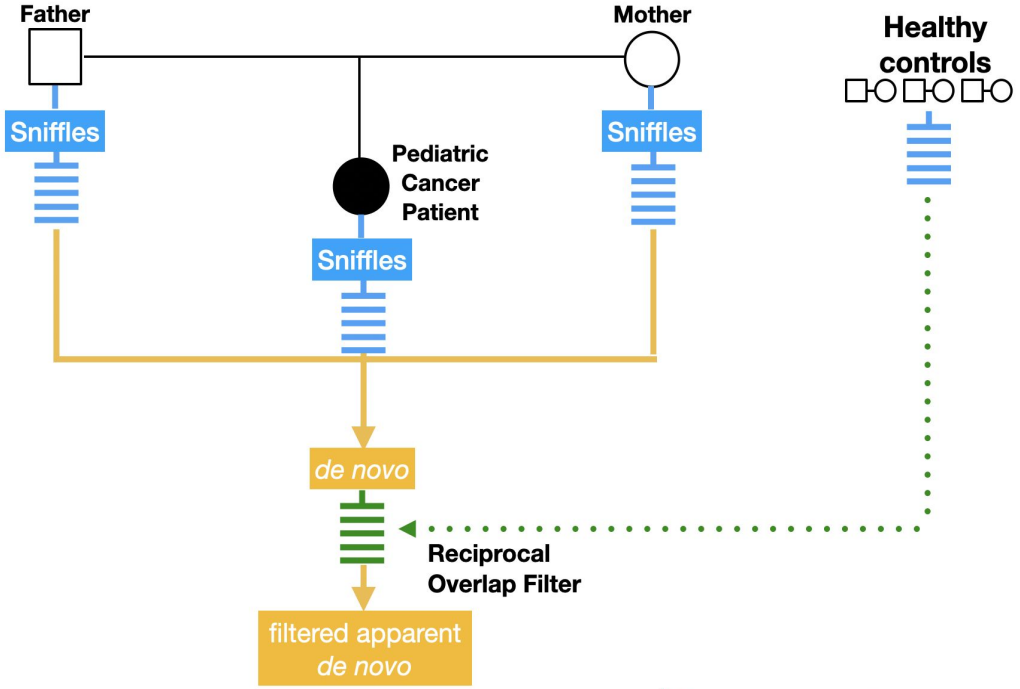


Allows for the comparison of long-read and short-read structural variant calling

Algorithms being utilized:







Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

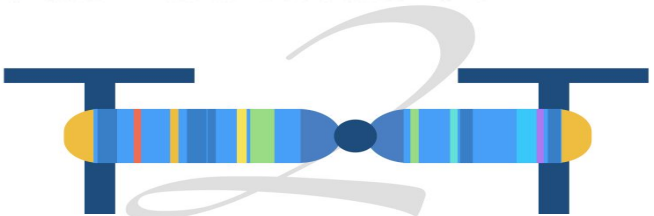
HUDSONALPHA
INSTITUTE FOR BIOTECHNOLOGY

Merker JD, et al. 2017

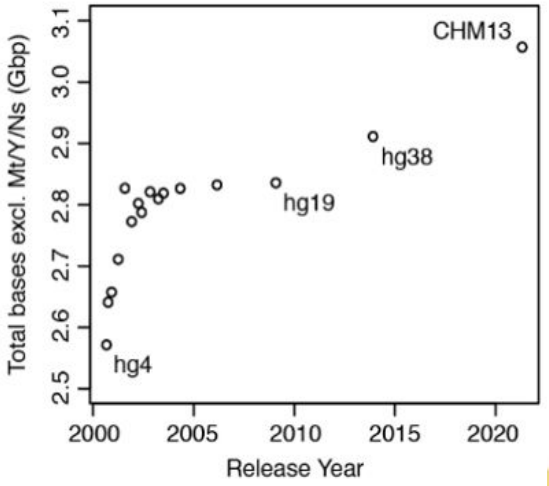
Novel CHM13 reference genome by the Telomere to Telomere (T2T) consortium

The complete sequence of a human genome

Sergey Nurk[1,*], Sergey Koren[1,*], Arang Rhie[1,*], Mikko Rautiainen[1,*], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[21], Tatiana Dvorkina[3], Ian T. Fiddes[22], Giulio Formenti[23,24], Robert S. Fulton[25], Arkarachai Fungtammasan[18], Erik Garrison[11,26], Patrick G.S. Grady[10], Tina A. Graves-Lindsay[27], Ira M. Hall[28], Nancy F. Hansen[29], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[30], Chirag Jain[1,31], Miten Jain[11], Erich D. Jarvis[23,24], Peter Kerpedjiev[32], Melanie Kirsche[9], Mikhail Kolmogorov[33], Jonas Korlach[30], Milinn Kremitzki[27], Heng Li[16,17], Valerie V. Maduro[34], Tobias Marschall[35], Ann M. McCartney[1], Jennifer McDaniel[36], Danny E. Miller[4,37], James C. Mullikin[14,29], Eugene W. Myers[38], Nathan D. Olson[36], Benedict Paten[11], Paul Peluso[30], Pavel A. Pevzner[33], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,39,40], Jeffrey A. Rosenfeld[41], Steven L. Salzberg[9,42], Valerie A. Schneider[43], Fritz J. Sedlazeck[44], Kishwar Shafin[11], Colin J. Shew[20], Alaina Shumate[42], Yumi Sims[19], Arian F. A. Smit[45], Daniela C. Soto[20], Ivan Sović[30,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[43], James Torrance[19], Justin Wagner[36], Brian P. Walenz[1], Aaron Wenger[30], Jonathan M. D. Wood[19], Chunlin Xiao[43], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[20], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13], Rachel J. O'Neill[10], Winston Timp[8,42], Justin M. Zook[36], Michael C. Schatz[9,49], Evan E. Eichler[4,24,†], Karen H. Miga[11,†], Adam M. Phillippy[1,†]



TELOMERE-TO-TELOMERE CONSORTIUM



doi.org/10.1101/2021.05.26.445798

The tools are uploaded in different languages across platforms

SevenBridges

Children's Hospital of Philadelphia
Center for Data Driven Discovery in Biomedicine

CAVATICA

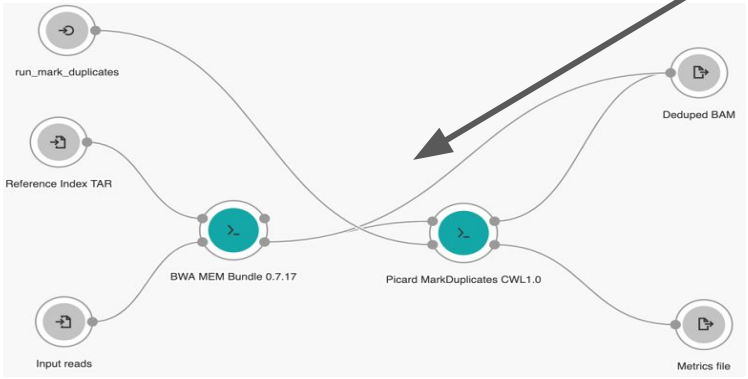
Terra

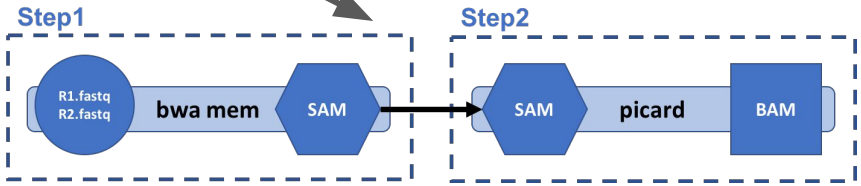
AnVIL

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

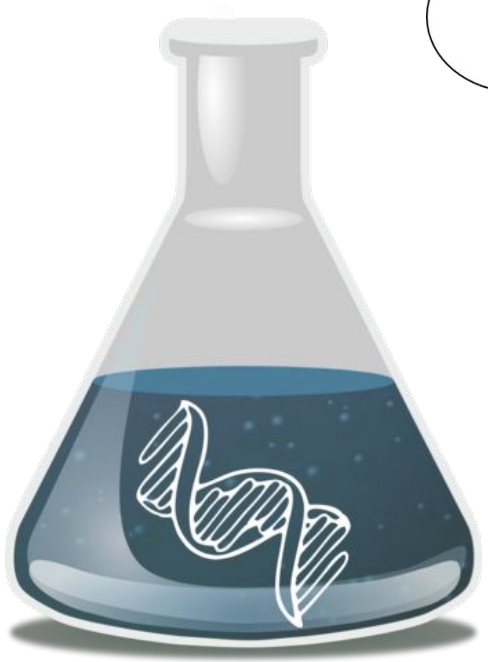
BWA MEM & Picard

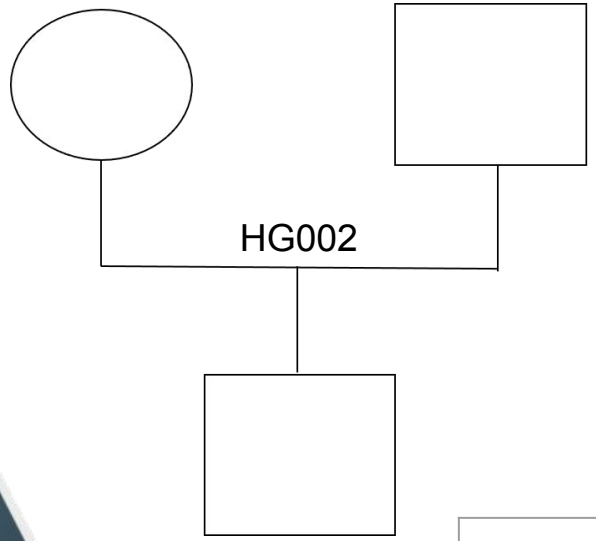
Tools are written in **Workflow Description Language (WDL)**





GIAB Benchmarking Data: HG002 Trio and Benchmarking Pipeline

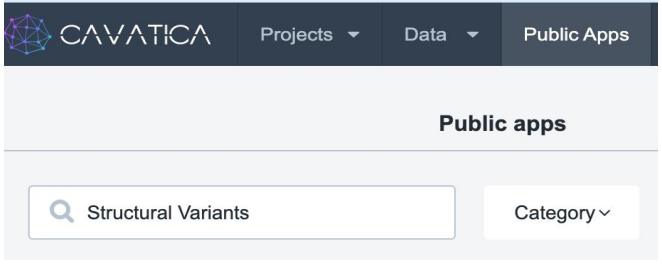


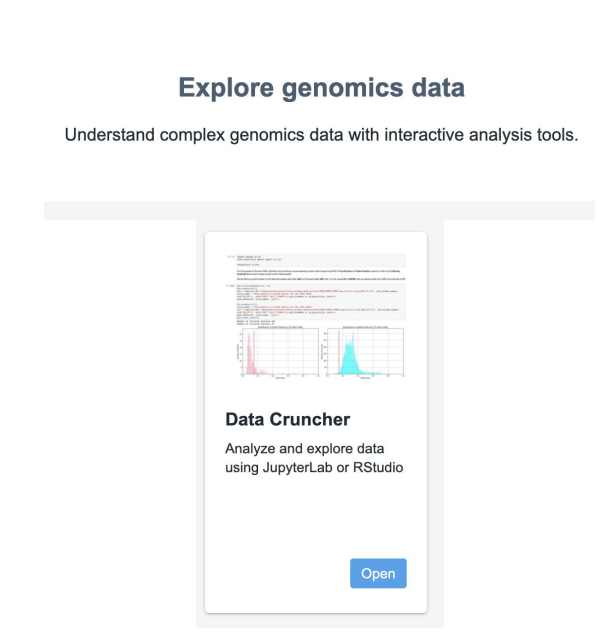


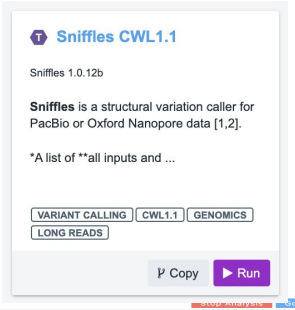
Long-Read Technology
PacBio Circular Consensus Sequencing (HiFi CCS)
Oxford Nanopore Promethion (ONT)
PacBio Continuous Long Read (CLR)

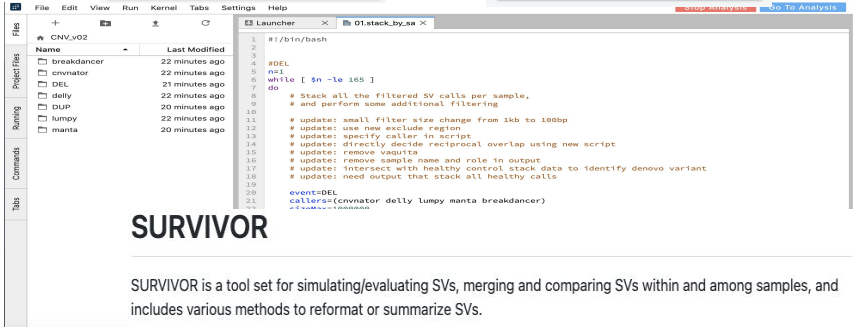
Tool	Version
Minimap2 (FASTQ Aligner)	2.17
Sniffles (Structural Variant Caller)	1.0.11
SURVIVOR (SV merging)	1.07

Creation of Long-Read SV Calling Pipeline on CAVATICA





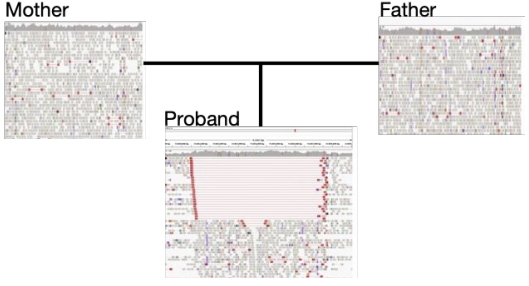




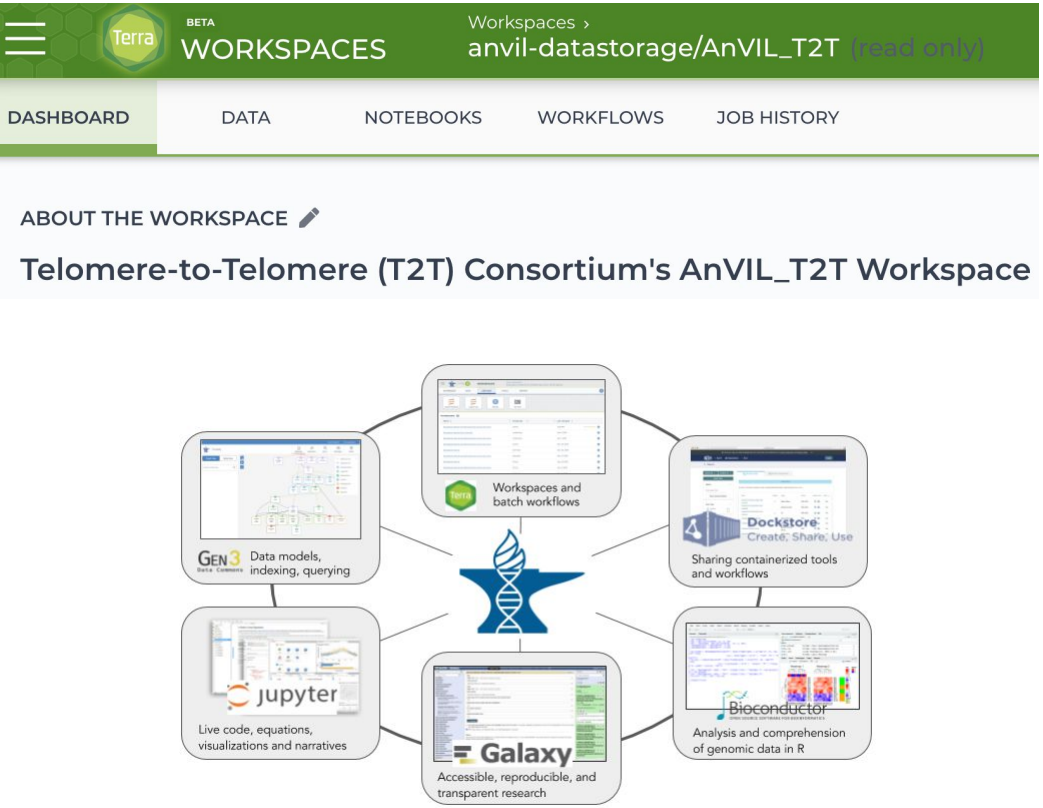
SURVIVOR

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



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:

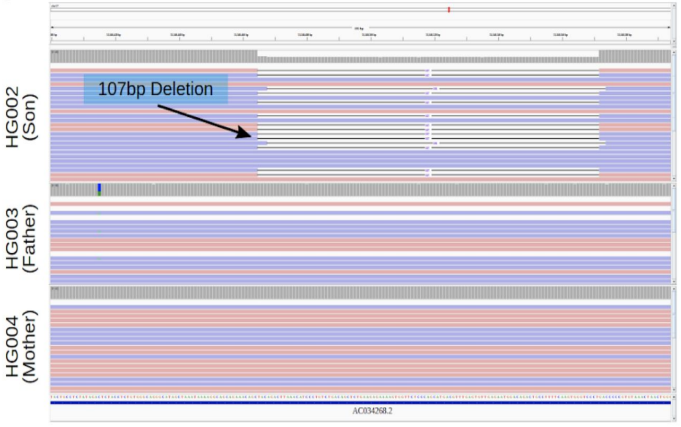
Sample	Raw Structural Variant Count (Terra.Bio)	Raw Structural Variant Count (CAVATICA)	Difference
HG002	92,350	96,977	+4,627
HG003	90,357	94,361	+4,004
HG004	88,803	93,159	+4,356

Discordant variant calls using SURVIVOR:

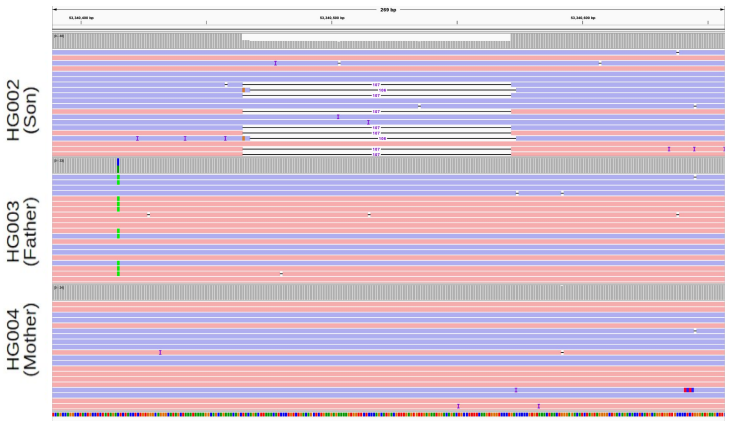
Sample	Variant Count (Terra.Bio)	Variant Count (CAVATICA)	Difference
Only in HG002	3,934	4,307	+373
Only in HG003	9,478	10,255	+777
Only in HG004	9,468	10,486	+1,018

De novo variants examples:

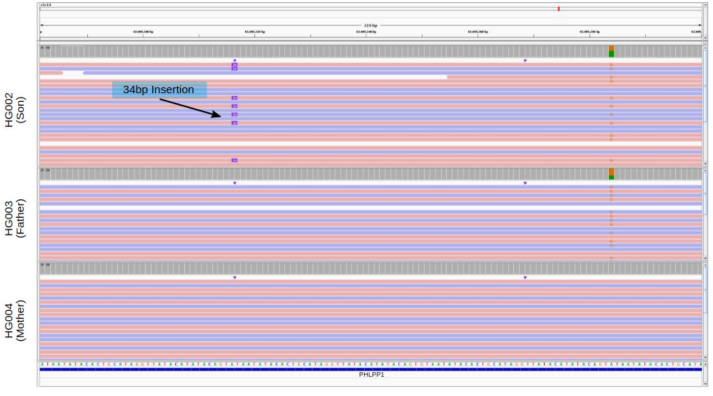
Deletion identified on the Terra platform



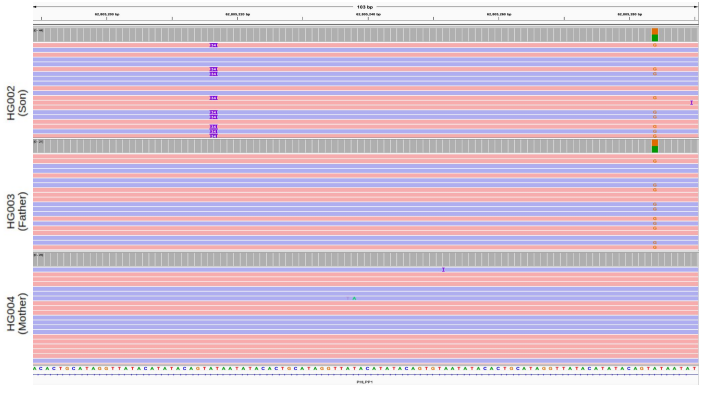
Deletion identified on the CAVATICA platform



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

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:

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

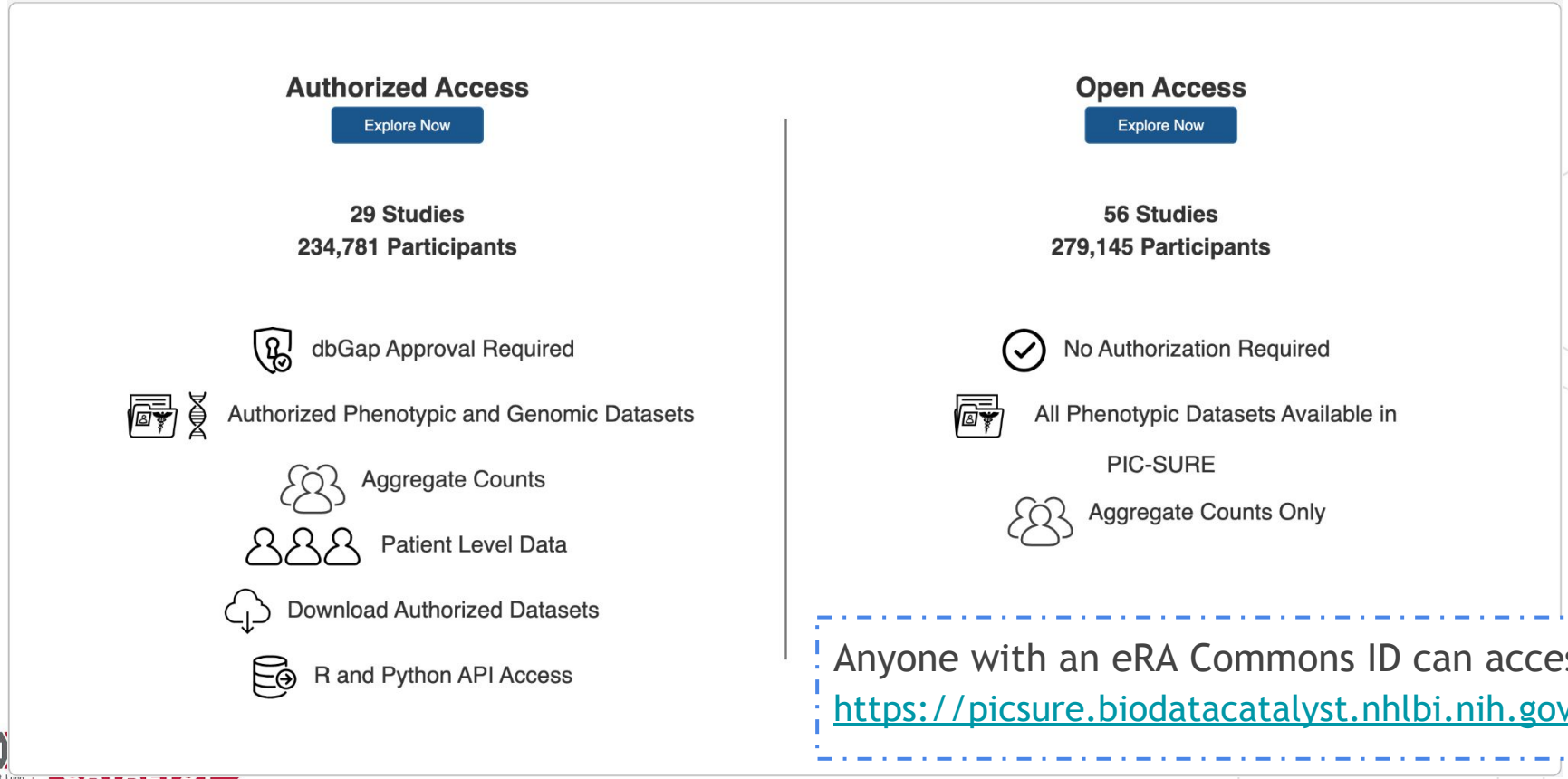
https://picsure.biodatacatalyst.nhlbi.nih.gov/

	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	Integrates clinical and genomic datasets across BioData Catalyst, including: ○ TOPMed and TOPMed-related studies ○ COVID-19 studies ○ BioLINCC 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	

NIH National Heart, Lung, and Blood Institute

BioData CATALYST

PIC-SURE Open and Authorized Access



Anyone with an eRA Commons ID can access!
https://picsure.biodatacatalyst.nhlbi.nih.gov/

Framingham Phenotype Datasets Table of Contents

Below is a listing of FHS SHARe 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 bwgt1_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_psysocalp_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_af4srv_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_alls vr_ctdates_2011_m_0715s vr_dates_2014_a_0912s vr_survaf_2014_a_0987s

Food Frequency with Derived Variables

vr_ffreq_ex01_3_0587s 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

cabg_2007s vr_cvdproc_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_survaf_2014_a_0987s

Endpoints: Cardiac/Cerebrovascular/Death

vr_soepedvt_2012_m_0756s vr_chfinit_2013_a_0828s vr_vte_2014_a_0913s vr_survcvd_2014_a_1023s vr_survdth_2014_a_1025s vr_soe4srv_2014_a_1027s vr_survstk_2014_a_1031s vr_soe_2016_a_1073s vr_soechf_2016_a_1070s

Stroke Related

psipl_2003s psipr_2003s vr_survstk_2014_a_1031s

Link to associated table

Bone Related

foapain_2001s 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

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

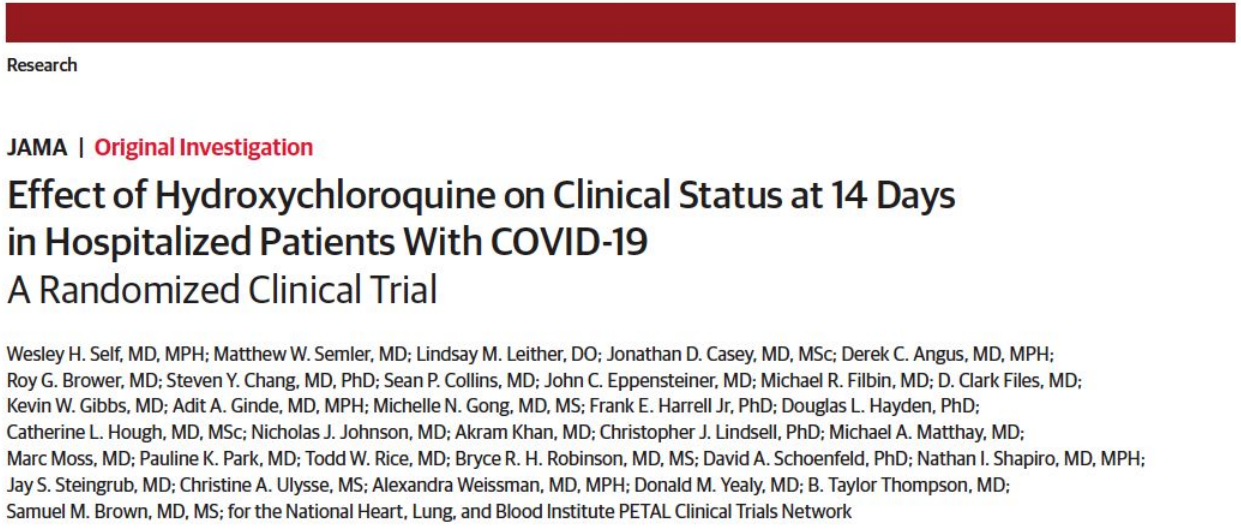
Use Case 1

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

JAMA | Original Investigation

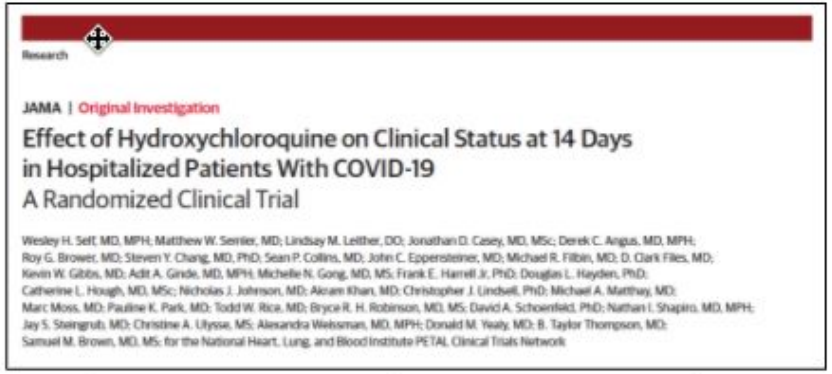
Effect of Hydroxychloroquine on Clinical Status at 14 Days in Hospitalized Patients With COVID-19

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

Published online:
November 9, 2020

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. 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

dbGaP

Patient level **Data**

Data **Dictionary**

Data management **code in SAS**

Analytic **code** *part 1 in R*

Analytic **code** *part 2 in **SAS***

Analytic **code** *part 3 in R*

PIC-SURE

API

jupyter

R

SevenBridges

Terra

NIH National Heart, Lung, and Blood Institute

BioData CATALYST

GEN3 Data Commons

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

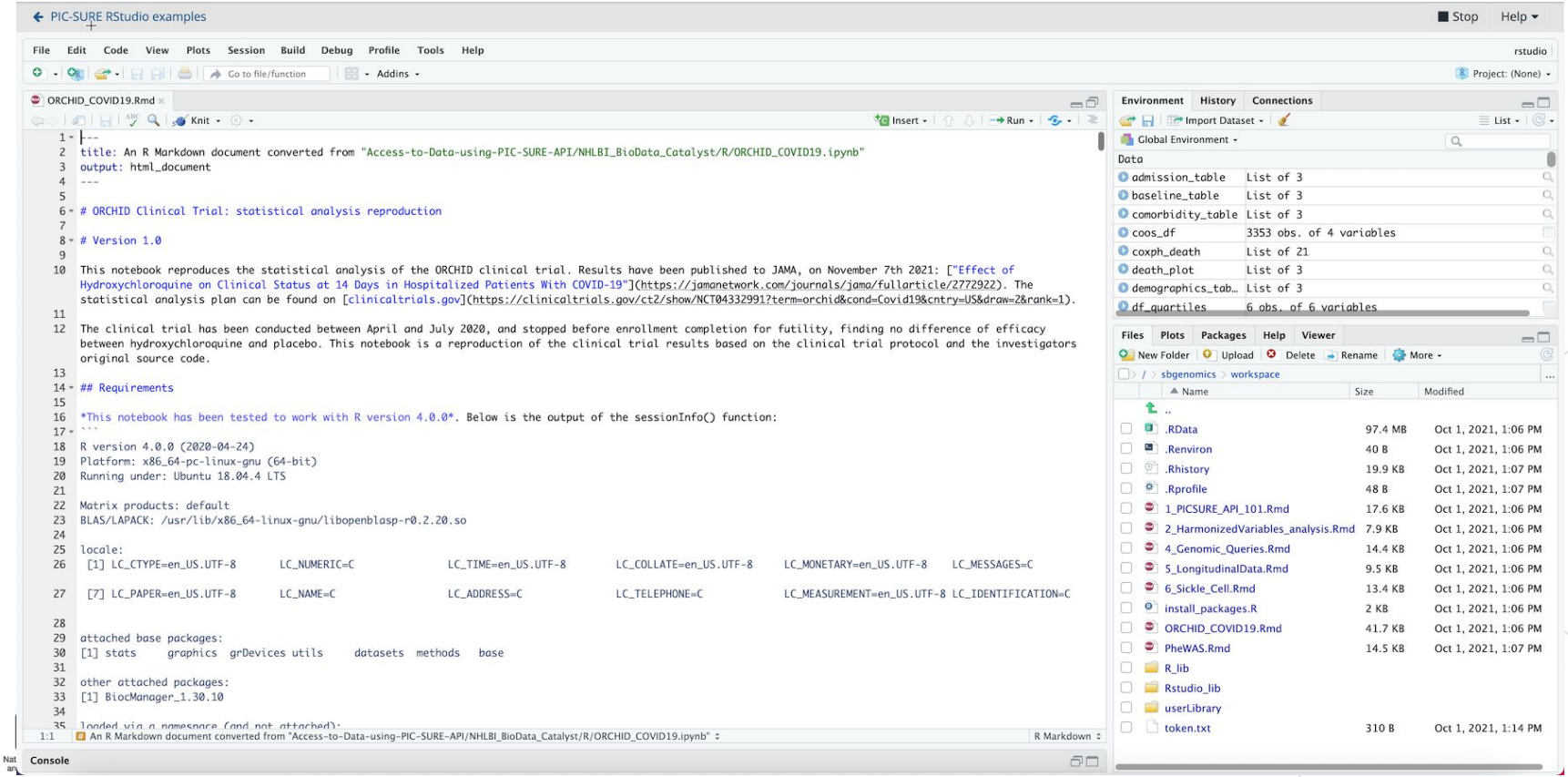
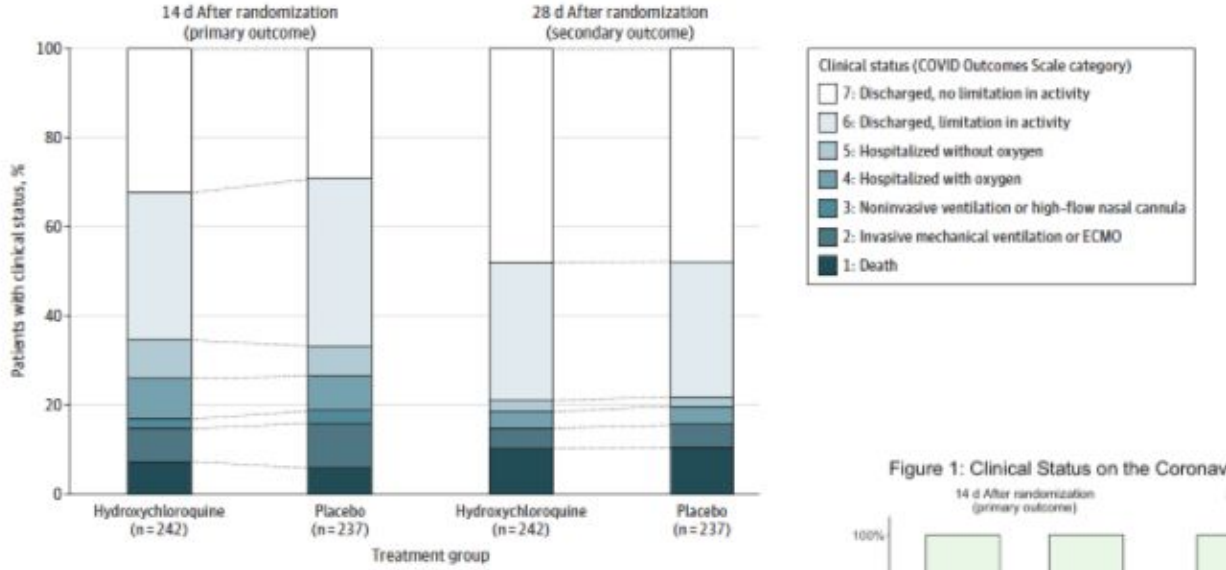


Figure 2. Clinical Status on the Coronavirus Disease (COVID) Outcomes Scale 14 Days and 28 Days After Randomization



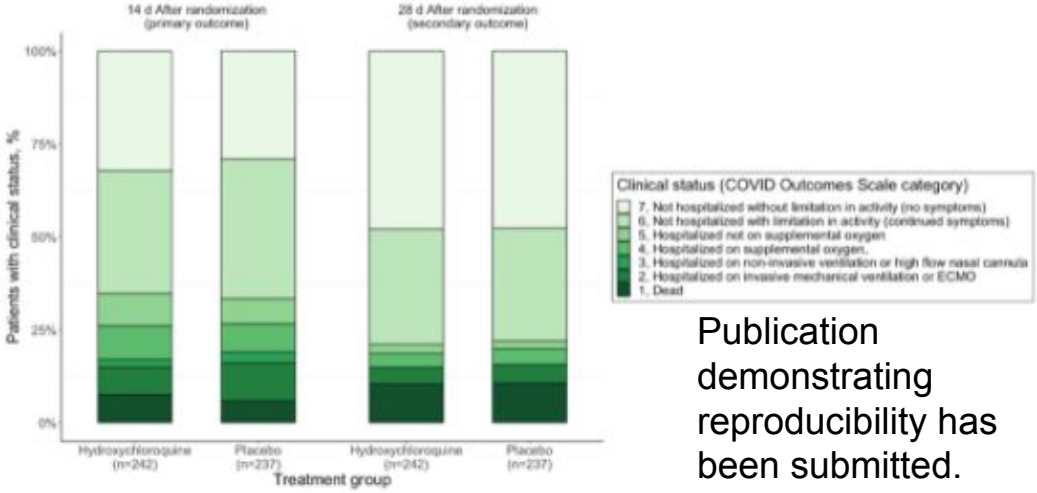
JAMA

Published online:
November 9, 2020

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

December 9, 2020

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



Publication demonstrating reproducibility has been submitted.

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

Use Case 2

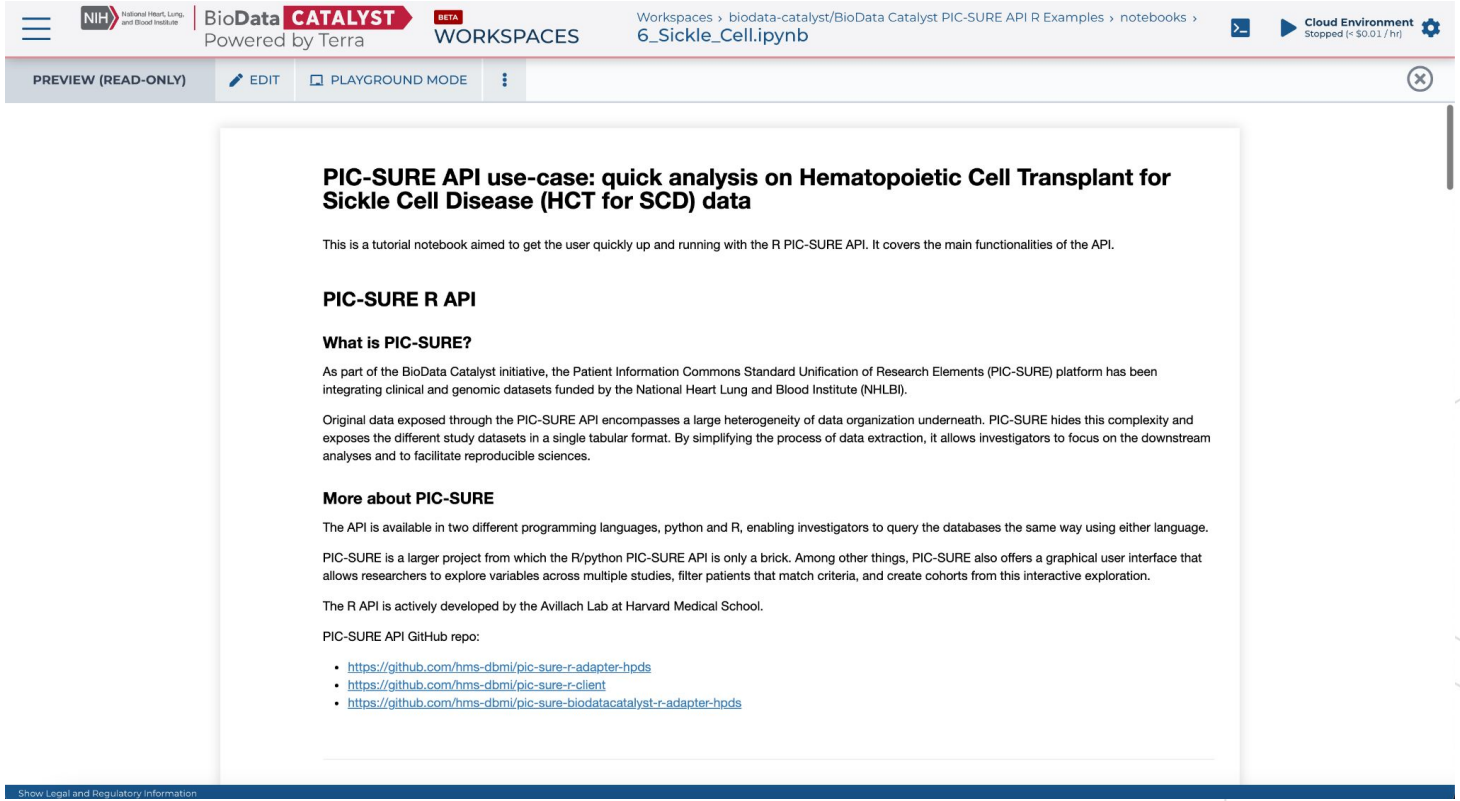
Using PIC-SURE to reproduce and expand analysis of the HCT for SCD Study on Terra

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

Portion of Notebook used now publicly available for others:
https://github.com/hms-dbmi/Access-to-Data-using-PIC-SURE-API/blob/master/NHLBI_BioData_Catalyst/R/6_Sickle_Cell.ipynb

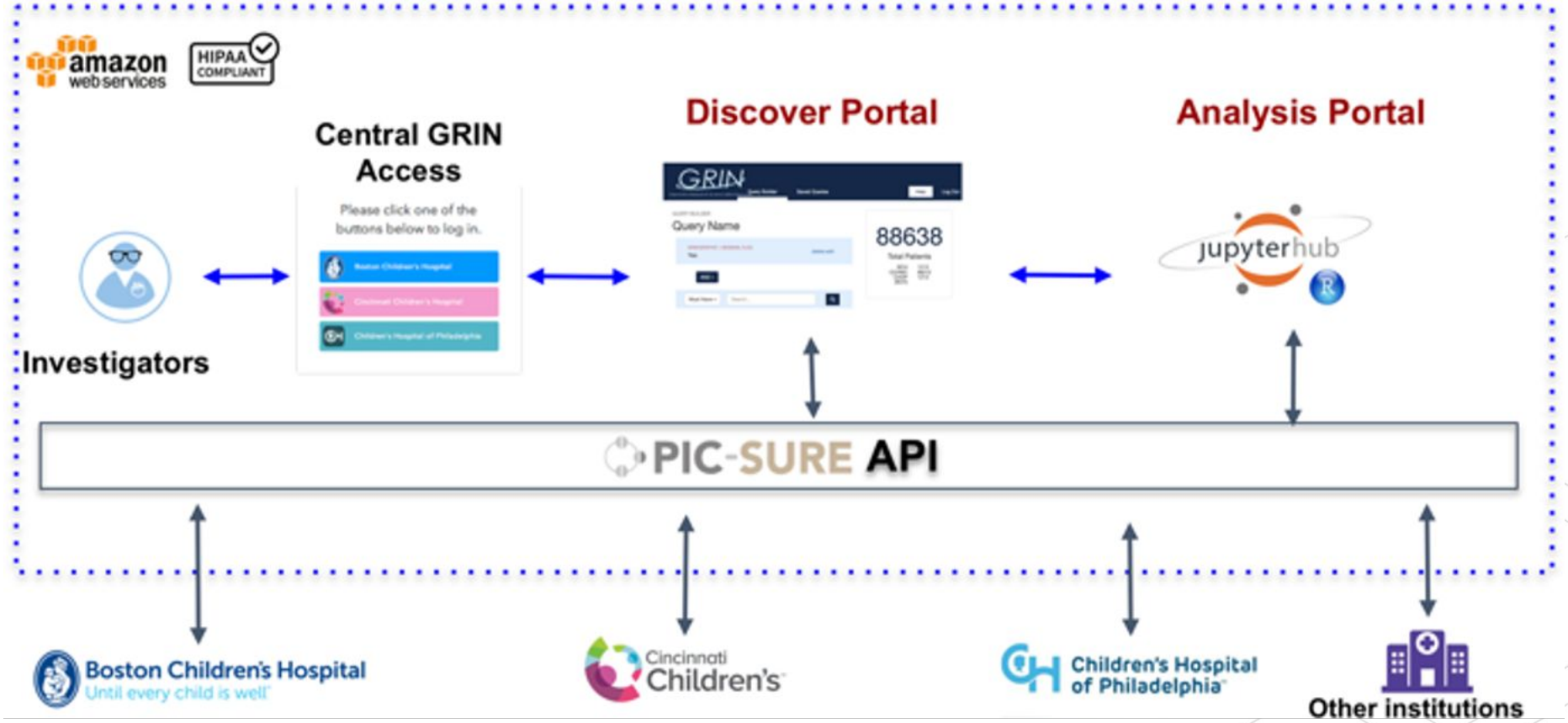
HCT for SCD Study Example Available in BioData Catalyst Powered by Terra



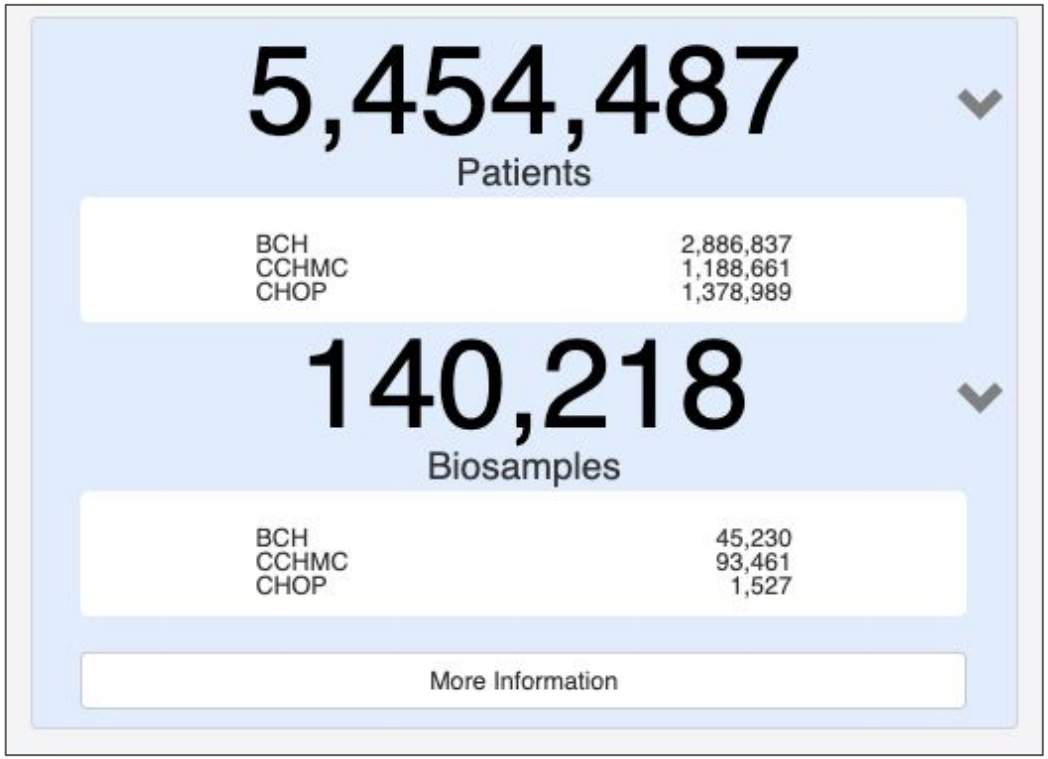
Genomic
Information
Commons

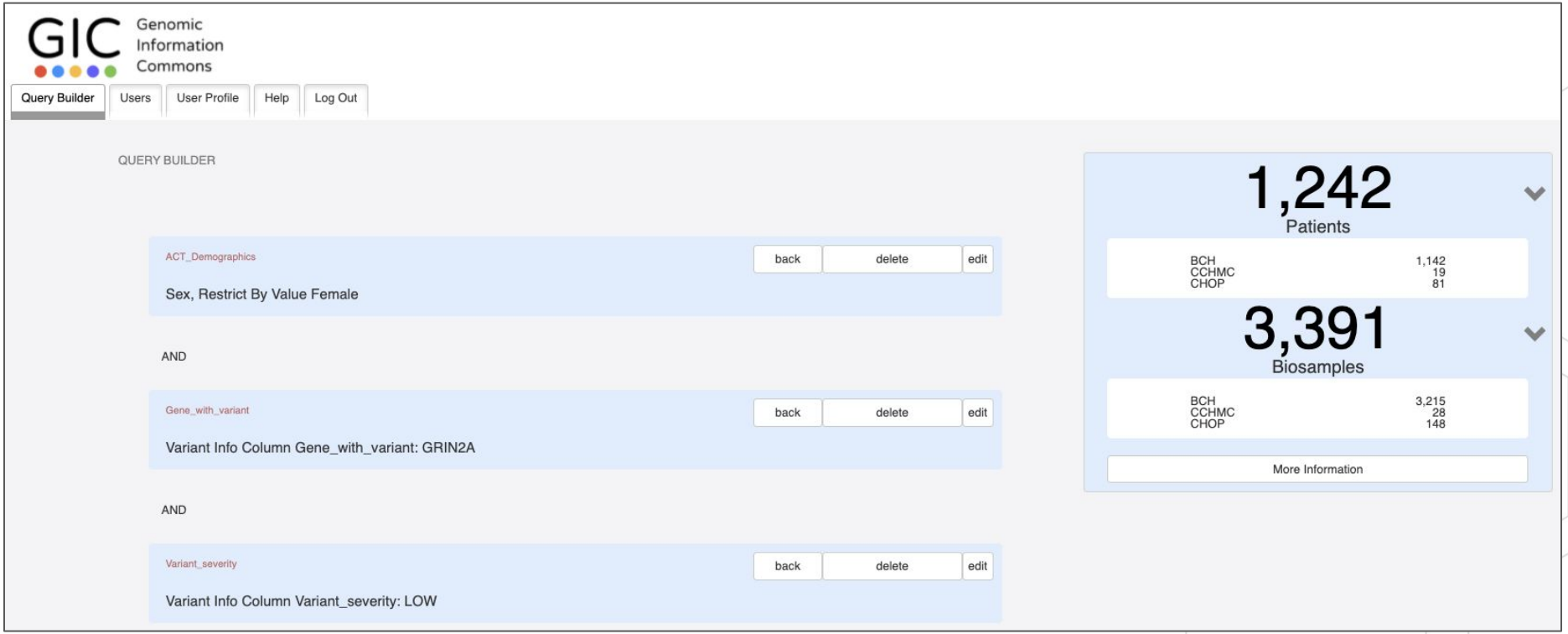
U01

National Center
for Advancing
Translational Sciences



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


5,454,487
Patients
BCH 2,886,837
CCHMC 1,188,661
CHOP 1,378,989
140,218
Biosamples
BCH 45,230
CCHMC 93,461
CHOP 1,527
More Information


GIC
Genomic
Information
Commons
Query Builder
Users
User Profile
Help
Log Out
QUERY BUILDER
ACT_Demographics
back
delete
edit
Sex, Restrict By Value Female
AND
Gene_with_variant
back
delete
edit
Variant Info Column Gene_with_variant: GRIN2A
AND
Variant_severity
back
delete
edit
Variant Info Column Variant_severity: LOW
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

API

PIC-SURE API
Application Programming Interface

PIC-SURE HPDS
High Performance Data Store

Patient level data

AnVIL · NIH National Heart, Lung, and Blood Institute · BioData CATALYST · All of Us Research Program · Gabriella Miller Kids First Pediatric Research Program · National COVID Cohort Collaborative

AnVIL · NIH National Heart, Lung, and Blood Institute · BioData CATALYST · All of Us Research Program · Gabriella Miller Kids First Pediatric Research Program · National COVID Cohort Collaborative · Service Workbench on aws

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

PIC-SURE HPDS
High Performance Data Store

PIC-SURE HPDS
High Performance Data Store

PIC-SURE HPDS
High Performance Data Store

PIC-SURE HPDS
High Performance Data Store

PIC-SURE HPDS
High Performance Data Store

AnVIL

NIH National Heart, Lung, and Blood Institute

BioData CATALYST

All of Us RESEARCH PROGRAM

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

National COVID Cohort Collaborative

NIH National Heart, Lung, and Blood Institute

BioData CATALYST

AnVIL

All of Us RESEARCH PROGRAM

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

Service Workbench on aws

National COVID Cohort Collaborative

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

PIC-SURE API
Application Programming Interface

PIC-SURE HPDS
High Performance Data Store

Patient level data stays in each platform

Index of all files centralized

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

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**
- We leverage existing APIs from each platform

PIC-SURE API
Application Programming Interface

AnVIL API

Terra, SBG, Gen3, PIC-SURE API

All of us API

Kids First API

N3C API

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!