Day 2: Tuesday, May 4

11:00am-12:30pm – Welcome and Community Interoperability Talks

Welcome - Melissa Haendel (U of Colorado) and Tanja Davidsen (NCI)

Notetaker: Alan Zheng (NCI)

Notes:

[Melissa on welcome]

- Welcome to Day 2 of the NCPI Spring Workshop
- Thanks to Tanja for organizing the meeting, to Sam for MCing yesterday and filling in for me today

[Tanja on logistics]

- Use WebEx application to access breakout sessions. Just click on the link, don't copy and paste the link directly in the browser
- Please mute yourself if not talking
- We will record all sessions except for the breakout sessions
- We will take notes for each session
- Speakers please turn on your camera when speaking, if possible
- Please register for this meeting if you have not done so (link on agenda)
- Please take the poll on Fall meeting dates (we have only 37 responses out of 177 registrants)
- Links to the slides and notes will be available on agenda after the meeting
- Each talk is 15 min (12 min talk and 3 min Q&A; at 9 min a warning is given)
- If you have questions, please use chat or raise hand

[Melissa presentation]

- We had some terrific conversations yesterday in the breakout sessions around search, data models, harmonization and RAS. Today's session is about community where we really put interoperability into action
- When we think about where we are with NCPI, we can think about what we want interoperability to do for us, and what that means to each one of us. If we do our jobs right, what will we be able to achieve collectively
- I think it is important for a community like this to understand the goals of interoperability. Being interoperable is good, but we should prioritize and direct our attention to the things that offer most bang for the buck.
- I'd like to go over a few ideas that summarize yesterday's discussion, and put these ideas in your head as we go through today's community presentations and future discussions about different types of interoperability
    - Legal and Licensing:
        - Has not been brought up before in NCPI discussion

        - Data licensing has been a big challenge across the community
        - Data is actually a licensed product. People don't really think a lot about what licenses they put on their data
        - When you mash up data, you cannot just redistribute according to the same license, unless the license of the original data was permissively licensed. Restrictively licensed data can only be combined with permissively licensed data
    - Regulatory:
        - We've talked about RAS, governance group, and access control. We need to make sure permissions on any given dataset match the access permission the user has been granted
        - This is really important and it is the goal of the passport work and GA4GH
    - System:
        - We spent a lot of time talking about system level interoperability and making sure that platforms and tools can actually talk to each other, so you can move data, analytical tools, from one place to another
        - Portability across systems is the key
        - APIs that function in the same way across an ecosystem of data is also important
        - All the security implications that come with interoperability need to adhere to the same common standards
    - Data:
        - Data is often not coded at all, or coded in different data models or terminology. We have a lot of free text fields that have to be mapped to something to create data interoperability
        - We need to think at a deeper level about the fundamental data model harmonization and the terminology harmonization
        - NCPI could help the community by working more upstream to define common data models, to define common terminologies, to help build tools, to be compliant further upstream, and not solely rely on downstream groups to do the data harmonization
- I am hoping as we go through the community presentations we can think about how they have achieved the success, based on some of these facets of data interoperability:
    - Ontology (the semantic, the data context)
        - There are many many ontologies and predefined concepts we use in our ecosystems such as Mondo, the human phenotype, Snomed, MCIt
    - Data Models (syntactic, data language)

        - These are the predefined models and data structures that the end data dictionaries, ontology terms, and enumerations can slot into
        - For example, how we related patients to a sample is different across different systems. The identifiers associated with [the sample and the patient], how we create those associations, the temporality of the collection of the sample, and the downstream analysis of the sample, all depend on the basic structure, which we don't have commonality for
    - Formats (system, data presentation)
        - This is how we present the content to the system. It is just how we encode and decode and represent the actual content
        - A few examples: OWL, RDF, VCF, FASTA, PBF
    - Exchange (structural, data architecture)
        - Help move data around via networks, computers, applications, web services, APIs, Docker
- With these data interoperability thoughts, we will move on to our first presentation. Think about what interoperability has meant in the context of the real-world applications. I'd like to introduce Tim Majarian from the Broad Institute, who is going to talk about "the proof of concept of interoperable approaches for improving outcomes of pediatric disease"

Proof of concept of interoperable approaches for improving outcomes of pediatric diseases - Tim Majarian (Broad)

Notetaker: Durga Addepalli (NCI)

Notes:

- Effort towards improving outcomes in pediatric diseases
- Combine data from three different Cloud Resources where we combined different studies, some for cases and some of controls.
- Leverage individual level data specifically in context of rare variations.
- GWAS study, cases were from KidsFriist, PCGC consortia –and Controls from TOPMed and Jackson Heart Study
- 1st Pediatric consortium to identify genetic variations associated with CHD
    - KidsFirst and TOPMed
- CHD is most common diagnosis at the time of birth. ¼ cases is critical. There are a large number of diagnosis which are distinct genetically and have heterogenous outcome. The challenge is there is no known shared genetic basis.
    - Another reason for CHDs is chromosomal abnormalities, in about 11% patients

- Looking at genetics of child and parents, case-parent rather than case-control.
- Main disadvantage is that it is easier to recruit all these people rather than to get 100 families with both parents and child.
- Our solution was to combine multiple datasets- larger disease focused studies. Used 2 studies for cases + 2 studies for controls.
- What interoperability efforts have done are
    - Cloud has been a challenge initially, but once everything was developed we are now capable of much more streamlined work
    - Were able to get DRS URI into our workspace which available for all data we used. This might lead to a push button format to get data to your platform
    - Good if available in other cloud environments also with no upload and upload – as of now on Terra
- Study Population - Multiple races and ancestries represented, with Controls from 2 studies from TOPMed
- Removed unrelated, and combined all samples and diagnoses -7000 individuals
- Actual framework – Proxy on slide
    - Count number of non-synonymous alleles and synonymous alleles – look for differences between the two
    - For synonymous used only synonymous variance annotation
    - Raw number of variants might be different – but we used ratio, so we don't need to worry about case control balance
    - 1 P value for each gene tested
- We saw no significant gene-based association but don't just see inflation or evidence of confounding either
- Non-CHD genes showed no enrichment for p values, therefore combining datasets was not an issue
- Didn't see associations because of lot of heterogeneity in the datasets. It groups them into one large category which is not good to do – we had to combine all in spite of race and ancestry and only looked into SNPs and indels and no structure variants
- Following up with gene expression analysis to enrich associations and being done on cloud platform
- Easy to do did it in the cloud and act on it
- What is important is more samples and more actual population diversity has actual impact

Q &A

- Terra space without the control data where we can see the methodology?
    - We can get it
- Are you taking VCF?

- Probands only. Available now iis Data trio VCFs or individual levels VDFs

Q&A from Chat

from Valerie Cotton to Everyone: 11:19 AM

NIH's copy of (most of) the data we are discussing are pre-competitive constrained only by terms of the DUC based on consent (not licensing per se but some folks refer this as licensing since redistribution is prohibited and DULs apply), and open-access data can be re-distributed (and often is in papers etc).

from Digant Shah (internal) to Everyone: 11:26 AM

@brian the poll is open now

from Adelaide Rhodes to Everyone: 11:26 AM

Is there a workspace without the controlled data to see the methodology?

from Ian Fore (internal) to Everyone: 11:28 AM

Excluding anyone related - mentioned in passing, but guessing that wasn't trivial across studies. Did it rely on identifying relatives via genomics? Or via. PHI?

from Kurt W. Rodarmer to Everyone: 11:31 AM

Is all external data accessed by importing into compute environment?

from Asiyah Lin to Everyone: 11:31 AM

What are the statistical considerations when designing the case-control study

from Asiyah Lin to Everyone: 11:32 AM

@ Tim, What are the statistical considerations when designing this case-control study?

from Anne Thessen to Everyone: 11:32 AM

I wonder if there is an environmental component to CHD that is not represented

from Tanja Davidsen (internal) to Everyone: 11:33 AM

Fall survey is back up!

from Tim Majarian to Everyone: 11:33 AM

@Ian Fore - We used KING to identify individuals closer than 2nd cousins & remove them.

Kids First and Multi-Cloud BASIC3 - Sharon Plon (BCM) & Owen Hirschi (BCM)
Notetaker: Jay Ronquillo (NCI)
Notes:

- BASIC3
    - NIH-funded cohort
    - “BCM (Baylor College of Medicine) Advancing Sequencing Into Childhood Cancer Care”
- Probands from NIH-funded BASIC3 have undergone clinical germline and somatic WES
    - Goal: characterize the diagnostic yield of combined tumor and germline WES for children with solid tumors (N=287)
- 120 probands from BASIC3 selected for trio WGS
    - Goal: identify de novo SVs, SNVs, and putative pathogenic variants in known cancer genes missed by whole exome sequencing
    - CAVATICA allowed easy integration and configuration of various tools, like Platypus and Variant Effect Predictor
- CAVATICA expedited de novo structural variant analysis/discovery
    - Leveraged multiple features of platform
    - Allowed up to 80 tasks to run simultaneously
    - Made easy to upload new tools
    - Allows moving from short reads to long reads, using “Sniffles” caller
        - PacBio long-read sequencing of BASIC3
- Kids First and CAVATICA enabled BASIC3 analysis
    - Quickly/efficiently upload tools and analyze BASIC3 short-read WGS for de novo SNVs
    - CAVATICA allowed use of pre-existing applications and terminal interface to create novel pipeline for analysis of structured variants in BASIC3
    - Kids First worked with BCM et al to upload tools in preparation for analysis of BASIC3 long-read WGS
- Output and dataflow of BASIC3 analysis
    - Germline WGS → Kids First and CAVATICA
    - Clinical variant interpretation → ClinVar
    - Germline exome, tumor exome, transcriptome → dbGaP
    - Follow-up study (germline, tumor exome, transcriptome) → “Kids Can Seq” into AnVIL
    - Other resources → St. Jude Cloud (PeCan)
- Key takeaways from chat and Q&A:

    - Time savings from CAVATICA allowed running of up to 80 tasks at once: estimated to be roughly 30 hours for tasks instead of 100+ hours on local machine
    - “Data never dies” – as tools and techniques improve, data gets analyzed and re-analyzed with new questions/tools over time

Analyzing Gene Fusions on NCI and St Jude Cloud - Jinghui Zhang (St. Judes)
Notetaker: Marcia Fournier (NCI)
Notes:

Goal: Use CICERO at St Jude Cloud to perform analysis of gene fusions in pediatric cancer to optimize analysis time/ cost for fast analysis of clinical samples.

- Performed analysis of components including discovery cohort (pediatric cancer), validation cohort (adult care), and negative control (health adult)
- Used gene fusion as biomarker for cancer diagnostics and treatment
    - Gene-re-arrangements (e.g. translocations)
        - Drug target
        - Risk assessment
        - Prognostic info
- 1.2 Pb of raw data (>11,000 patients), 93 institutions, 20 countries
- RNA seq BAM files – workflow looked at splice junction analysis, oncogenic activations, single exon deletions, gene fusion – fusion annotation and ranking.
- Challenge of meeting timeline of 15 days from lab, mapping, analysis processes to allow for rapid sequencing of clinical real time samples.
- Optimized processes using CICERO on St Cloud Cloud to allow rapid sequencing.
- Challenges with “ timeout “ (data failure) were resolved with optimizations of CICERO
- Reduced time by bringing mapping and analysis to the cloud to meet timeline.
- Looked at alterations on MAP3K8 which targeted therapies are available
- Comparisons of exon 8 vs exon 9 using TCGA melanoma data for validation using NCI cloud genomics and CWL
- Tested in real life – 170 benchmark samples – reduced time to 2.5hrs/ $3.96 per sample
- NCI cloud performed with similar performance (~5hrs /$5 per sample)
- Planning to bring data to St Judes but now knowing of NCPI capabilities it makes more sense to interact with the NCIP ecosystem for interoperability

Q&A / Chat:

from Ian Fore (internal) to everyone: 11:48 AM

Kim, I believe this has something to do with what I raised yesterday about SRA access to data in buckets owned by the NCPI platforms.

from Valerie Cotton to everyone: 11:56 AM

Is CICERO accessible across NCPI platforms/workspaces?

from Michael Schatz to everyone: 11:57 AM

You should consider trying the new CHM13 T2T assembly which resolves the centromeres and telomeres. https://genomeinformatics.github.io/CHM13v1/

from Sharon Plon to everyone: 11:58 AM

There is increasing calls for avoiding the term "blacklist" given the potential historical connations of racism. Would suggest using term like filtered list. I believe that ENCODE has decided to make this change.

from Heidi Sofia to everyone: 12:03 PM

I agree wiht the need to avoid "blacklist" and "whitelist". How about "stoplist" and "golist"?

from Allison Heath to everyone: 12:03 PM

Noted your use of the BAM slicing feature, I was wondering if you felt this was a generally useful feature, especially for viewing/reprocessing read-level data for specific genes instead of the whole BAM/CRAM file?

from Ian Fore (internal) to everyone: 12:04 PM

Where was CICERO run in the Cancer Genomics Cloud? Seven Bridges? Would be a good candidate to try out through GA4GH WES.

from Kathy Reinold to everyone: 12:04 PM

Good point Sharon!

from Allison Heath to everyone: 12:06 PM

Are there plans to make the cicero on CGC a public app?

from Clay McLeod to everyone: 12:06 PM

Yes^

from Allison Heath to everyone: 12:06 PM

Thoughts on when? Would be a great test case across the NCPI? :)

from Ben Heavner he/him to everyone: 12:07 PM

more on moving away from racist language in technical vocabulary: https://thenewstack.io/words-matter-finally-tech-looks-at-removing-exclusionary-language/

from Brian Furner to everyone: 12:07 PM

"deny list" and "allow list" are terms i've seen in circulation

from Jinghui Zhang to everyone: 12:07 PM

Yes bam-slicing is very useful. We recommended this feature during the development of GDC so we are very happy to use that feature.

from Sharon Plon to everyone: 12:07 PM

Owen pointed out that this is what ENCODE now says. “The DAC Exclusion List Regions (previously named "DAC Blacklisted Regions”)”

from Jinghui Zhang to everyone: 12:07 PM

OK. Exclusion list.

Cloud-Based Whole Genome Sequencing Analysis Workflow - Xihong Lin (Harvard)
Notetaker: Melissa Cook (NCI)
Notes:

Overview of Whole Genome Sequencing in US
In next few years, will have tens of millions of whole genome sequence available

- Need for platform

Overview of Pipeline:
Must do analysis en masse (not on individuals)

- raw data functional annotation (more cost effective)
- Then enter into STAAR analysis pipeline
- Summary statistics for meta-analysis later

FAVOR (favor.genohub.org)
- Background database
- Then can use Online or Offline annotation

FAVOR Annotator Workflow
- Compression rate = 1000 times is more scalable and cost efficient
- FAVOR Database (3 billion positions)
- Then create aGDS

Annotator – scripts
- Backend database – run scripts on multiple sources

WGS Association Analysis Workflow
- Input
- Common variants single SNV analysis
- Rare Variants SNV-Set Analysis to ensure novel discoveries

STAARPipeline Workflow for RV analysis
- Gene-Centric analysis

Implemented in Terra
- Raw genotype (VCF) file
- Use FAVOR backend database to annotate
- WGS Association Analysis – use dbGaP for Phenotype
- Then perform the workflow analysis thru STAARpipeline
- Have a STAARtopmed applet – used widely

Demonstrated STAAR App in analysis commons
- Working on it with Biodata Catalyst and topmed

Can finish analysis in less than a day

Cost estimate & time estimate analysis compared by Method

- Total cost ~$1000

Challenges

- Data access – labor intensive (dbGaP, data use restriction – can take a long time to get letter of agreement, phenotype harmonization very labor intensive)
- Cost –data storage and computing costs are much more than Computing Clusters
- Analytic platforms – need for

Challenges on their side –

- Visualization – Whole Genome Sequencing – working on this
- Meta-analysis – efficient and scalable workflow for rare variant meta-analysis
    - ortals
    - Summary statistics
    - Working on this, ex. Type 2 Diabetes Knowledge Portal
- Other analysis portals

See slides
Questions will be taken later

Chat:

from Michael Schatz to Everyone: 11:57 AM

You should consider trying the new CHM13 T2T assembly which resolves the centromeres and telomeres. https://genomeinformatics.github.io/CHM13v1/

from Sharon Plon to Everyone: 11:58 AM

There is increasing calls for avoiding the term "blacklist" given the potential historical connations of racism. Would suggest using term like filtered list. I believe that ENCODE has decided to make this change.

from Heidi Sofia to Everyone: 12:03 PM

I agree with the need to avoid "blacklist" and "whitelist". How about "stoplist" and "golist"?

from Allison Heath to Everyone: 12:03 PM

Noted your use of the BAM slicing feature, I was wondering if you felt this was a generally useful feature, especially for viewing/reprocessing read-level data for specific genes instead of the whole BAM/CRAM file?

from Ian Fore (internal) to Everyone: 12:04 PM

Where was CICERO run in the Cancer Genomics Cloud? Seven Bridges? Would be a good candidate to try out through GA4GH WES.

from Kathy Reinold to Everyone: 12:04 PM

Good point Sharon!

from Allison Heath to Everyone: 12:06 PM

Are there plans to make the cicero on CGC a public app?

from Clay McLeod to Everyone: 12:06 PM

Yes^

from Allison Heath to Everyone: 12:06 PM

Thoughts on when? Would be a great test case across the NCPI? :)

from Ben Heavner he/him to Everyone: 12:07 PM

more on moving away from racist language in technical vocabulary:
https://thenewstack.io/words-matter-finally-tech-looks-at-removing-exclusionary-language/

from Brian Furner to Everyone: 12:07 PM

"deny list" and "allow list" are terms i've seen in circulation

from Jinghui Zhang to Everyone: 12:07 PM

Yes bam-slicing is very useful. We recommended this feature during the development of GDC so we are very happy to use that feature.

from Sharon Plon to Everyone: 12:07 PM

Owen pointed out that this is what ENCODE now says. “The DAC Exclusion List Regions (previously named "DAC Blacklisted Regions”)”

from Jinghui Zhang to Everyone: 12:07 PM

OK. Exclusion list.

from Clay McLeod to Everyone: 12:08 PM

It's largely implemented, we're currently in the process of doing the final scientific validation to be sure it produces the results we expect. I would expect 1.5-2 months before it's public.

from Allison Heath to Everyone: 12:10 PM

@Jinghui and we implemented it ;) but I haven't see anywhere else other than GDC for NCPI (or GA4GH), it's always been about the whole files.. but it's relatively easy using byte offset, so was curious if others had these use cases, or if people just deal with getting the whole file... ?

from Allison Heath to Everyone: 12:10 PM

@Clay great!

from Jinghui Zhang to Everyone: 12:11 PM

Allison, that sounds great. We will try it. We are working on getting access to KidsFirst data now:)

from Ian Fore (internal) to Everyone: 12:11 PM

@Allison - would htsget do it?

from Allison Heath to Everyone: 12:15 PM

@Ian ah yeah, you're right htsget has start and end parameters, so it should, so maybe it's about implementing htsget to move more towards a standard

from Ian Fore (internal) to Everyone: 12:18 PM

@Allison NCPI has lots of DRS implementations at the moment - htsget probably worth a look.

NCI CRDC Center for Cancer Data Harmonization efforts - Sam Volchenboum (UChicago)

Notetaker: Annie Kuan (Broad)

Slides link

Notes:

- When building a data commons ecosystem, start with data models before beginning to collect data, but the CRDC started as more of a grassroots program with each node having its own data model, tools, and APIs
- In general, there is a disparity between the existing data commons and the resources in terms of how the data is modeled and harmonized.
- **CCDH's goal:** To build a common data model and tools and services to help harmonize data across the CRDC, especially important for downstream applications such as aggregation via CDA then further use by the CRs
- CCDH has been focusing on:
    - Harmonized data model
    - Building terminology services
    - Building integrated terminology and code sets
    - Aiming for data generated by the nodes to use these model and services, so then it will be streamlined for CDA for integrated search across the nodes
- CRDC-H:
    - Took information from the data models from each of the nodes. Built and aggregated data model, then refactored into a common data model, and currently, refactoring again to make the first release of CRDC-H coming out end of May
    - This doesn't mean every group needs to use this model, but the CCDH team will be advocating for nodes to use CRDC-H
- Next few weeks (end of May): Releasing Biospecimen and Administrative subdomain entities, along with select Clinical subdomain entities. Terminology bindings will be included
- How to harmonize data across the NCPI?
    - Want to be mindful of the ramifications of moving data from one place to another during harmonization to a common data model.
    - Make sure we bring in all the data modalities.

    - Want consensus around how to pull everything together and how we represent individual patients across the ecosystem
- LinkML: Data modeling language, helps data be “born interoperable”
    - Has a Simple YAML as the source of truth, but can generate many different “flavors” (e.g. JSON Schema, Python Dataclasses, etc.)
    - Focusing on modeling the components we can agree upon as a community
    - Learn more about LinkML at the github: https://linkml.github.io
    - LinkML has the RDF hidden in plain sight. Uses RDF to deliver the different aforementioned “flavors”
    - One issue we are addressing: Not having access to the resolved enumerations. In LinkML, URIs resolve to the enumerated identifiers
    - Terminology Services: TCCM (Terminology Common Core Model) has full logic interoperability
    - Value Mapping Graph Model: Mapping files have full provenance relating codesets and values
    - LinkML will be able to also push to things like FHIR, so can use this as the central component to relate models and FHIR/Big Data tool/etc. Also working on adding tabular formats.
    - Using the CRDC-H LinkML - will support the transformation of existing CRDC data into this common data model. Looking forward to working with PFB in the validation step.
- Main takeaways: See slide deck

From the chat:

rom Ben Heavner he/him to Everyone: 12:27 PM

Are there general purpose tools for validating that submitted data aligns to these models, or do validation tools need to be developed by some responsible party?

from Sharon Plon to Everyone: 12:31 PM

Helpful to think about challenge when linking data on the same participant in multiple places (like BASIC3 and other studies) as well as how to link unique data sets that are used for comparison studies.

from Adelaide Rhodes to Everyone: 12:31 PM

@Ben I think it is the latter depending on the use case

from Brian Walsh to Everyone: 12:34 PM

Great presentation. Love the mapped_from -> mapped_to.

from Brian Walsh to Everyone: 12:35 PM

Assume that is generating a FHIR schema?

from Sweta Ladwa to Everyone: 12:36 PM

+1 on using FHIR for FHIR exchange and creating a schema to support the semantic mapping! Thanks Melissa

from Ben Heavner he/him to Everyone: 12:36 PM

:D to validation tools

1:00-1:20pm – Community Interoperability Talks Discussion

Group discussions on topics covered in morning Community Interoperability talks led by Adam Resnick (CHOP)

Notetaker: Marcia Fournier (NCI)

Notes:

Intro:

- Community focus: the major purpose of the NCPI interoperability is to support the community. Foster collaborative research and implementation across the data.
- Focus of the last six months: Focus on suggestions, recommendations, implementations for community based – discovery.
- Common themes: Common approaches, combined data sets, and annotations.
- Generation of "data products": e.g. workflows, CICERO (portable?), fusion analysis and copy number integration, etc

Discussions: Adam Resnick, Sharon Plon, Owen Hirschi, Allison Health, Sam Volchenboum, Brian Furner.

- Clinical applications (e.g. St Jude) – while these are clinical studies done under informed consent - building tools and analysis that can impact patient care.
- Data products - Community activities bringing impact on patients.
- Industry/ commercial labs are major players but often left out of the "community"
- Data product: Having community facing aspect of the data is important. The community is ready for it based on high volume of cases at St Jude.
- Portable applications: nice about CAVATICA – create a CAVATICA application and build up from there
- How do you make data searchable – empower analysis
- Suggested sub-committee /working group to ensure data product is searchable and usable – not only data sources.

- Case level: How to search at case level with data being in different places. Same patient with unique/universal identifiers. Having a way to identify samples related to the same patient.
- Track data across platforms
- Identifiers: Concept of longitudinally.
- Key unmet need – being able to -re-query/ re-look data sets across platforms
    - These talks are ongoing
    - Trying to bring groups together
    - Example of CICERO ID – same patient ID is not the same in other systems.
- Concern with generation of "new" datasets by combining data sets (e.g. genomic and proteomics) – patient informed consent may not cover use of combined data.
- Future consideration: Search infrastructure vs. pipeline perspective

3:00-3:20pm - The Future of Interoperability - Speaker: Brian O'Connor (Broad)
Notetaker: Annie Kuan (Broad)
Notes:

- Goal of NCPI - Breakdown data silos across the NIH, moving towards a federated data ecosystem
- Starting point in 2020 was data portals connected intra-IC with analysis systems (workspaces). So wanted to move toward data portals connecting *inter*-IC and have workspace access data inter-IC. A cross platform way to handoff data within NCPI
- Using the GA4GH standard for handing off data using DRS and PFB. This came out of the FHIR working group to assess best way enable interoperable data sharing
- Going into the end of 2020, see both BioData Catalyst and AnVIL enable PBF handoff = 417k subjects. All 4 platforms now support data access via DRS 1.1 = 7.6PB of data. Also seeing AuthN via RAS being rolled out across the platforms
- In early 2021, we have now demonstrated handoff from all 4 portals to Terra and SBG workspaces. DRS has been key in enabling this work
- Supported Researcher Use Case:
    - Tim Majarian's cross data analysis
    - He was able to look at and create groups of patients from AnVIL and Kids First, compute on the data using CWL in Terra.
    - The work that NCPI has done so far made Tim's work that much easier and streamlined to execute

- Focus for 2021
    - From the last FunRetroBoard at the last NCPI workshop, themes were:
    - **1) Authorization and policy** (happening now) - RAS, data access via Passport+Visas. Need to make technical decisions for the design and policy decisions
    - **2) Search** (next 6 months) - User should be able to search across NCPI systems. Want to leverage common API standards for searching across projects. Need to continue to assess best search API. Still need to scope out context, common data model, etc
    - **3) Portable compute** (next 12 months) - What about data enclaves where data can't exit or want to avoid egress? Can we enable sending algorithms to the data? Maybe leverage workflow execution from GA4GH. Can we make our workspaces mobile (beyond just workflows)? Need to think about how to package the full compute environment beyond what Docker can do right now for tools
    - Need to expand use cases for both individual researchers as well as cross institute. Use these to drive work forward

From the chat:

from Ben Heavner he/him to Everyone: 3:10 PM

Is the change in scope from "search for data sets" to "search for data" in search intentional?

from Ben Heavner he/him to Everyone: 3:11 PM

"search for data elements" is much larger scope, particularly given dbGaP operational unit of "consent group".

from brandi to Everyone: 3:18 PM

In the framework of portable compute / workspaces where do the results of an analysis (derived data) end up? what about when you combine data from multiple sources--- this likely touches on both governance as well as the tech

from Ankit Malhotra AWS to Everyone: 3:19 PM

link to the retroboard ?

from Valentina Di Francesco to Everyone: 3:20 PM

https://easyretro.io/publicboard/hFgFa2Yzo2XY8FPRtn8YHgvQ5Ur1/0304c214-7227-42ef-aee7-13fff66bcafe?sort=votes

from Brian Walsh to Everyone: 3:20 PM

subdivision by underlying tech? search w/ fhir ; search

from Brian Walsh to Everyone: 3:20 PM

subdivision by underlying tech? search w/ fhir ; search w/ ga4gh

from Kathy Reinold to Everyone: 3:21 PM

Probably several use cases there.

from Kathy Reinold to Everyone: 3:22 PM

In some cases, it might be a list of subject IDs or sample IDS or DRS IDs and in other cases PFBs

from Kurt W. Rodarmer to Everyone: 3:22 PM

Where do I read about PFB?

3:20-4:20pm – Breakout Groups Report Back

Topic 1: Data/Tools/Workflow/Compute Interoperability and Functional Equivalence - Speakers: Jack DiGiovanna (Seven Bridges) & Michael Schatz (JHU)

Notetaker: Jack DiGiovanna (SBG)

Notes:

Presentation (DiGiovanna + Schatz)

- Representative use cases
    - Different workflow languages supported on different platforms (exclusively?)
- How do I interact with data on different platforms?
    - Bring me the data: egress, enclaves, and bears!
    - Rewrite all the code in OTHER_DESCRIPTION_LANGUAGE
    - Don't use the data?
    - Send the compute to OTHER_PLATFORM?
    - Universal workflow engine?
        - Interested to develop it, but many months to years away
- How to send compute?
    - Establishing accounts is straightforward
    - AuthN/Z is working (RAS)
    - How to talk to other platforms?
        - WES: “correct choice” but potentially less powerful than native APIs (e.g. Terra API, SB API, Gen3 API)
- Does the pipeline exist?
    - Featured workflows, Dockstore
- Are the pipelines functionally equivalent?
    - “Exact matches” should be interchangeable
    - But often there is not an exact match
    - How do we even measure functional equivalence?

        - Sequence the same sample (GIAB) multiple times to measure sequencing variance
        - Process the same sample using multiple pipelines to measure variance in compute

Postdoc hat

- What are the most important use cases? Alison Health
    - Long Reads are an important use case
    - What does functional equivalence mean?
        - How to consider multiple reference genomes, annotations, sex chromosome
        - Housekeeping genes tend to be more stable
        - How to compare kids first to topmed
        - Move to new reference genomes is a huge barrier
            - Still get asked to use hg19
    - Kathy Reinold
        - We don't use the same names for each reference
        - Time and costs for rerunning analyses
    - Kyle Ellrott
        - Need a whole knowledgebase to effectively liftover
        - WES - abstract workflows
        - TES - input/outputs fully defined, command lines define
            - Cromwell TES, Galaxy TES, Snakemake TES
            - Need to define TES endpoints for different environments
        - Multiple technical solutions here, but the user would need to know where it was running.
    - Ian Fore
        - Workflow challenges are analogous to language choices
        - Can docker help solve this challenge?
            - Need to layer the analysis
    - Xihong Lin
        - CCDG/TopMed (Freeze 2) - high agreement on **callset (99.6%) after QC**, this seems more important than individual read mapping or other aspects.
            - Broad called 60k samples, then shared the bam with U Mich.
            - U Mich did JC + QC.

        - Association results were very consistent. Functional equivalence based on research application: call set, associations observed
        - QC Pipelines are critical.
        - Joint calling required for all the samples, done at Broad
    - Usually the callset is Program-specific (e.g. KF callset, GTEx callset), this is a pain point

Reviewer 2 hat

- # of CPUs and Random seeds can change the output
- What can we learn from PCAWG?
- Cyclic testing: same input on different machines/time of day may produce different outputs
- Checker workflows (in Supplemental Materials)
    - Known inputs and workflows to run on multiple platforms)
    - Exact same code with new data may lead to different outcomes. So need some more test data that's known but novel (e.g. slightly different format in a vcf annotation
- Mike has investigated this deeply for a tool that is robust to this. However, overall it's a mess for most tools. Is this something NCPI should look into?
- [good datasets] Kids First has a set of highly validated samples for this
- GIAB is another common dataset.
- SNVs are most stable, indels are okay, SVs are a mess
    - PCRfree vs PCRplus will subtly change. Should NCPI look into this?
- State of the art in SNPs? KF is experimenting somatic-style, first pass with GATK then consensus calling considering other callers as well.

Reproducibility of SNP calling or reproducibility of downstream analytics?

- We should be concerned on both
- t-SNE is stochastic by design. Deep learning often gives slightly (or huge) differences when run/trained multiple times
- We are doing this to find patterns, which we expect/hope to validate in other domains
- Should we require replicates in the analysis? Inspired from bulk-RNA Seq community. They are required to understand intrinsic stochasticism.
    - This is costly, but it's not on all the data - it's a subset of known files.

- External datasets / external APIs are especially challenging as they may change at any time

Concerns about DRS pointing to the data accurately, theoretically it could, but are we sure. Not yet, but I think we could assume it's fine (or at least out of scope).

Scientifically, there's a lot of debate around how to represent different aspects and processing. Things are changing very rapidly.

- Can we provide a stable platform such that at least "we can say you were processing it correctly (or incorrectly but consistently) on Platform A vs Platform B"
- Run the ONT WGS pipeline on both platforms, what is the test data?
    - Obviously GIAB, but some limitations
- Run another Best Practice WF
- Key question for any of them, what's success? Likely we'll need to judge it for each WF. Is this a useful outcome
    - File conversions should look exactly the same, other "analyses" will likely show more variability
    - SOP for a few best practice workflows - what is the threshold for success, what datasets we should use?
        - GIAB for everything? Ideally there is at least one subset of data that's publicly available for this first check. Getting access even to GRU data could be a high activation cost.
        - KF annotated data, controlled access but GRU. Could be a collaborative research approach

Topic 2: Cloud Costs & Benchmarking – Speaker: Alex Baumann (Broad)
Notetaker: David Pot (GDIT)
Notes:

- Grant guidance, is there stock language that can go into a grant.
    - recommendations for funding agencies, how to put that into a grant
- Sharing of benchmarked results doing standard analyses. More than one datapoint
- **Guiding/educating our users on cloud costs & benchmarking**
    - **how do people educate**
        - Galaxy: popular tools, benchmarks? Systematic approach to benchmarking - lookup table/API - also look at historic data, and

        - use a tool called polyester to generate synthetic data, try combinations of inputs
        - Terra - open access data. exome, wgs. Run few times. Publish in featured work-spaces
- Educating users has to include a comparative analysis of on-premise vs cloud tradeoffs.
    - users need to be convinced versus on-prem - what advantage does cloud provide?
    - capital versus operational expenses of institutions
    - is it a top down directive to use the cloud, or grassroots?
        - cloud is a disruptive change - work required to get started - build it and they will come may not be enough
    - cost is just one criteria - if it's optional to move to the cloud, and they don't see advantage, it's hard to convince them to switch
    - highlight the advantages: sharing, one copy of data, existing tools
    - institutions have hidden costs that people don't consider when comparing to the cloud
    - early adopters needing scale or dataset via cloud
    - requester pays, helps egress
    - cloud expertise is an issue - white glove service has been enormously helpful here
- Carrots to come to the cloud - taxi meter running
    - consulting help
    - credits
    - flexibility
    - community
    - standardized and maintained pipelines - not tied to specific platform? could the same pipelines be run on-prem?
    - having a clear and predictable cost model
    - writing of grants
    - could NIH funding "price match" on prem costs with cloud costs? cloud insurance
    - white paper describing why move to the cloud won't work, what is relevant to me?
    - institutions can start charging for on-prem to get people used to model / show cloud is cheaper
    - need to see the costs as you spend them - reduce delays

        - technical solutions - monitoring virtual machines actively rather than wait for delay
    - time to do compute is a benefit
- Existing algorithms - benefits
    - methods readily available, don't need to start from nothing
    - machines can quickly been started/stop - flexibility
- Present back to user community of benefits. Time and compute/storage costs. CGC has knock-on effects (for example).
- Most success (Terra) - people doing something new. Is there existing content for it, with a little bit of customization (tweaking of parameters of existing pipeline).
- Early adopters are needed to seed the process - post-doc, getting something running. Needs hand-holding
- non-linearity for cost estimations can be problematic
- knowing what costs are most significant (SSD)
- setting alerts and alarms
- What are some success stories attracting people to the cloud?
    - how much has "existing content" worked out of the box?
    - Manisha: whiteglove support, then the researcher presenting back to their community with an apples to apples comparison of their experience - encouraging cost comparison, finding opportunities for them to speak - does it have viral effects? yes.
    - train the trainers - doing whiteglove for interested labs to get past initial hurdles - work with a small number of people who then train others in the lab
    - cloud credits that are well tracked so we can keep them on track for spending
    - less concern about "runaway spend" on storage
    - teaching people about the underlying infrastructure (GCP, AWS, etc)
- Recommendations
    - Increase NCPI training efforts, training on costs?
    - whiteglove help for viral growth to larger communities
    - keep up with free credits that are well tracked in near real time
    - example benchmarks across platforms for standard pipelines
        - error margin ok, what are costs for long term,
    - egress costs - secondary sharing of data - can $ be allocated, discounts?
    - find ways to make cloud solutions as equivalent as on-prem functionality
    - advertise availability of high value data (e.g. open access)
    - top down incentivization

        - language to write in grants - forcing requiring clouds be used for research
        - funding agencies provide additional costs if using cloud created surprise costs
        - price matching?
    - clearly communicate costs and define what error margins people are comfortable with
    - Anvil budget justification example: https://anvilproject.org/learn/investigators/budget-templates

Topic 3: Governance – Speaker: Bob Grossman (UChicago)

Notetaker: Stan Ahalt

Notes:

- Is there agreement on the two consideration:
    - Authorized environment consideration
    - Right to distribute consideration
- Yes, there seems to be general agreement about these two considerations but some wordsmithing is needed.
- Trust
    - Platform with the data (the data steward) have to make tough decision about deciding which platform have the security compliance and are known entities that can be trusted to distribute the data.
    - Trust is the responsibility of the data steward at the institute/center, usually the CISO (can be someone else).
    - David Burnett pointed out that most generally if you look at another system at NIH, if that has a much lower level of security should it be trusted?
    - As Data stewards such decisions have to be made
        - Plus side can interoperate
        - Minus if data should in general that flows into doesn't have the same level
    - RAS authorizes users, not system
- dbGap Agreements and Considerations
    - Nice in the future to clarify the compatibility of the systems.
    - Formalization of the usage of the NCPI platforms for analysis when signing the dbGap Agreement

- Opportunities
    - Once interoperability has been achieved, need trackability and traceability of the data once downloaded
    - There is interest in tracking data migration across systems and report back to stewards
- Clarifications for considerations
    - Being more specific about what platforms are trusted, which platforms can distribute the data
    - Being more explicit about what is needed in the current process support these considerations
- Take away
    - At the point where we are generating at the tail end of the concerns surrounding the consideration document
    - Many are agreeing to endorse theses
    - Some have reservations about the language that have been captured in the chat notes
    - Based on these suggestions, should be able to close out and have agreement
    - Broad announcement is needed once the tweaks have been made and publish it
- Discussion
    - Valerie - factors can be different for every system. Need to continue to think through and navigate

4:20-4:30pm – Conclusion of meeting - Melissa Haendel (U of Colorado) and Tanja Davidsen (NCI)
Notetaker: Alan Zheng (NCI)
Notes:

[Melissa]

- In this big data ecosystem that we are collectively trying to create, what should we think about the most important and most pressing issues, how can we prioritize the various aspects of interoperability in order to achieve our goals
- Fundamentally we need to really stop and think about what are the reasons, what are the other rationale, what are the requirements for what success looks

like. Interoperability for the sake of interoperability doesn't necessarily make science go better or faster

- We really need to think about what those requirements look like, and how we will know that we've achieved them. I urge everyone to fill the survey and give us your ideas about what this means to you, so we can plan the NCPI work, our collective resources work, and also next meeting for the fall
- Thanks everyone for the great ideas and presentations and demos

[Tanja]

- Thanks to all our speakers, our discussion leaders, breakout session leads, our note takers, and technical administrative support. And thank you Melissa and Sam for MC the past two days. And of course, thanks to all of the participants for the lively discussions.
- Reminder to our speakers - please send your presentation to me (Tanja). We will put all the slides together and make them available to all participants
- The dates for the fall NCPI workshop are **Oct 5th and 6th**, and it will be hosted by NHLBI and RENCI
- Please submit your feedback via the survey link

[Partial Survey Result]

- A lot of people like the breakout sessions and discussions
- Suggested topics for upcoming meeting
    - Interoperability
    - More data harmonization
    - Dataset search
    - More breakouts, more governance, junior investigator, real user experience, more postdoc involvement, and use case development
    - Integration with CRDC efforts
    - CCDH Slack channel is open to anyone
    - Ensure systems can access data from across NCPI platforms
    - Multimodal analytics
    - Enable examples like the external speakers make it genuinely open, let users decide where they want to work, including the components outside of NIH platforms
    - Make more data and tools accessible and usable