

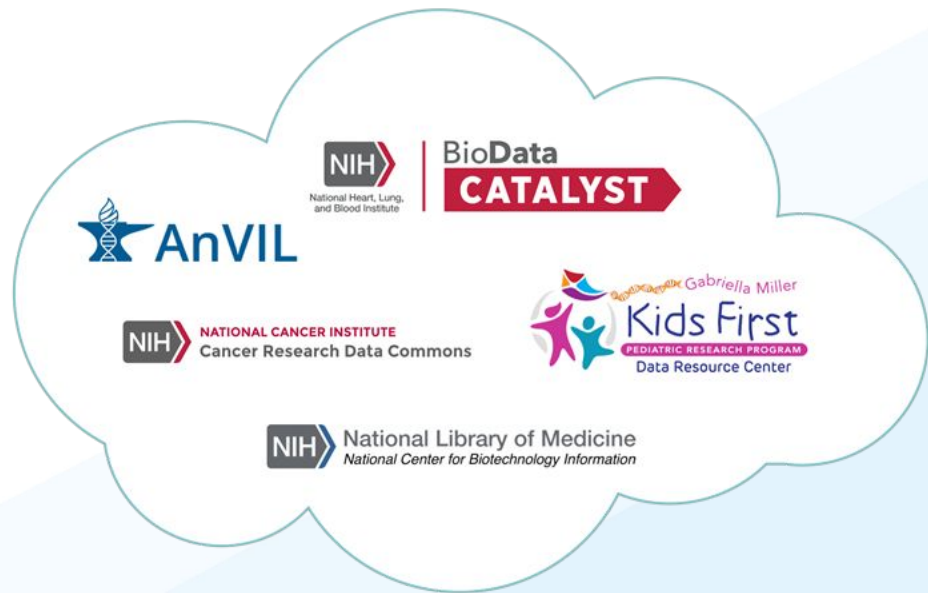
Welcome to the...

NIH Cloud Platforms Interoperability Spring 2021 Workshop

We'll be starting shortly!

May 3 & 4, 2021 11:00am-4:30pm EDT

tinyurl.com/NCPIagenda

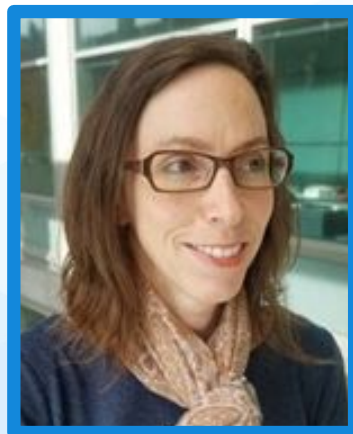


Welcome – NCPI Spring 2021 Workshop

Day 2

Melissa Haendel

University of Colorado



Tanja Davidsen

National Cancer Institute



Logistics

- Please use the **WebEx application** and not a browser
- Please mute when not speaking
- We will be recording all the sessions except the breakout sessions
- Notes will also be taken during the sessions
- Speakers please turn your camera on when speaking
- If you have not registered, please do: **tinyurl.com/NCPIregistration**
- Agenda: **tinyurl.com/NCPIagenda**
- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**

Agenda

Day 2: Tuesday, May 4

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

12:30-1:00pm – Break

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

1:20-2:30pm – Three Concurrent Breakout Groups

2:30-3:00pm – Break

3:00-3:20pm – The Future of Interoperability talk

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

4:20-4:30pm – Wrap Up

Interoperability is in the eye of the beholder



Legal/Licensing



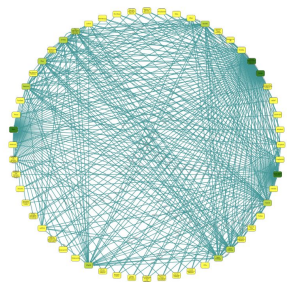
Regulatory



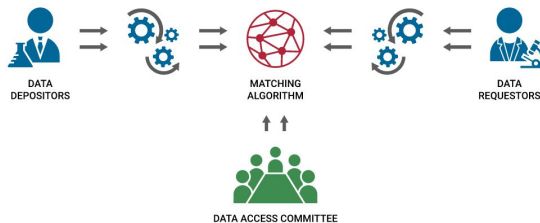
System



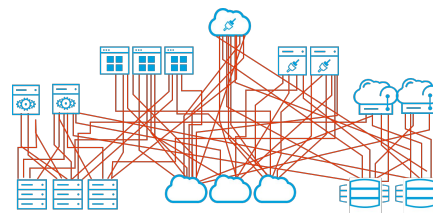
Data



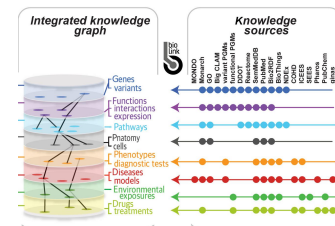
Restrictively licensed data can only be combined with permissively licensed data



Access control must match provenanced regulatory permissions



Platforms and tools often cannot talk to one another to move data and analyses



Data is often un-encoded or coded in different data models & terminologies, limiting search and integrated analytics

Achieving data interoperability

ONTOLOGIES	DATA MODELS	FORMATS	EXCHANGE
Semantic (data context)	Syntactic (data language)	System (data presentation)	Structural (data architecture)
...via pre-defined ontology concepts	...via pre-defined data models, data structures, data dictionaries, and data schemes	...via common data formats defined for encoding, decoding, and representation	...via networks, computers, applications and web services
Mondo, HPO, Snomed, Uberon, NCIt, ICD-O	OMOP, BRIDG, FHIR, LinkML, bioschemas, MIAME	OWL, RDF, VCF, FASTA, PFB	APIs, Docker

Proof of concept of interoperable approaches for improving outcomes of pediatric diseases

Tim Majarian

Computational Biologist, Broad Institute



Genetics of Congenital Heart Disease (CHD): improving outcomes of pediatric diseases

Study aims:

1. Identify, access, and summarize available genetic and phenotypic data through 3 cloud resources
2. Leverage individual-level data from multiple studies to assess the contribution of rare, exonic variants to CHD risk

Framework:

Genome Wide Association Study (GWAS)

Cases - KFDR PCGC CHD + TOPMed PCGC

Controls - TOPMed FHS & JHS

Follow up [TBD] - GTEx

Platform	Datasets	dbGaP	Sample	Use
AnVIL	GTEx	phs000424.v8.p2	980	In progress
KFDR	PCGC	phs001138.v3.p2	699	Case
NHLBI BioData Catalyst	TOPMed PCGC	phs001735, phs001194.v2.p2	1,901	Case
	FHS	phs000974.v4.p3, phs000007.v30.p11	4,155	Control
	JHS	phs000964.v4.p1	2,777	Control

Pediatric Cardiac Genomics Consortium

NHLBI-sponsored consortium focused on:

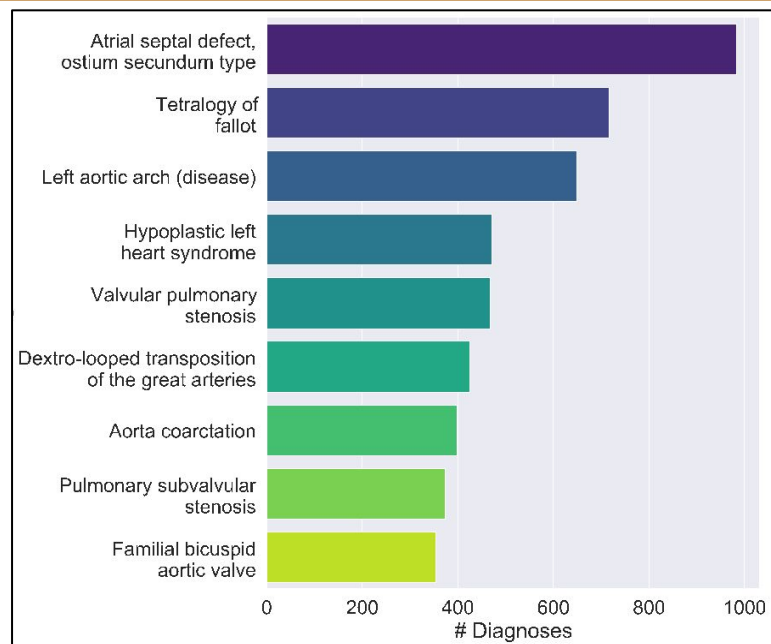
- Discovery of genes responsible for CHD
- Identification of genetic variants associated with CHD

Gabriella Miller Kids First Pediatric Research Program TOPMed

Congenital heart defects (CHD)

- Most common major human birth malformation
- 4-10/1000 live births
- 1 in 4 CHD cases is critical – require surgery or other procedures in 1st year of life
- Heterogeneous disease
- AHA lists at least **18 distinct types** of CHD
- Many cases of CHD due to chromosomal abnormalities (11% of patients)
 - ex: DiGeorge syndrome (60-70% have CHD)

CHD encompasses diverse clinical phenotypes



Previous studies focused on a case-parent trio framework rather than case-control

Advantages to Case-parent:

- Investigation of maternal + inherited genetic effects
- Avoid population structure + ancestral background confounding
- Shared environment

Disadvantages:

- Difficult to obtain large sample size

Solution leveraging interoperability:

- Combine datasets across multiple disease-focused studies
- Utilize large set of healthy controls through other consortia

Published: 09 October 2017


Contribution of rare inherited and *de novo* variants in 2,871 congenital heart disease probands

Sheng Chih Jin, Jason Homsy, [...] Martina Brueckner 

Nature Genetics 49, 1593–1601(2017) | [Cite this article](#)

6995 Accesses | 243 Citations | 200 Altmetric | [Metrics](#)

De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects

James R. Priest, Kazutoyo Osoegawa, Nebil Mohammed, Vivek Nanda, Ramendra Kundu, Kathleen Schultz, Edward J. Lammer †, Santhosh Girirajan, Todd Scheetz, Daryl Waggott, Francois Haddad, Sushma Reddy, Daniel Bernstein, [...] Euan A. Ashley  [[view all](#)]

Published: April 8, 2016 • <https://doi.org/10.1371/journal.pgen.1005963>

REPORT

De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies

Jason Homsy^{1,2,*}, Samir Zaidi^{3,*}, Yufeng Shen^{4,*}, James S. Ware^{1,5,6,*}, Kaitlin E. Samocha^{1,7}, Konrad J. Karczewski^{1,7}, Steve...

* See all authors and affiliations

Science 04 Dec 2015:
Vol. 350, Issue 6265, pp. 1262-1266
DOI: 10.1126/science.aac9396

A case-control study utilizing multiple cohorts: then vs. now

Pre-interopability effort

Data authorization

- Obtain dbGaP access
- Log into dbGaP
- Create download request

Access and localization to cloud platform

- Manual download & upload to cloud storage
- Access through cloud workspace

Data preprocessing & Final analysis

- Single cloud workspace

Current paradigms



Data authorization

- Obtain dbGaP access

Access and localization to cloud platform

- ERA credentials through Gen3 or KFDR
- Combination manual & automated data import to cloud workspace
- **DRS URIs available for all genetic data**
- **But requires manual upload & download of manifest**

Data preprocessing & Final analysis

- Separate workspaces within individual cloud ecosystems
- Export preprocessed files to single cloud workspace

Future

Data authorization

- Obtain dbGaP access

Access and localization to cloud platform

- Fully automated for multiple data repositories (BDC, AnVIL, KFDR)
- Through a UI in Terra

Data preprocessing

- One cloud workspace for all data
- Accessible through Seven Bridges or Terra

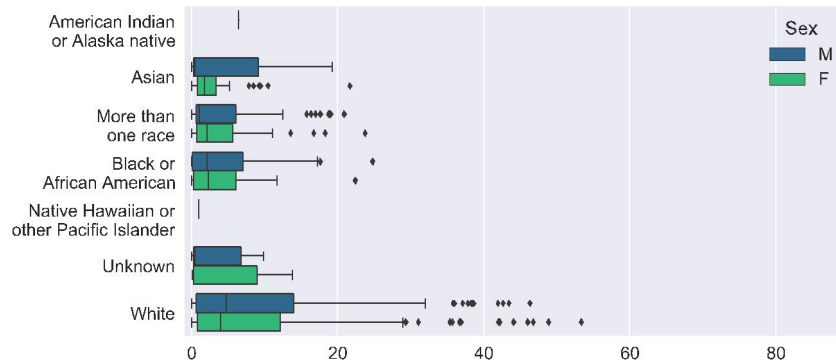
Final analysis

- One cloud workspace workspace
- No download and upload

Study population – PCGC, the Jackson Heart Study, and the Framingham Heart Study

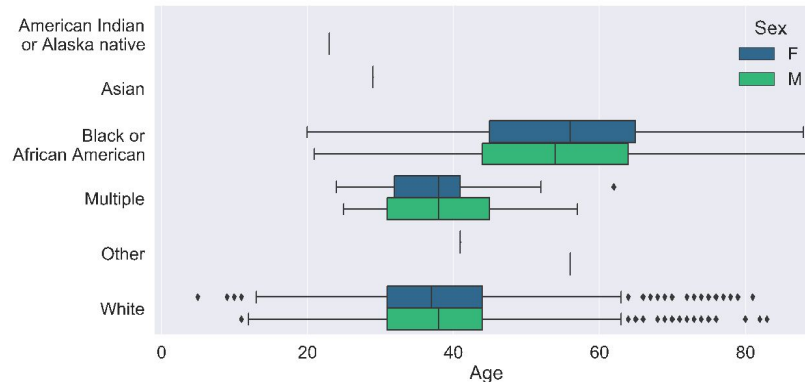
TOPMed PCGC & KFDR PCGC

- Probands only
- Unrelated (2nd degree or closer removed)
- Combined all samples & clinical diagnoses
- Whole Genome Sequence
 - Genotype & variant calling performed separately
- N = 1130



TOPMed JHS & FHS

- Unrelated (2nd degree or closer removed)
- Combine all samples with phenotypic data
- Whole Genome Sequence
- N = 6943

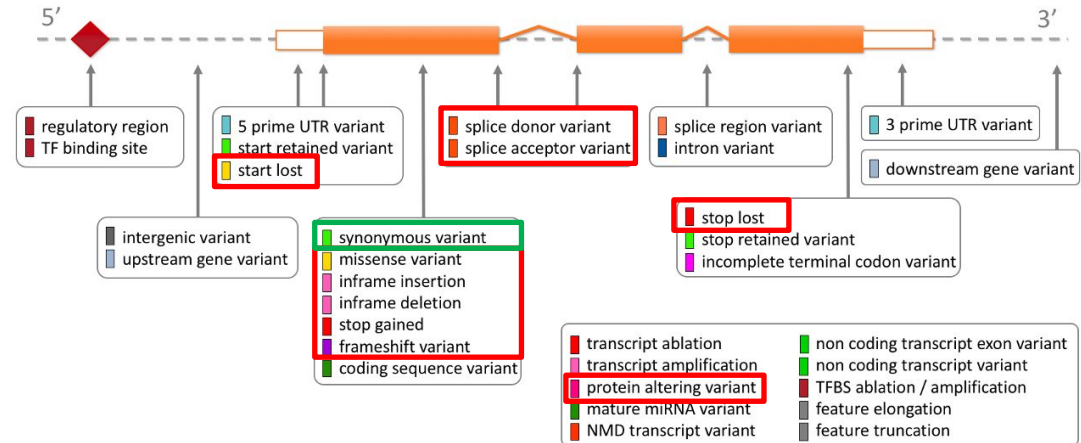


ProxECAT – Gene-based association testing using internal cases and external controls

Proxy External Controls Association Test

- Test for enrichment of rare variants within gene regions
- One p-value for each gene tested
- Non-synonymous (NS) alleles: VEP high + moderate impact
- Synonymous alleles (SYN) alleles serve as a *proxy* for how well variants are sequenced within the region

$$H_0: \frac{\# NS}{\# SYN} \text{ in cases} = \frac{\# NS}{\# SYN} \text{ in controls}$$



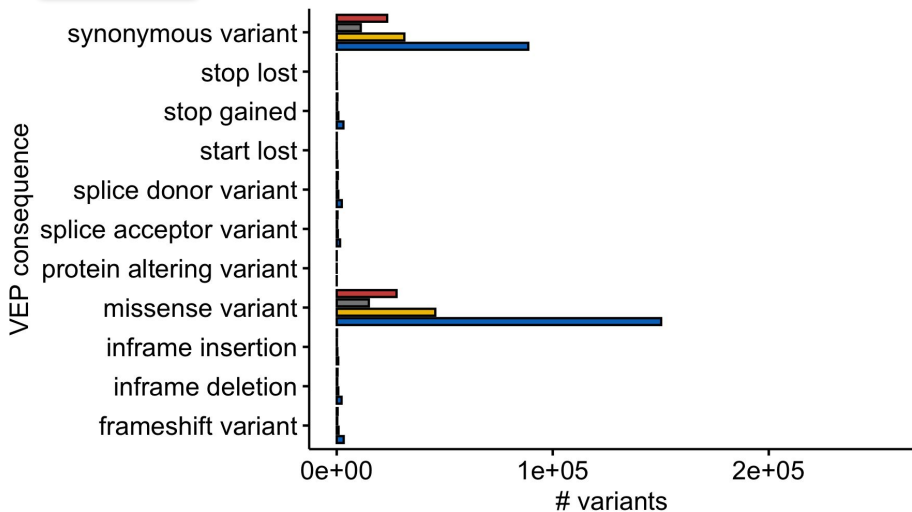
Cases and controls show similar patterns of allele frequency distribution among annotations

Variant annotation, aggregation, and ProxECAT association analysis performed in a Terra using the Hail software and genome aggregation database (gnomAD)

All data were imported using DRS from BioData Catalyst Powered by Gen3 and KFDR

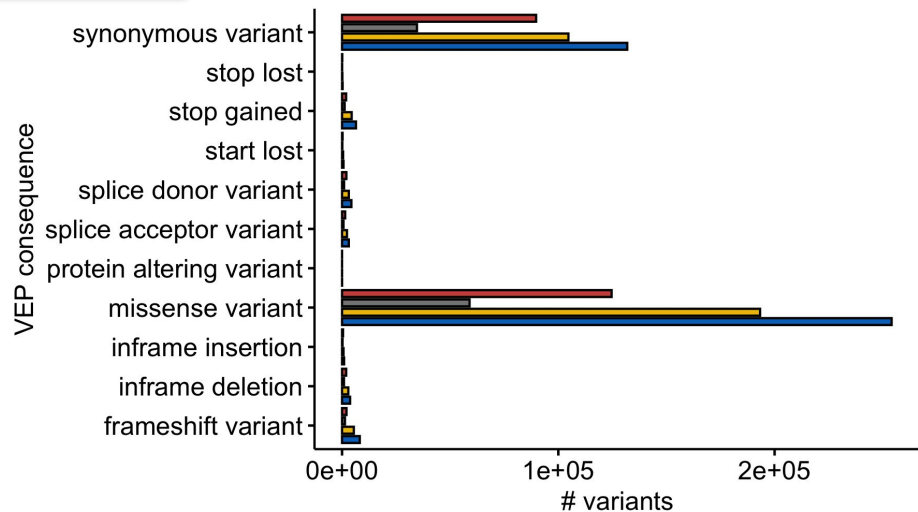
Cases

Frequency ■ MAC = 1 ■ MAC <= 3 ■ MAC <= 5 ■ MAF < 0.01



Controls

Frequency ■ MAC = 1 ■ MAC <= 3 ■ MAC <= 5 ■ MAF < 0.01



ProxECAT analysis shows no inflation, yields no significant gene-based associations

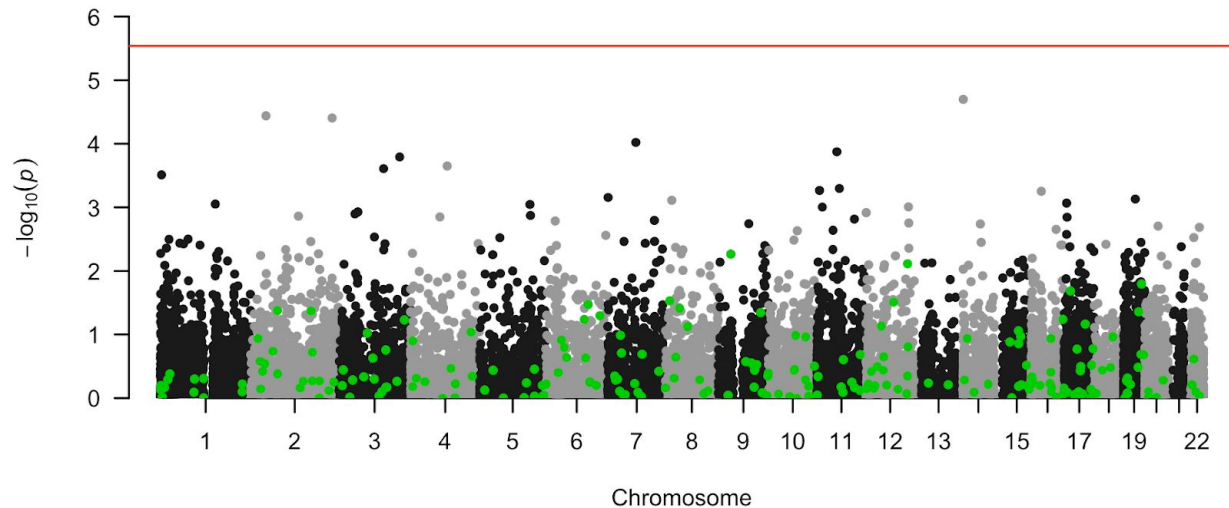
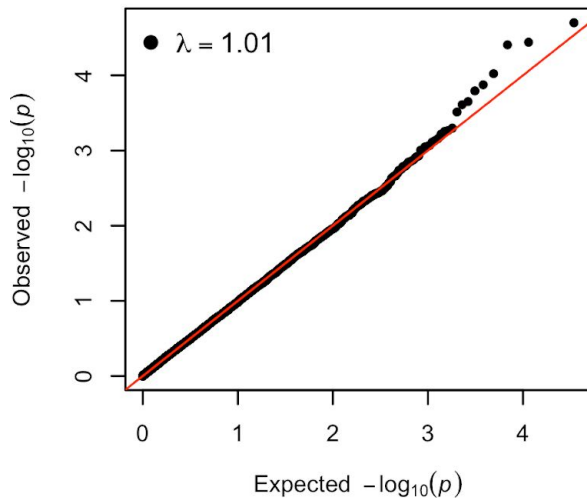
1130 cases, 6943 controls

18k genes tested

1.2M variants (780K NS, 420K SYN)

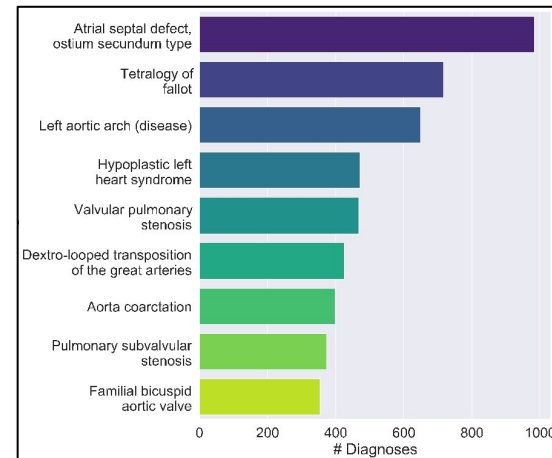
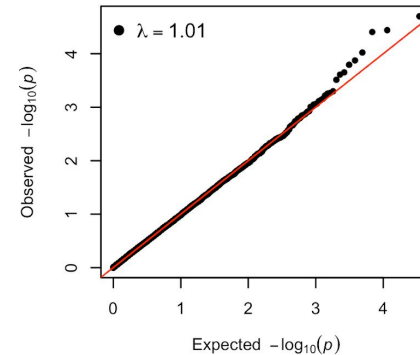
No significant associations; no evidence of confounding (GC = 1.01)

CHD genes from: Jin SC, et al. (2017) Nat Genet.



Why didn't we see any associations – heterogeneity and sample size

- No evidence of inflation in the test statistic
- CHD encompasses a diverse set of clinical phenotypes
 - Should we assume that these have a common genetic basis?
- Our statistical framework does not allow for covariate adjustment
 - population structure
 - ancestral background
- Relatively small sample size (1130 cases)
- SNVs and short INDELS only, no structural variants or chromosomal abnormalities





Conclusion and future work



Successfully leveraged genetic and phenotypic data from multiple cohorts to investigate the contribution of rare, exonic variants to clinically identified CHD

Interoperability tools allowed for data access and computation across distinct cloud platforms

- Most data access was automated (AnVIL, NHLBI BioData Catalyst)
- Some was manual (KFDR), although this should be automated soon

No associations observed using the ProxECAT framework but well behaved statistical analyses

More samples + more population diversity are needed to perform GWAS on CHD and CHD subtypes

- **More interoperability = more data sharing = more clinically relevant findings**

GTEx follow up analysis may yield further insights towards tissue and pathway enrichment of nominally significant associations



Acknowledgements



Alisa Manning
Brian O'Connor
Asia Mieczkowska
Becky Boyles
Patrick Patton
Steven Cox
Michael Baumann
Andrew Rula
Alex Baumann
Allison Heath
David Higgins
Maia Nguyen

Gabriella Miller Kids First Pediatric Research
Program of the Pediatric Cardiac Genetics
Consortium (PCGC)
Pediatric Cardiac Genomics Consortium (PCGC)
Genotype-Tissue Expression (GTEx) project
TOPMed's PCGC's Congenital Heart Disease
Biobank
Framingham Heart Study
Jackson Heart Study
BioData Catalyst Consortium
AnVIL

Community Interoperability Talk

Analysis of Childhood Cancer Patients (BASIC3 study) on the Kids First CAVATICA Platform and Other Clouds

Sharon E. Plon, M.D., Ph.D.

Baylor College of Medicine



Owen Hirschi

Baylor College of Medicine



Probands from 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

Outcome:

Initial diagnostic germline findings from WES

Autosomal dominant (P/LP)	N=26	19 different genes
Genes associated w/ specific childhood cancer	15	<i>Examples include DICER1, VHLx3, MSH2, WT1x2, TP53x3</i>
Genes not previously associated w/ specific childhood cancer	11	<i>Examples include BRCA1x2, BRCA2, PALB2, CHEK2x2, FLCN, SMARCA4</i>
Autosomal recessive (biallelic)	N=1	TJP2
No one gene was reported in more than 3 BASIC3 patients: 3 each for <i>VHL</i> and <i>TP53</i> .		

120 probands from BASIC3 selected for trio WGS



BASIC³

BCM Advancing Sequencing
Into Childhood Cancer Care

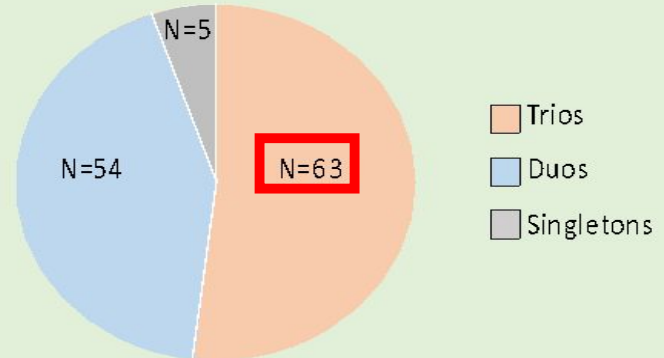
Goal: identify *de novo* SVs, SNVs, and putative pathogenic variants in known cancer genes missed by whole exome sequencing



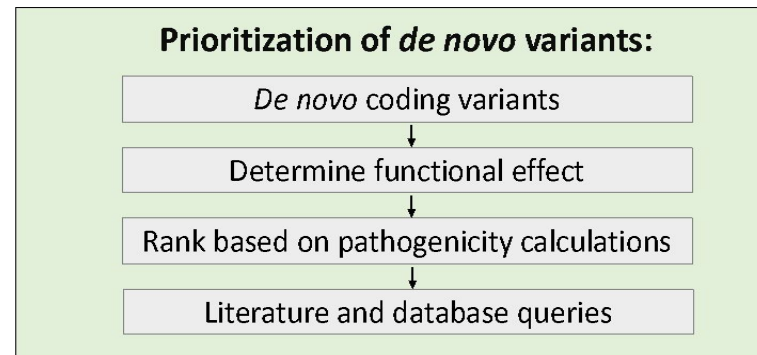
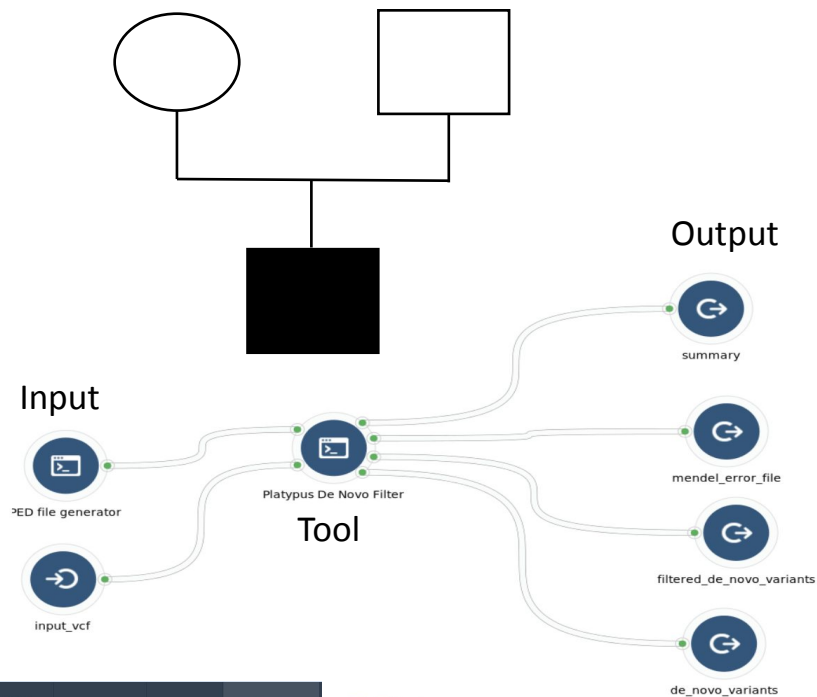
Breakdown of cancer type in cohort

Cancer Type	Frequency
CNS tumor	56
Non-CNS tumor	94

Breakdown of trios, duos, and singletons in cohort



Analysis on CAVATICA expedited *de novo* variant discovery



Outcome:

- SNV analysis completed on 54 proband-parent trios
- The pipeline resulted in an expected number of variants per trio

Variant Type	Frequency
Genome-wide <i>de novo</i>	60 to 190
Coding <i>de novo</i>	0 to 4

CAVATICA Projects Data Public Apps

Variant Effect Predictor

Created by admin on Oct. 18, 2018 11:26
Revision note: "label without version"

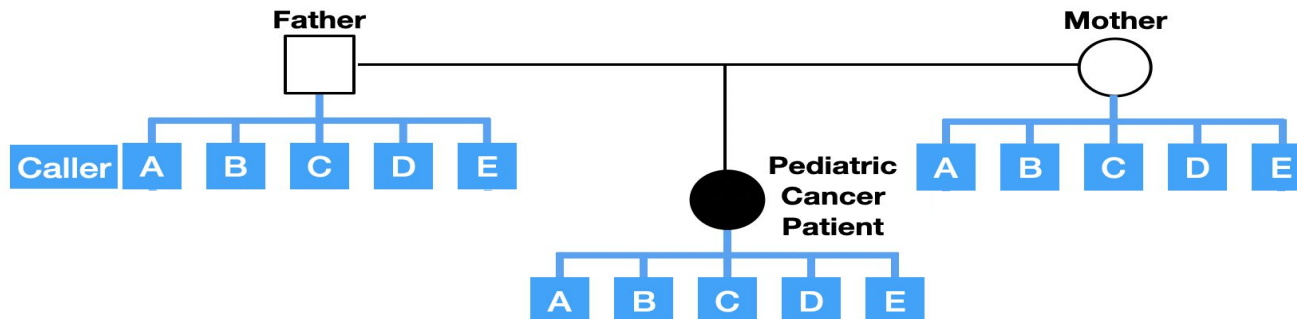
Platypus

Created by vojislav_varjasic on Mar. 12, 2018 06:46 • Last edited by vojislav_varjasic on Aug. 15, 2018 06:41
Revision note: "typo in JS fixed"

Description

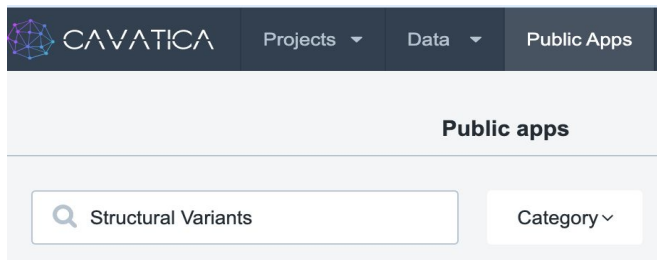
Platypus is a tool designed for efficient and accurate variant-detection in high-throughput sequencing data. Platypus reads data from BAM files, and outputs a single VCF file containing a list of identified variants, and genotype calls and likelihoods for all samples.

De novo structural variant analysis on CAVATICA



Caller A, B, C, D, & E:
Lumpy, Manta, Delly,
Breakdancer, & CNVnator

Analysis of SVs on CAVATICA requires multiple features of the platform



Explore genomics data

Understand complex genomics data with interactive analysis tools.



Data Cruncher

Analyze and explore data using JupyterLab or RStudio

Open

Completed

BATCH 165 Delly - Call run - 01-25-20 21:17:01

Executed on Jan. 25, 2020 15:22 by [owenhirschi](#) | Batch by: File

Spot Instances: **On** | Memoization: **Off** | Price: **\$68.37**

App: **Delly - Call - Revision: 0**

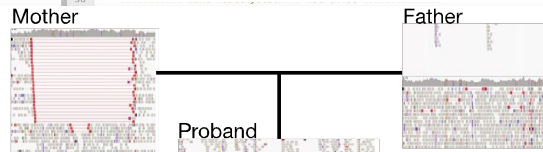
```
1 #!/bin/bash
2
3
4 #DEL
5 #=1
6 while [ $n -le 165 ]
7 do
8
9   # Stack all the filtered SV calls per sample,
10  # and perform some additional filtering
11
12  # update: small filter size change from 1kb to 100bp
13  # update: use new exclude region
14  # update: specify caller in script
15  # update: directly decide reciprocal overlap using new script
16  # update: remove sample name and role in output
17  # update: intersect with healthy control stack data to identify denovo variant
18  # update: need output that stack all healthy calls
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
```

Jasmine

JASMINE: Jointly Accurate Sv Merging with Intersample Network Edges

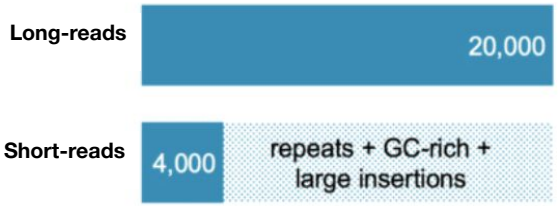
Version 1.0.11

IGV images



Long-read sequencing allows for greater detection of SV

Structural Variants Observed



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

Algorithms being utilized:

minimap2

minimap2 v2.17

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

ALIGNMENT | GENOMICS | LONG READS

CWL1.0

Copy Run

Sniffles CWL1.1

Sniffles 1.0.12b

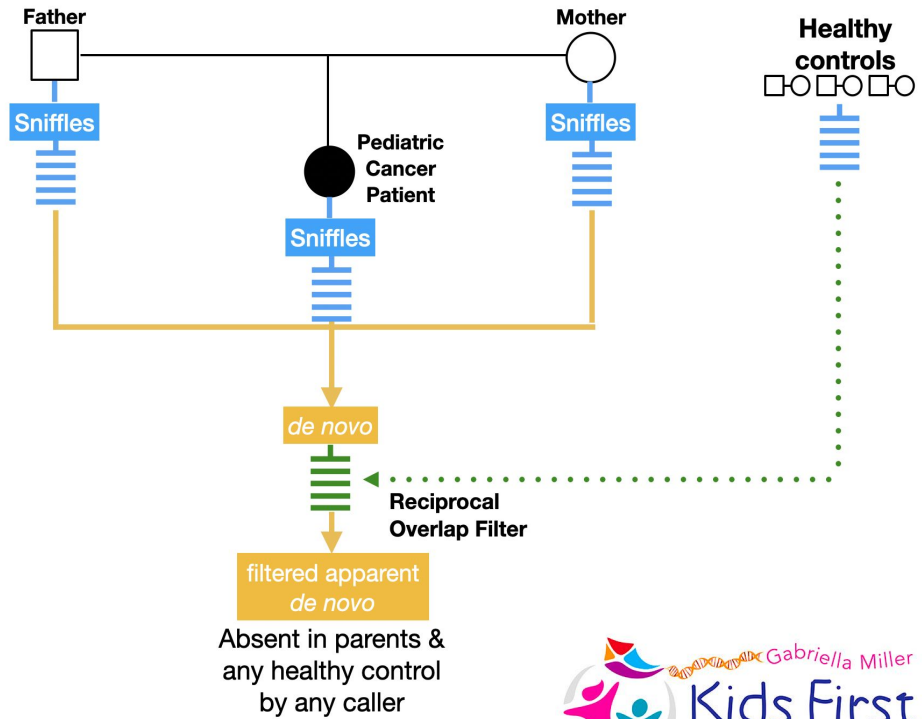
Sniffles is a structural variation caller for PacBio or Oxford Nanopore data [1,2].

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

VARIANT CALLING | CWL1.1 | GENOMICS

LONG READS

Copy Run



Kids First and CAVATICA enabled BASIC3 analysis

- We have been able to quickly and efficiently upload tools and analyze BASIC3 short-read WGS for *de novo* SNVs
- The CAVATICA platform has allowed us to use pre-existing applications and the terminal interface to create a novel pipeline for the analysis of structural variants in BASIC3
- Kids First has worked with us and others to upload tools in preparation for the analysis of BASIC3 long-read WGS

How to expand BASIC3 genomic analyses?

NIH Workshop on Cloud-Based Platforms Interoperability



BASIC³

BCM Advancing Sequencing
Into Childhood Cancer Care

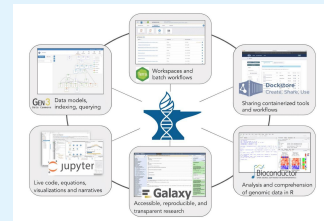


Clinical Variant
Interpretation

Germline Exome
Tumor Exome
Transcriptome

Other important resources

Follow-up study
Germline, Tumor
Exome
Transcriptome



Analyzing Gene Fusions on NCI and St Jude Cloud

Jinghui Zhang (St Jude)

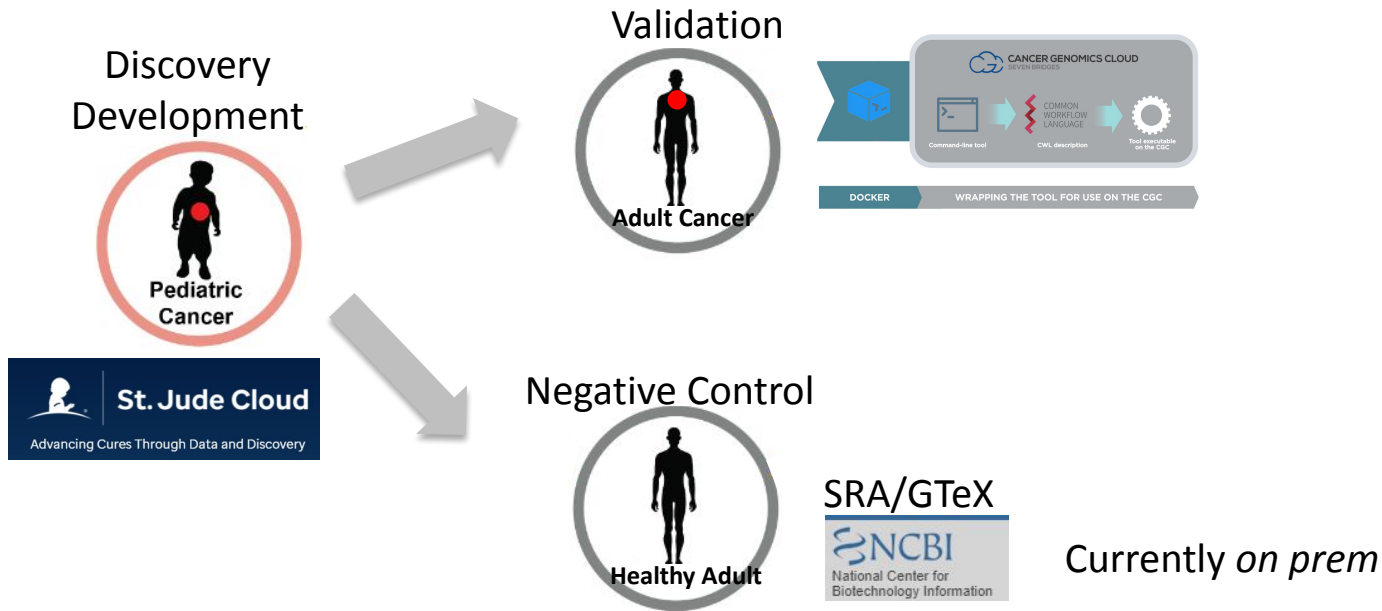
Use Case: Analyzing Gene Fusions on NCI Genomics Cloud and St Jude Cloud

Jinghui Zhang, PhD

Chair, Member

Department of Computational Biology
St. Jude Children's Research Hospital

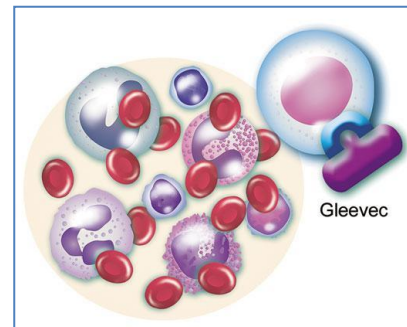
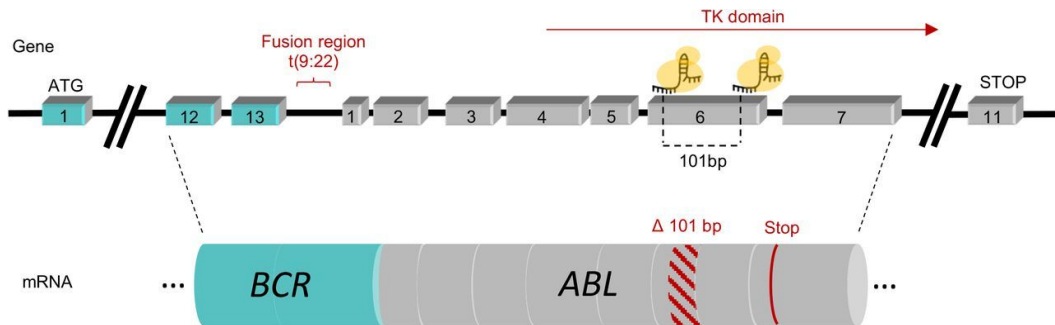
Overall Workflow



We demonstrate this process using gene fusion detection as an example

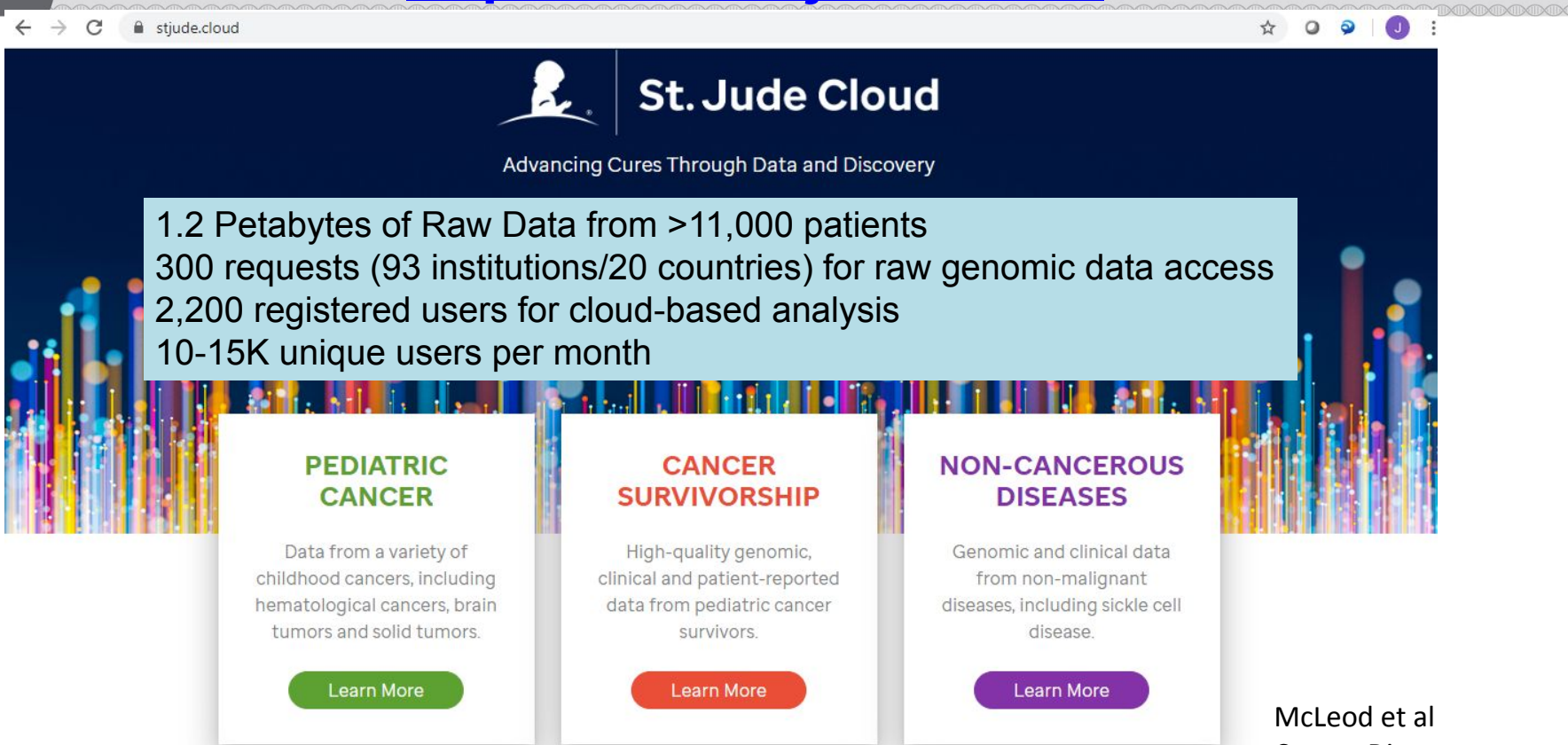
Why Gene Fusion?

- Gene fusions are Important biomarkers for cancer diagnosis and treatment
 - ✓ They can be cancer initiating event resulting from chromosomal re-arrangements (e.g. translocation, inversion, tandem duplication).
 - ✓ Used for risk stratification/ subtype classification in pediatric cancer treatment
 - ✓ They are one of the most targets for precision oncology



Data Sharing on St Jude Cloud

<https://www.stjude.cloud>



The screenshot shows the St. Jude Cloud website interface. At the top, there is a navigation bar with the St. Jude logo and the text "St. Jude Cloud" and "Advancing Cures Through Data and Discovery". Below this, a light blue box contains the following statistics:

- 1.2 Petabytes of Raw Data from >11,000 patients
- 300 requests (93 institutions/20 countries) for raw genomic data access
- 2,200 registered users for cloud-based analysis
- 10-15K unique users per month

Below the statistics, there are three white boxes with colored headers and "Learn More" buttons:

- PEDIATRIC CANCER** (green header): Data from a variety of childhood cancers, including hematological cancers, brain tumors and solid tumors.
- CANCER SURVIVORSHIP** (red header): High-quality genomic, clinical and patient-reported data from pediatric cancer survivors.
- NON-CANCEROUS DISEASES** (purple header): Genomic and clinical data from non-malignant diseases, including sickle cell disease.

McLeod et al
Cancer Discovery

CICERO for Complex Fusion Detection

Tian et al. *Genome Biology* (2020) 21:126
https://doi.org/10.1186/s13059-020-02043-x

Genome Biology

METHOD

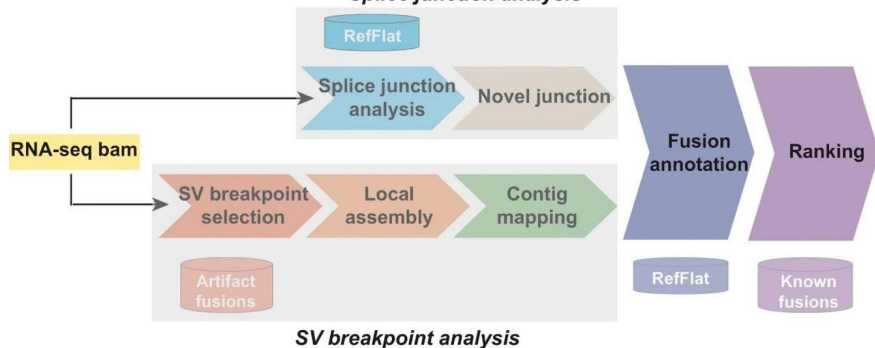
Open Access

CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data



Liqing Tian^{1†}, Yongjin Li^{1†}, Michael N. Edmonson¹, Xin Zhou¹, Scott Newman¹, Clay McLeod¹, Andrew Thrasher¹, Yu Liu^{1,2}, Bo Tang³, Michael C. Rusch¹, John Easton¹, Jing Ma³, Eric Davis¹, Austyn Trull¹, J. Robert Michael¹, Karol Szczytla¹, Charles Mullighan³, Suzanne J. Baker⁴, James R. Downing³, David W. Ellison³ and Jinghui Zhang^{1*}

Splice junction analysis



The NEW ENGLAND
JOURNAL of MEDICINE

HOME

ARTICLES & MULTIMEDIA ▾

ISSUES ▾

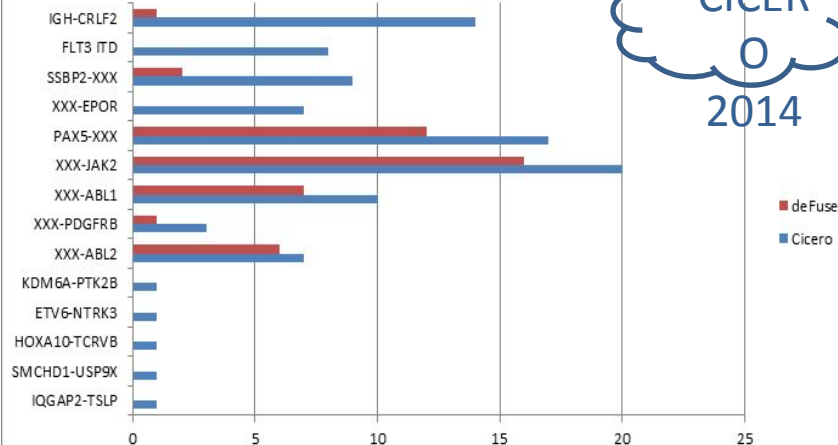
SPECIALTIES & TOPICS ▾

FOR AUTHORS ▾

CME ▾

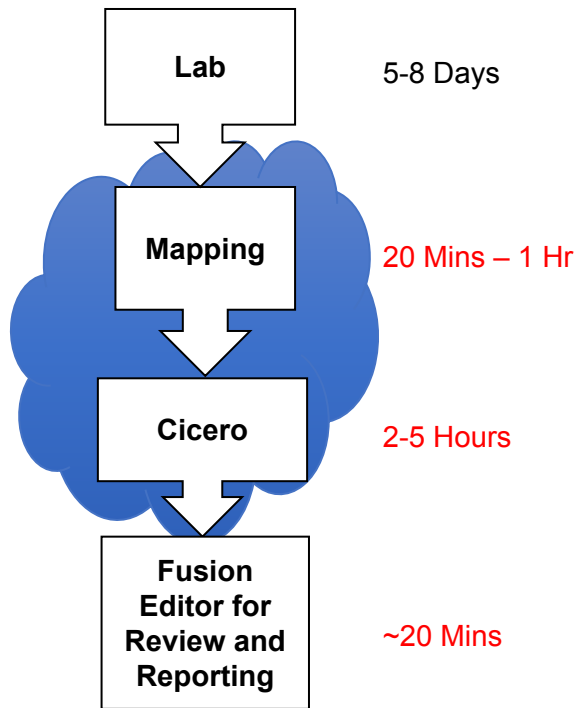
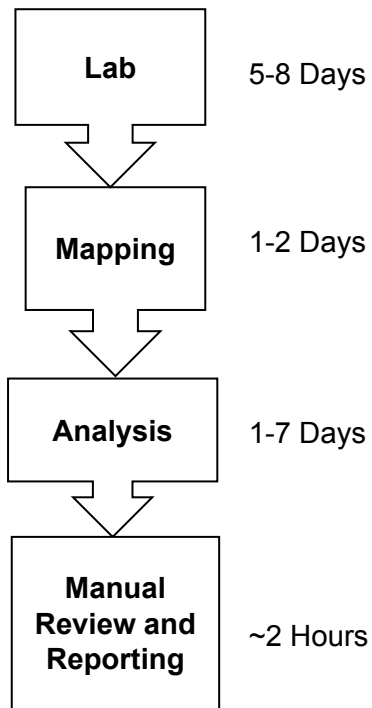
ORIGINAL ARTICLE

Targetable Kinase-Activating Lesions in Ph-like Acute Lymphoblastic Leukemia





Deploy CICERO on St Jude Cloud for Rapid RNA-seq Analysis in Clinic



- Meet timeline of 15 days
- Cost \$10 per run
- Set up a time-out of 15 hours per sample
- In production since 2017 to support the Total XVII clinical trial
- Multiple iterations of performance improvement

Rapid RNA-seq Fail/Timed Out for a Subset of Samples

Sample	Target	Project	RRS_Fail_reason	Public RRS (BAM)	Public RRS (FastQ)	Reads
SJEPD031786_D2	TRANSCRIPTOME	Clinical/2020	Timeout	Timeout	Timeout	148,185,112
SJAML031434_D1	TRANSCRIPTOME	Clinical/2019	None	Success	Success	150,270,874

Data analyzed on-prem were able to complete successfully

This does not appear to be totally related to # of reads

We have performed a down-sampling “experiment” and found that the majority of the samples remain to be timed-out

Optimization Implemented in 2020

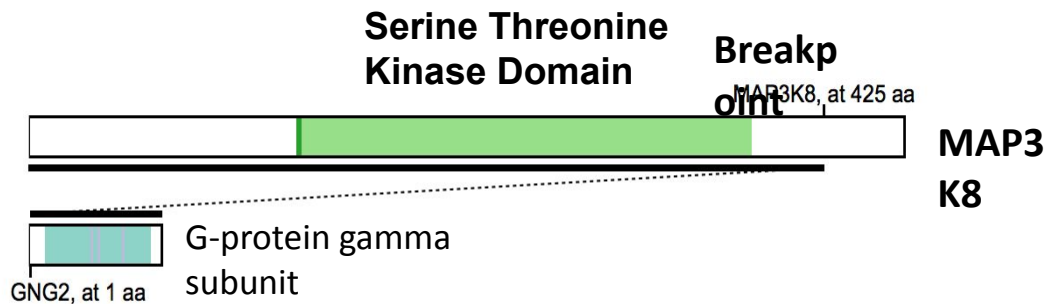
- Exclusion sites
 - Centromere / Telomere
 - Problematic sites (Seasult)
- Increase minimum soft clip read support when number of sites exceeds a threshold
- Read BLAT results once per query
- Removed unused or unnecessary calls (particularly subshells)
- Other updates
 - Updates to sv_inframe.pl
 - Update soft clip clustering distance
 - Label complex regions before recurrence check

Performance with Optimization

- 170 benchmark samples
 - Before updates, average runtime ~6 hr (\$9.504 at \$1.5840 / hr)
 - After updates, average runtime ~2.5 hr (\$3.96 at \$1.5840 / hr)
- 30 time-out samples from clinical service
 - 26 now complete under 15 hours
 - 29 complete under 20 hours

Therapy Change Based on Clinical Sequencing

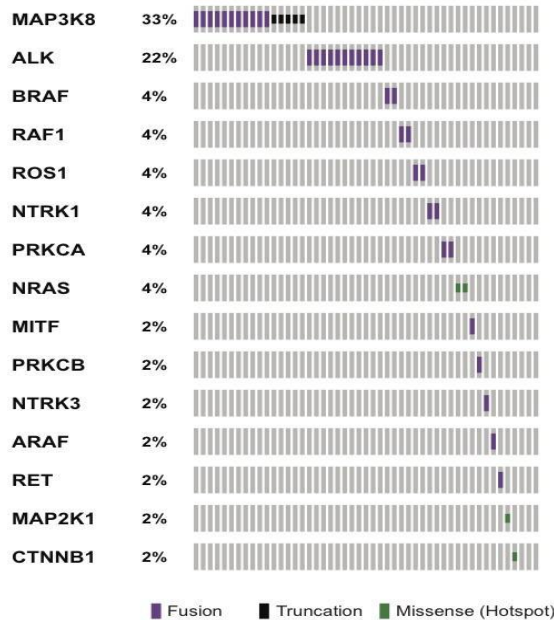
- JAK inhibitor for an ALL of IGH/EPOR known to activate JAK-STAT pathway
- MEK inhibitor for a spitzoid melanoma with *MAP3K8-GNG2* fusion predicted to activate MAP kinase signaling independent of BRAF



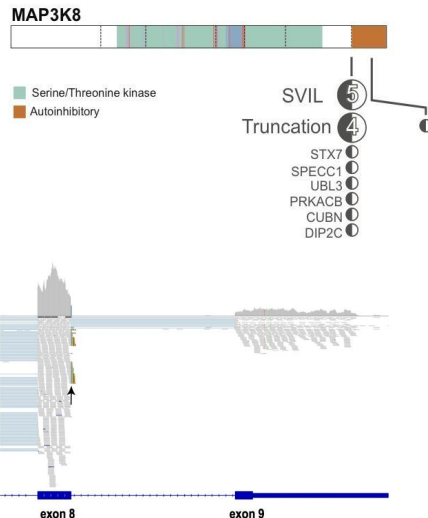
- BMT for an ALL with CREBBP mutation, predicted to have poor outcome
- Immunotherapy for two high grade gliomas with a hypermutator phenotype

Recurrent Screening by RNA-seq of 49 FFPE Spitzoid Melanoma

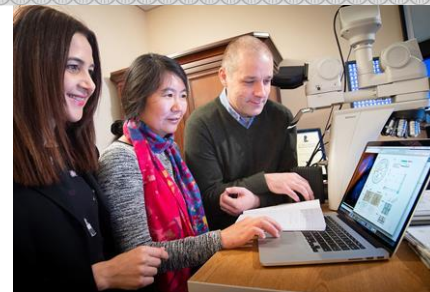
MAP3K8 has the highest mutation prevalence (33%)



Truncations/fusions cause loss of exon 9

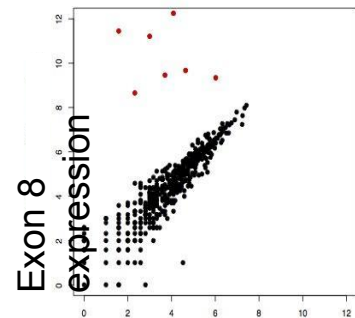


Ongoing collaboration to test new compounds targeting MAP3K8



Newman et al, Nature Medicine 2019

472 TCGA melanoma

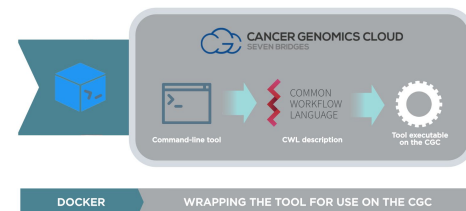


Exon 9 expression

Deploy CICERO to NCI Cancer Genomics Cloud

Wrapping CICERO with CWL

- Cancer Genomics Cloud (CGC) requires Common Workflow Language (CWL) for software implementations
- The native CICERO implementation is a 5 step workflow with a complicated working directory structure
- In CGC, the CWL workflow is one step and runs on a single, multi-core node
- GNU parallel provides on-node parallelization across available cores



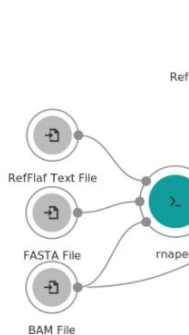
Running CICERO in NCI Cancer Genomics Cloud

1. Select RNApeg + CICERO workflow

RNApeg + CICERO
Created by an thrasher on Apr. 19, 2021 11:45 - Last edited
Revision note: "Point to GH docker image"

Description

No description.



20 TCGA GBM samples
Average runtime: 1395 minutes
Median runtime: 1068 minutes
Average cost: \$7.16
Median cost: \$4.30

Inputs

ID	Label
in_bam	BAM File
in_fasta	FASTA File
in_refflat	RefFlat Text File
in_ref	Reference Directory

2. Select inputs

DRAFT RNApeg + CICERO run - 05-03-21 17:06:42

Last update by an thrasher on May 3, 2021 13:06
App: RNApeg + CICERO - Revision: 5

Task Inputs Execution Settings

Inputs

Batching Off

BAM File *
No files selected

FASTA File *
No files selected

This field is required and cannot be empty.

App Settings

cicero (f/cicero)

Genome *

This field is required and cannot be empty.

ncores

05-03-21 13:38:29

Progress: Running.

3. Wait for workflow to complete

Search task names Status: All

Task Name

- RNApeg + CICERO run - 05-03-21 13:38:29: file: 212ae104-ecfc4888-a4e0-2e9be74ac1f1_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: 48271f49-4bbb-4b63-abcfc6c981a6383d1_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: f72f58ea-8d37-4beb-b334-20206cda1d4_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: 593b8463-f60e-44b9-ad21-8952608ac325_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: a288ce0e-2b96-4788-9d2b-92e83e6a0d40_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: 492ef759-a33e-4b20-b0f1-18a551e11b1c_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: 33e402b4-9a88-45d7-a184-b109774e0003_gdc_realn_rehead.bam
- RNApeg + CICERO run - 05-03-21 13:38:29: file: fc534c42-60c7-495c-926b-39768a06d4c3_gdc_realn_rehead.bam

Analyze fusions with FusionEditor

7.0GiB BAM file produces a 180 candidate fusions (144kb in size). This output can then be visualized in St. Jude Cloud with FusionEditor for manual curation

Welcome to the ProteinPaint/FusionEditor!

Select the reference genome and upload your CICERO output file to begin. Or, [try out an example](#).

141 genes | 1 sample | Gene expression | Parameter cutoff | Legend | Export data | Help

9637636a-1fb7-401a-85ed-b619512b56c3_gdc_realn_rehead

◦ HQ **In-frame** 6

chr7	(11)	EGFR	SEPTIN14	chr7	HC IN Fusion	coding coding	5896	No recurrence
chr4		FGFR3	TACC3	chr4	HC IN Fusion	coding coding	604	No recurrence
chr5		PIK3R1	PIK3R1	chr5	HC IN ITD	coding coding	114	No recurrence
chr17		PSME3	TTYH1	chr19	HC IN Fusion	coding coding	48	No recurrence
chr19		C19orf48	ATP6V1E1	chr22	HC IN Fusion	coding coding	35	No recurrence
chr1		ARID1A	NCAN	chr19	HC IN Fusion	coding coding	3	No recurrence

- HQ **truncation** 7
- LQ **truncation** 102
- LQ **others** 12
- Read-through **In-frame** 3
- Read-through **truncation** 29
- Read-through **others** 11



Ongoing Work

- Improving accuracy by applying CICERO to SRA GTeX samples to profile patterns that resemble false positives
- Running this locally as there is already a local copy of GTeX on SRM
- Can this be done using SRA Cloud?

Acknowledgement

Andrew Thrasher

Liqing Tian

Clay McLeod

Mike Edmonson

Cloud-Based Whole Genome Sequencing Analysis Workflow

Xihong Lin (Harvard)

Cloud-Based Whole Genome Sequencing Analysis Workflow

Xihong Lin

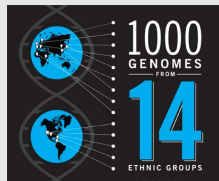
Department of Biostatistics and Department of Statistics

Harvard University

NHGRI Genome Sequencing Program

NHLBI TOPMed

Need: Develop Cloud Platforms for Scalable WGS Analysis



1000 Genomes
N=1000

2008



GSP(NHGRI)
N=360,000

2016

NATIONAL
CANCER
INSTITUTE

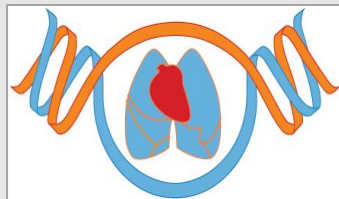
IHCS(NCI)
N=200,000

2019

Large Scale Whole Genome/Exome Sequencing Timeline

2015

TOPMed
(NHLBI)
N=150,000

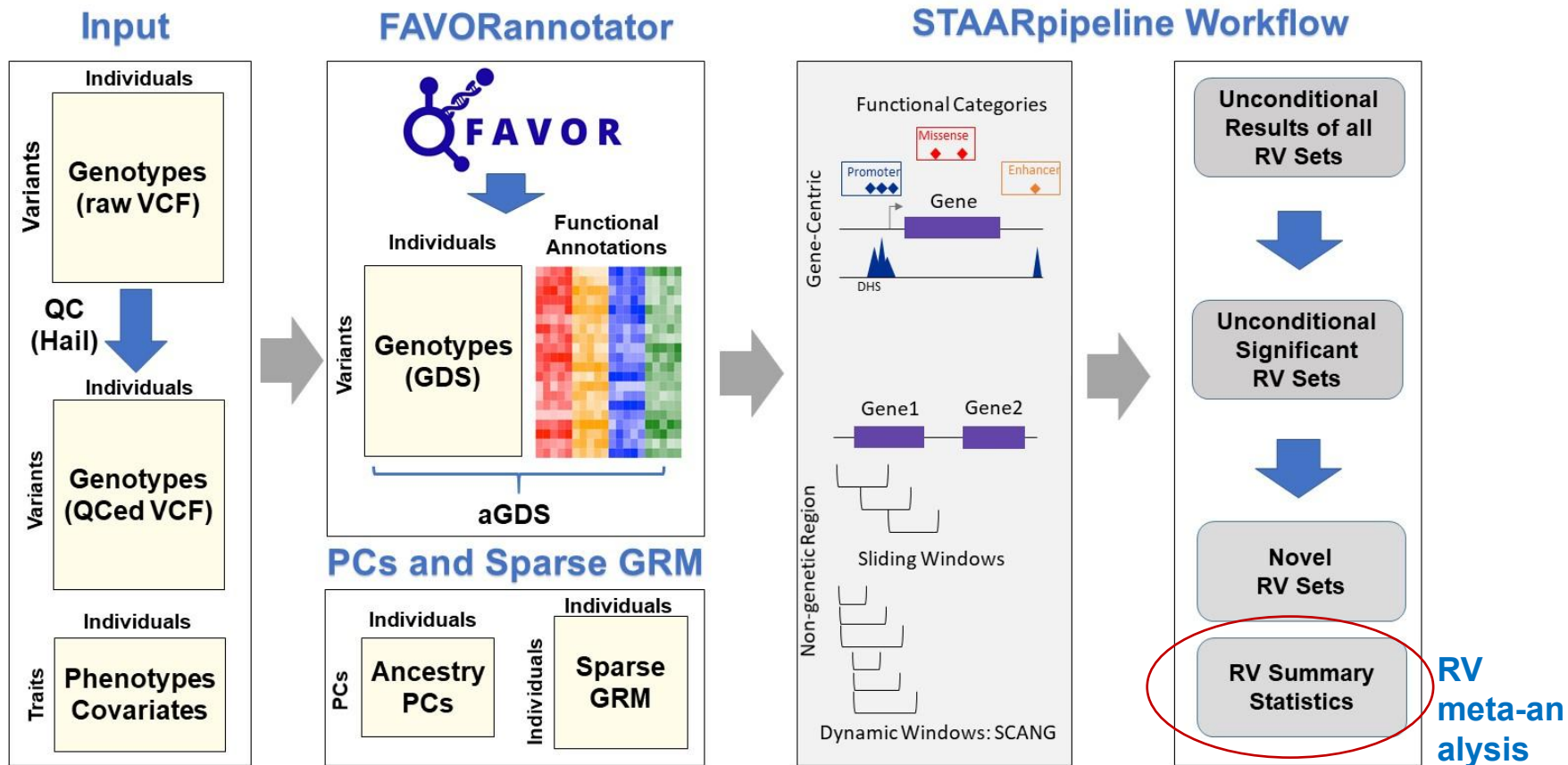


2018

Biobanks
(N=tens of
millions in a
few years)



Overview of WGS Analysis Pipeline (Functional Annotator + Rare Variant Analysis Workflow)





Functional Annotation of Variants Online Resources (FAVOR)

Online Portal (Web UI)

Offline or
Online
Annotator

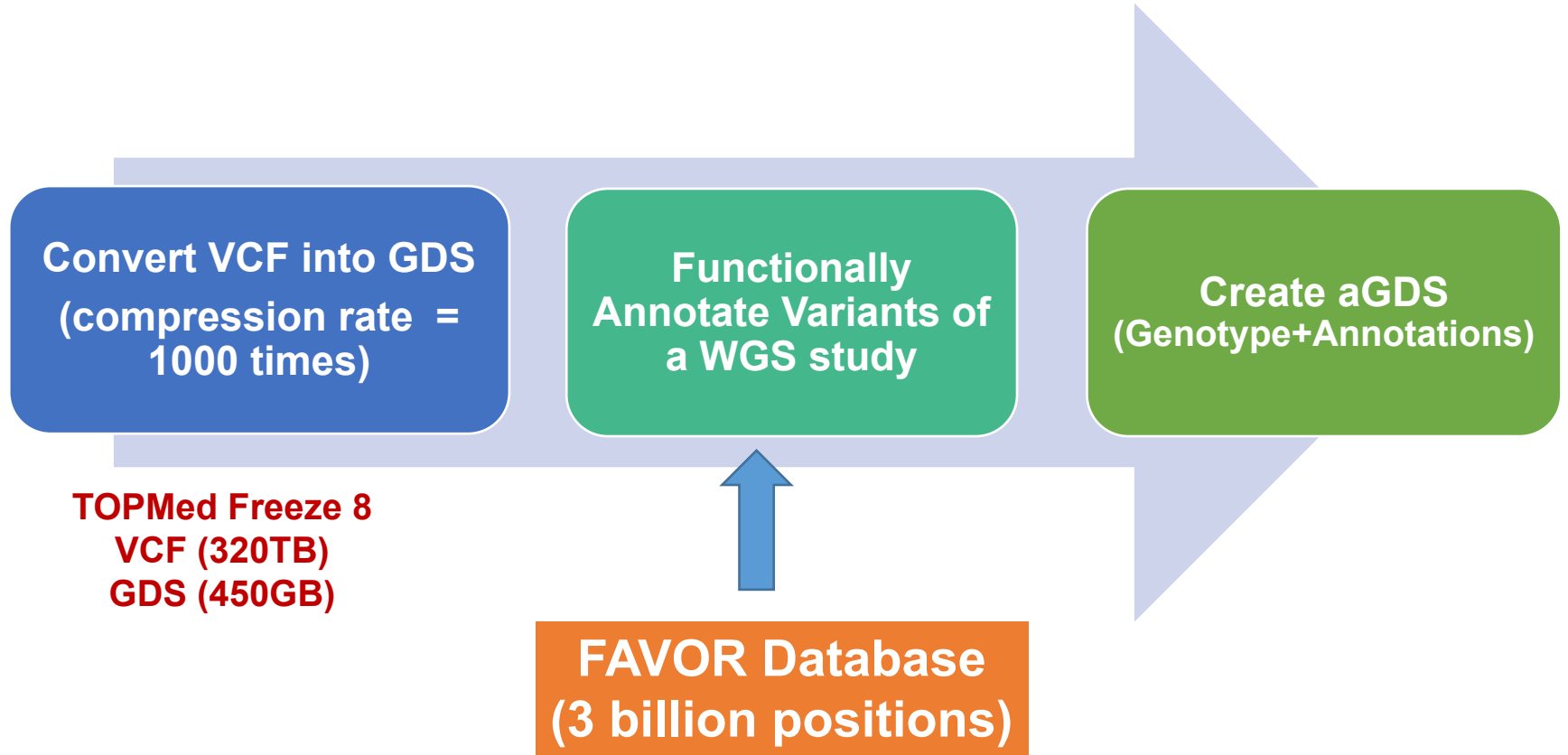
Single
variant-based
Query

Region- /
Gene-based
Query

Batch Annotation
for small variant
sets

Create aGDS of
variants for a WGS
study

FAVOR Annotator Workflow



How FAVOR annotator works (scripts)

- Backend database: functional annotation of 9 billion SNVs
- Install the FAVOR V2 SQL database in local computer or cloud platforms
- Run FAVORannotator scripts or use BigQuery:

Rscript `FAVORannotator.R` `Input/vcf.gz` `Output/annotated.gds`

WGS Association Analysis Workflow

Input

Common Variants
Single SNV Analysis

Rare Variants
SNV-Set Analysis

- Phenotype
- Covariates/PCs
- aGDS

Gene-Centric
Analysis

Genetic
Region
Analysis

Coding
Masks

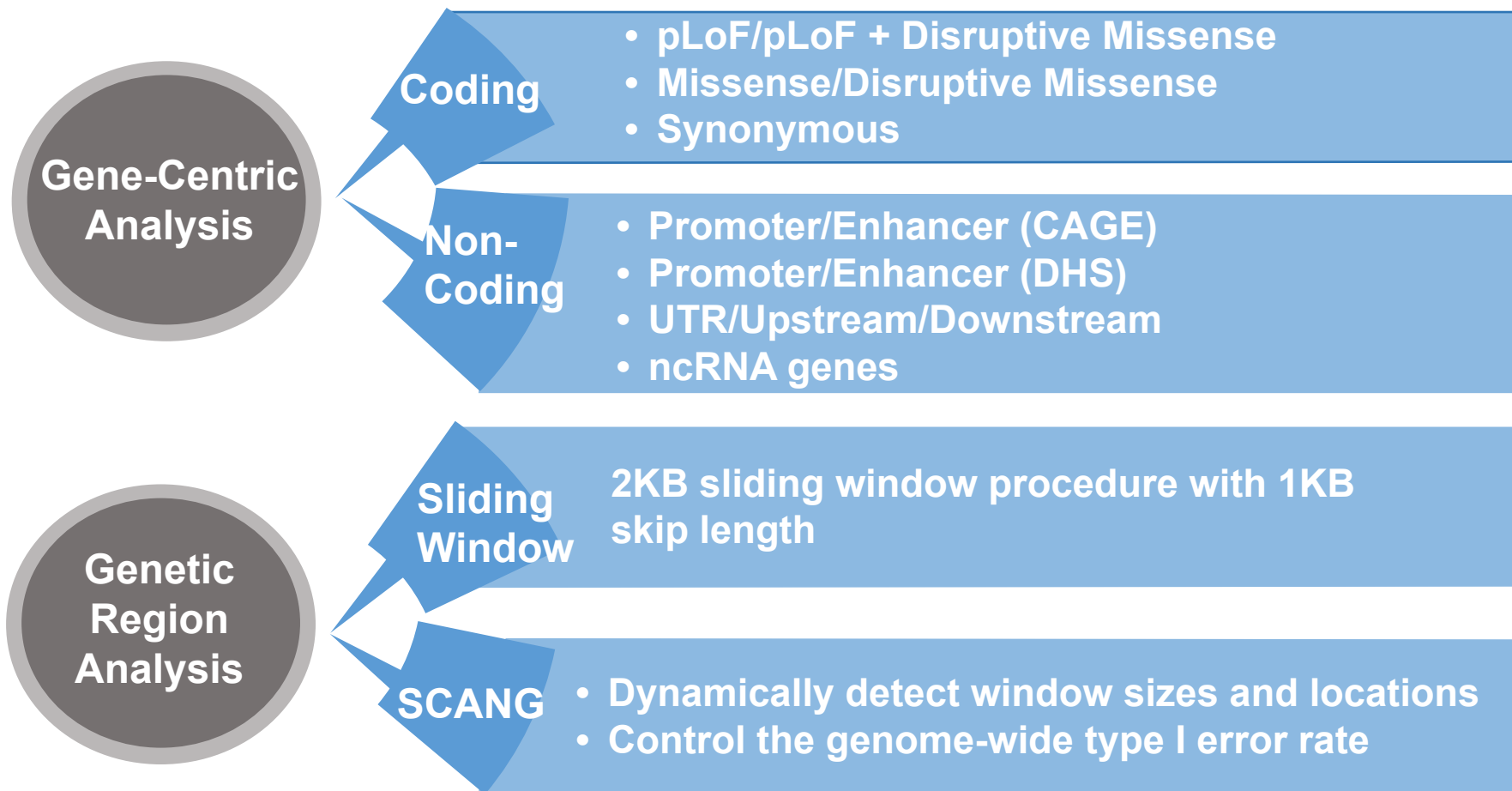
Noncoding
Masks

Sliding
Window
Analysis

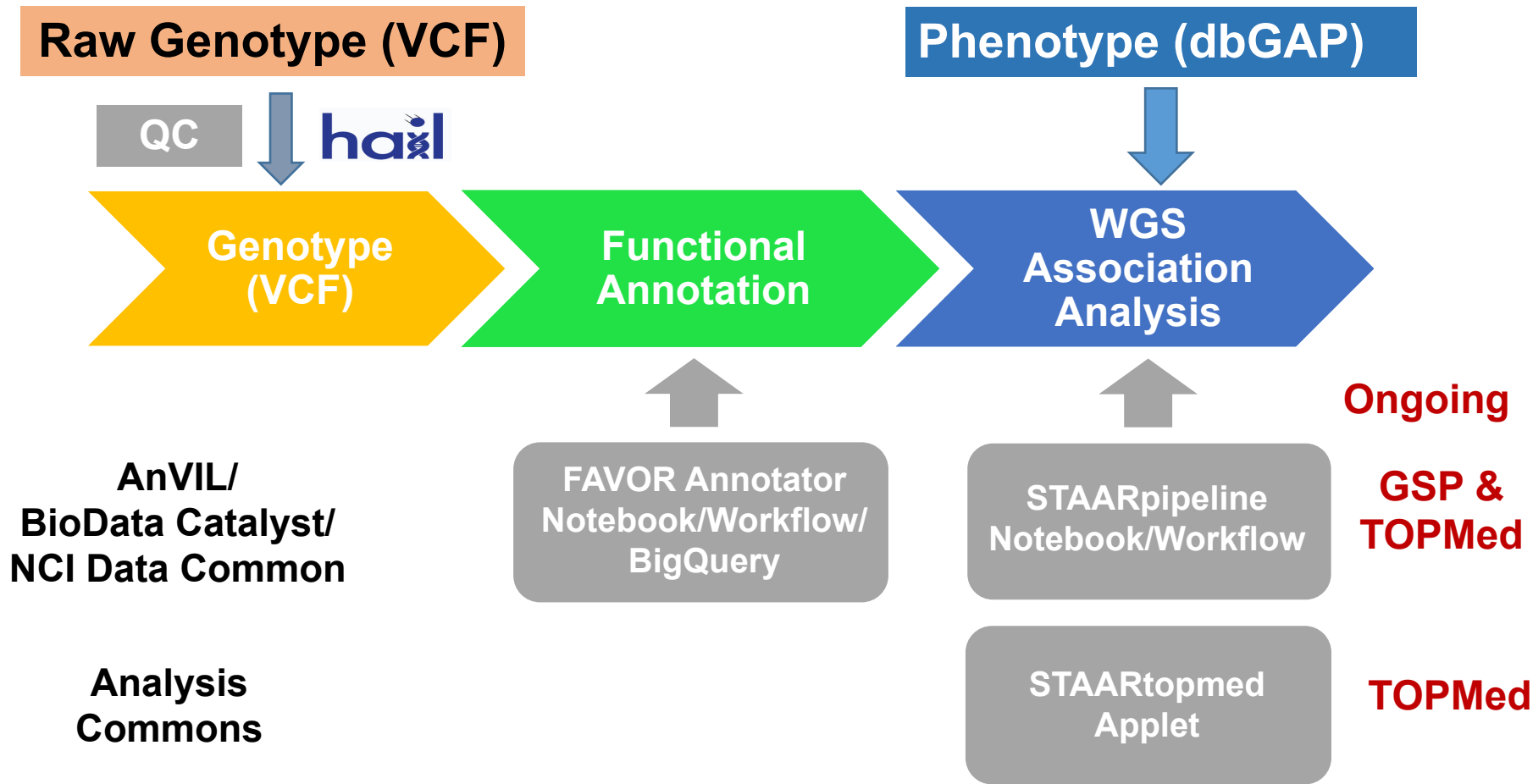
Dynamic
Window
(SCANG)

Unconditional and conditional common and rare variant analysis

STAARPipeline Workflow for RV Analysis (MAF < 1%)



Implementation of Annotator and Analysis Workflow in Terra



STAAR App in Analysis Commons and BioData Catalyst (Ongoing)

STAARtopmed Applet in Analysis Commons

STAAR in BioData Catalyst

RUN "STAAR PROCEDURE FOR ANALYZING TOPMED WGS DATA" AS ANALYSIS

View job progress in the Monitor tab.

STAAR Procedure for Analyzing TOPMed WGS Data

1 app
unconfigured

Workflow Actions ▾

▶ Run as Analysis... ⚙

Inputs

App

Outputs

*.csv | A.csv file saving the phenot...

STAAR Procedure for An...
configure params

results [array]

*RData *Rdata *Rda | An R object saving the relate...

*RData *Rdata *Rda | An R object saving the fitted...

*gds | Genotype and functional an...

Close

BioData CATALYST
Powered by Terra

WORKSPACES
staar_rare_variant_pipeline

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

← Back to list
staar_rare_variant_pipeline

Version: master

Source: github.com/ahellegyrod/staar_rare_variant_pipeline/master

Synopsis:
No documentation provided

Run workflow with inputs defined by file paths
 Run workflow(s) with inputs defined by data table

Use call caching Delete intermediate outputs Use reference disk

SCRIPT INPUS OUTPUTS RUN ANALYSIS

SAVE CANCEL

Hide optional inputs Download json | Drag or click to upload json

Task name	Variable	Type	Attribute
STAAR_analysis	geno_files	Array[File]	Required
STAAR_analysis	results_file	String	Required
run_analysis	agds_annot_channels	String	Optional
run_analysis	agds_file	String	Optional
run_analysis_annotfree	annot_file	File	Optional
STAAR_analysis	agds_annot_channels	String	Optional
STAAR_analysis	agds_file	String	Optional
STAAR_analysis	agg_file	File	Optional
STAAR_analysis	annot_files	Array[File]	Optional
STAAR_analysis	cond_file	File	Optional

Benchmarking: STAAR Analysis of TOPMed Freeze 5 Fasting Glucose and Insulin Traits (n=23-26K) in BioData Catalyst

Task	Time
1- Null model	<1 hr
2- Gene centric (100 6GB cores)	1 hr
2- Genetic region (100 6GB cores)	13-14 hr
3- Summarizing	<1 hr

Benchmarking: Total Cost of WGS of n=62,000 Individuals Using STAARtopmed Workflow Applet in Analysis Commons

Method	WG Cost Est.*	WG Computation Time (hr)**
Null Model	\$0.17	1
Individual	\$111.54	5
Coding	\$48.62	2
Noncoding	\$258.61	5
ncRNA	\$32.23	1.5
2kb Sliding Window	\$570.13	9
Total	\$1021.30	23.5

* The cost and time is based on analyzing TOPMed F8 LDL trait (n = 62,000)

** The Applet run analysis separately for each chromosome. The time is benchmarked by chromosome 1 (longest)

Challenges and Opportunities

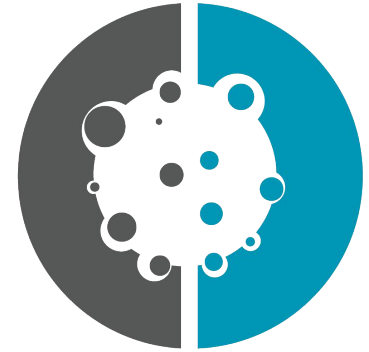
- **Data:** Tedious dbGAP approval, letters of collaboration, phenotype harmonization
- **Cost:** Cloud data storage & computing costs are much more than Computing Clusters
- **Analytic platforms:** Need for supporting developing analytic tool & resource in cloud
- **Visualization** of WGS RV analysis results ([ongoing](#))
- **MetaSTAAR ([ongoing](#)):** a cloud based efficient & scalable workflow for rare variant meta-analysis
 - RVAS summary statistics (score statistics and covariance).
 - Standards and portal for rare variant summary statistics catalog
 - Collaboration with Type 2 Diabetes Knowledge Portal
- **Phe-STAAR ([ongoing](#)):** A cloud based efficient & scalable workflow for phenome-wide rare variant analysis for biobanks and metabolomics.
 - Portal for biobank RVAS summary statistics

NCI CRDC Center for Cancer Data Harmonization Efforts

**Melissa Haendel (U of Colorado)
Sam Volchenboum (UChicago)**

Bringing the CRDC data into harmony

Melissa Haendel
Sam Volchenboum



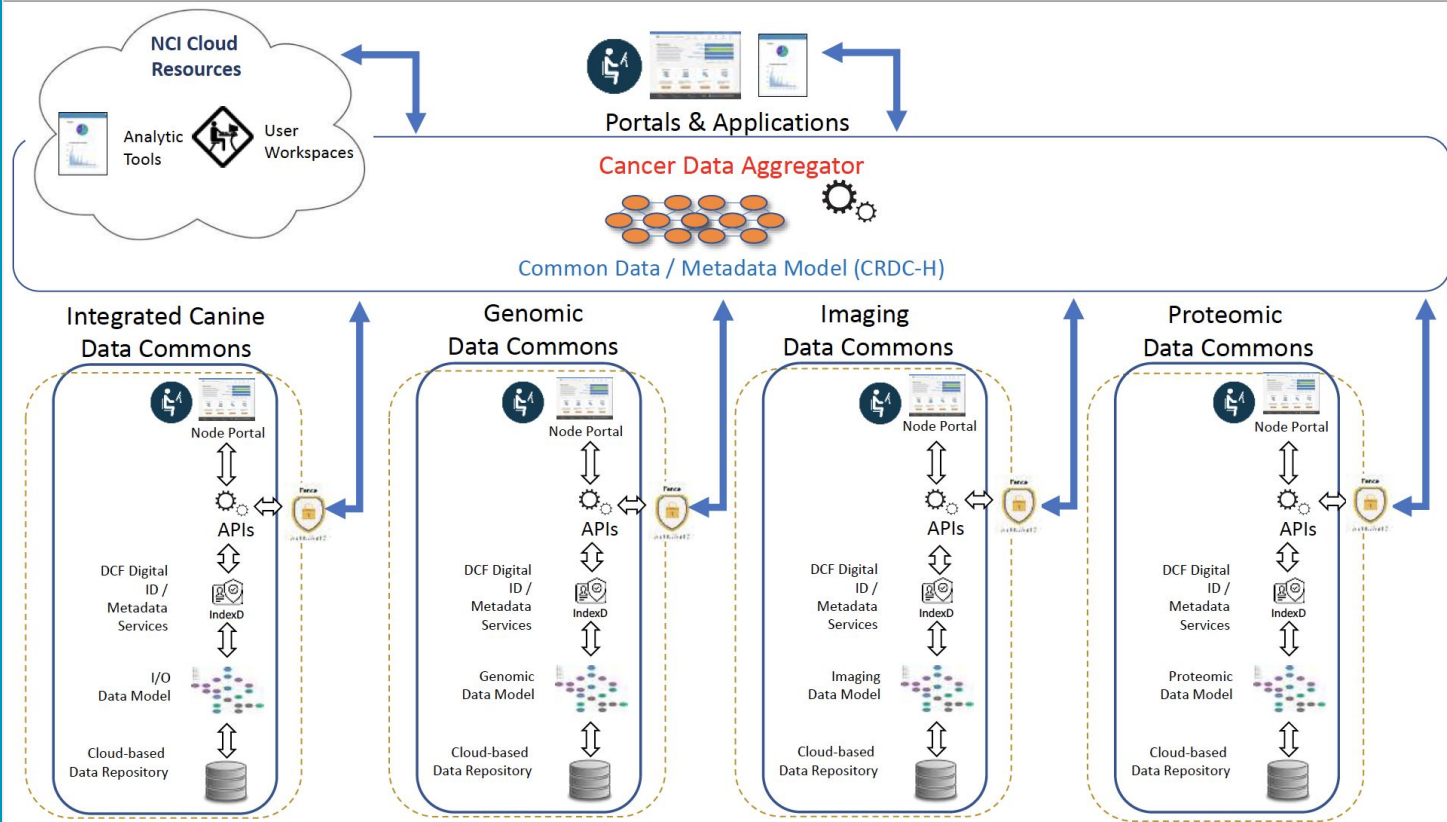
CENTER *for*
CANCER DATA
HARMONIZATION

ccd.h.cancer.gov

NCPI workshop
May 4, 2021

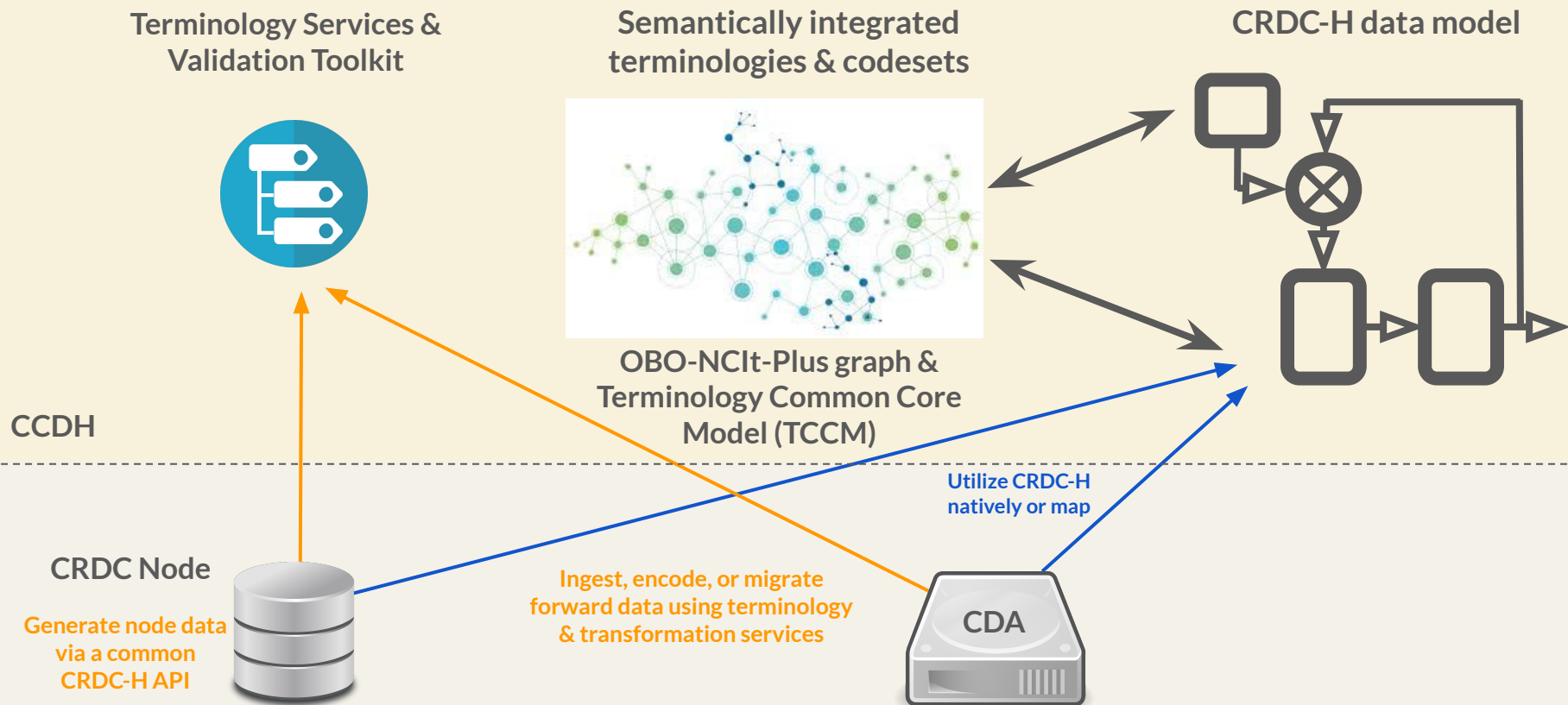
These slides: bit.ly/ccdh-ncpi-2021

CCDH: Building a common data model and services to help harmonize data across the CRDC

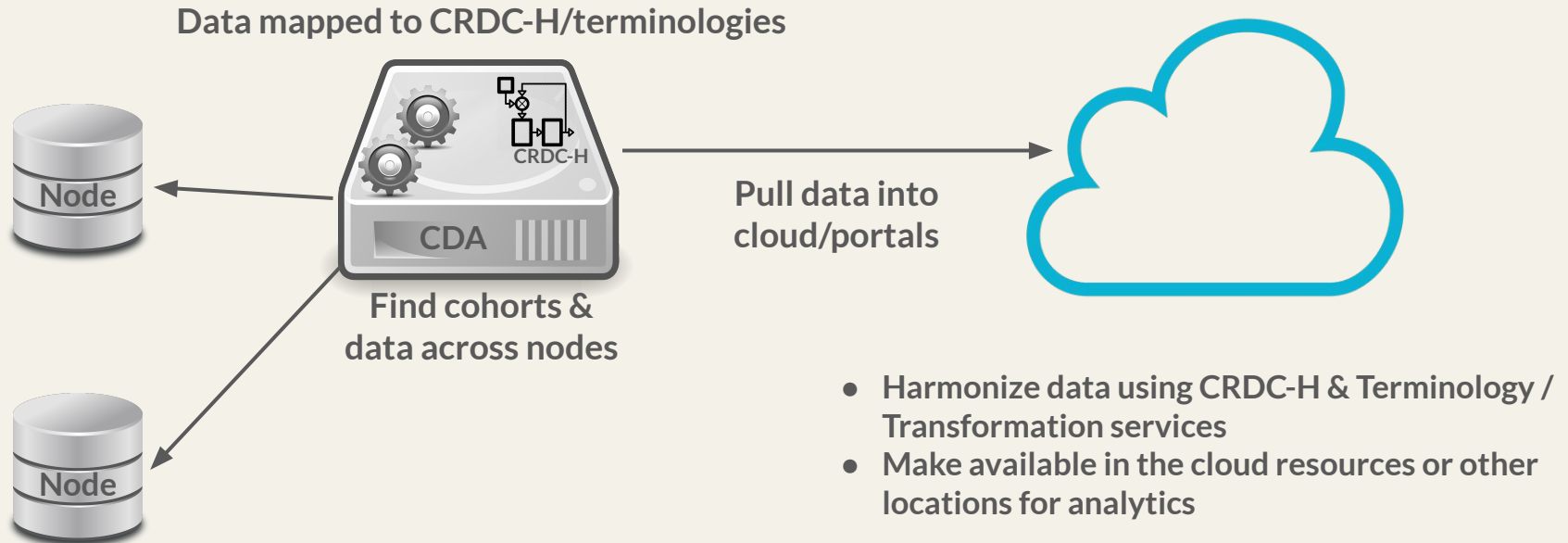


The CRDC resources do not use the same data model or terminological content, making query and analytics across them challenging

Enabling search across the CRDC ecosystem

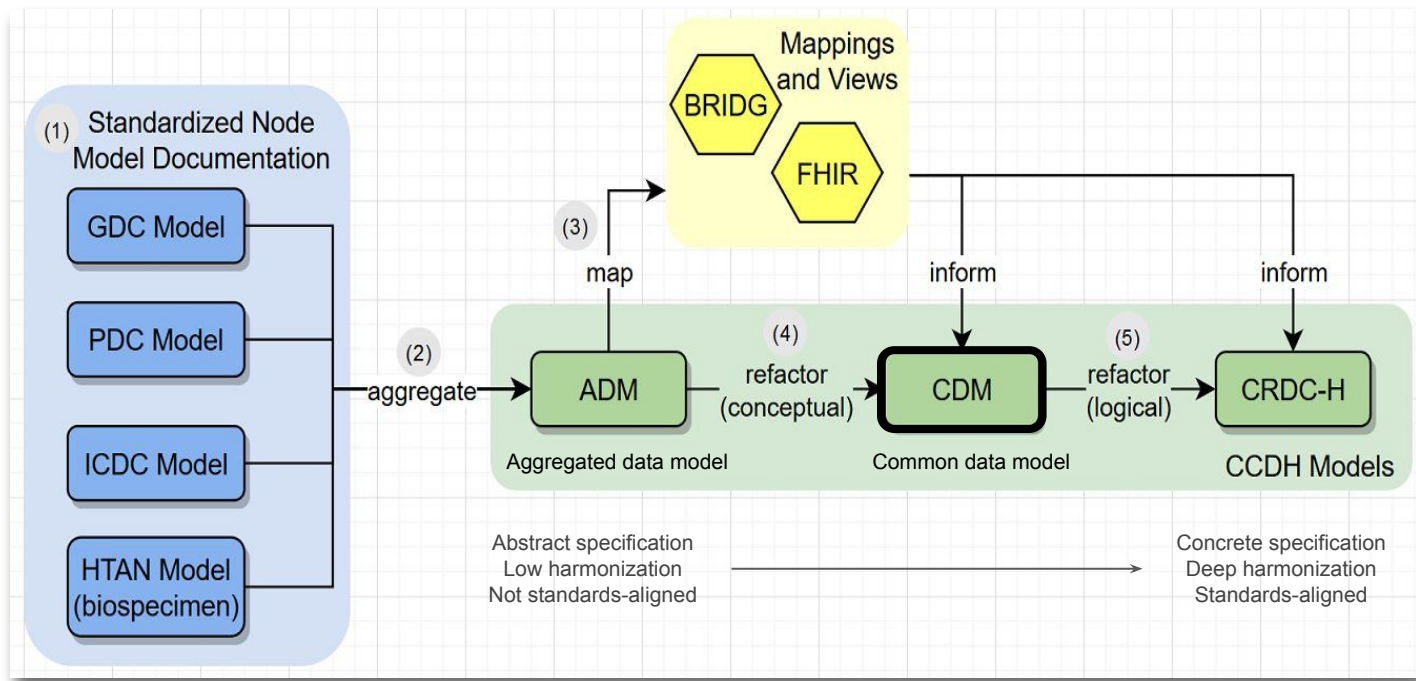


Getting the data, harmonizing it, using it



Introducing the CRDC harmonized data model (CRDC-H)

An iterative process where source model content is evaluated, aggregated, mapped, and refactored into a standards-aligned and harmonized data model, the **CRDC-H**

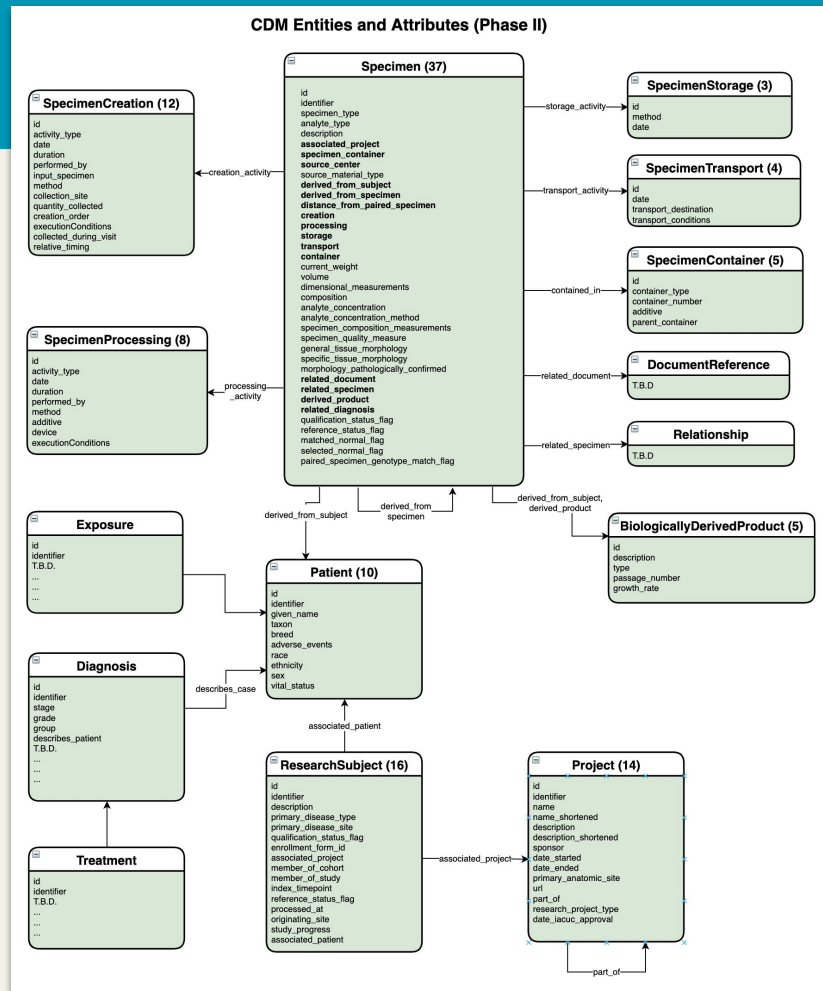


CRDC-H Scope

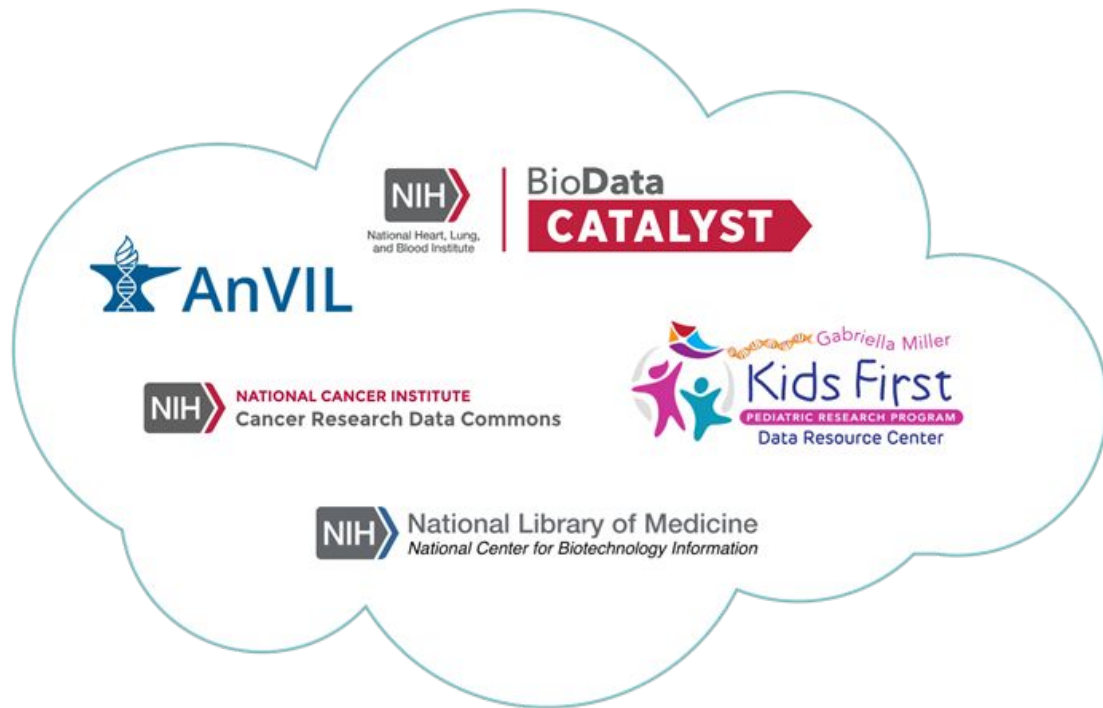
- End of May release will include *Biospecimen and Administrative subdomain entities*, along with *select Clinical subdomain entities*
 - Demographics
 - Diagnosis
 - Treatment
 - Exposure
- Terminology bindings will be included

Auto-generated CRDC-H at:

<https://cancerdhc.github.io/ccdhmodel/>



How to harmonize the DATA across the NCPI?



To date, NCPI has focused on system interoperability.

The use of common data models, terminologies, and standards for their use can enable data interoperability in support of search and multi-modal analytics.

How can we achieve this across heterogeneous resources and studies?

We need a semantics-friendly modeling language that can be realized in different instantiations

LinkML: “born interoperable” semantic data modeling framework designed for data dictionaries, data submission forms, data commons, and complex biomedical schemas

- Simple YAML as the source of truth
 - **Expressive:** but only use what you need
- Generate
 - **JSON Schema:** validation for JSON
 - **Python Dataclasses:** building Python APIs and writing ETL
 - **Java classes:** building Java APIs and writing ETL
 - **GraphQL:** building APIs on top of data stores
 - **SQL DDL:** (in progress)
 - **JSON-LD context:** RDF to JSON serialization
 - **RDF Turtle:** Semantic web, RDF graphs
 - **OWL:** reasoning, ontology generation
 - **Shape Expressions (ShEx):** validation of RDF graphs

A sample LinkML Schema

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
license: https://creativecommons.org/publicdomain/zero/1.0/
version: 0.0.1

prefixes:
  linkml: https://w3id.org/linkml/
  sdo: https://schema.org/
  ex: https://example.org/linkml/hello-world/

default_prefix: ex
default_curi_maps:
  - semweb_context

imports:
  - linkml:types

classes:
  Person:
    description: Minimal information about a person
    class_uri: sdo:Person
    attributes:
      id:
        identifier: true
        slot_uri: sdo:taxID
      first_name:
        required: true
        slot_uri: sdo:givenName
        multivalued: true
      last_name:
        required: true
        slot_uri: sdo:familyName
    knows:
      range: Person
      multivalued: true
      slot_uri: foaf:knows
```

Metadata

Namespaces

Dependencies

Actual Model

LinkML RDF is hidden in plain sight

```
id: https://example.org/linkml/hello-world  
title: Really basic LinkML model  
name: hello-world  
license: https://creativecommons.org/publicdomain/zero/1.0/  
version: 0.0.1
```

Metadata

```
prefixes:  
  linkml: https://w3id.org/linkml/  
  sdo: https://schema.org/  
  ex: https://example.org/linkml/hello-world/
```

Namespaces

```
default_prefix: ex  
default_curi_maps:  
  - semweb_context
```

```
imports:  
  - linkml:types
```

Dependencies

```
classes:  
  Person:  
    description: Minimal information about a person  
    class_uri: sdo:Person  
    attributes:  
      id:  
        identifier: true  
        slot_uri: sdo:taxID  
      first_name:  
        required: true  
        slot_uri: sdo:givenName  
        multivalued: true  
      last_name:  
        required: true  
        slot_uri: sdo:familyName  
      knows:  
        range: Person  
        multivalued: true  
        slot_uri: foaf:knows
```

Actual Model

Terminology Bindings within LinkML

```
enums:  
  lens_color:  
    description: Harmonized stoplight lens color  
    code_set: ORS:pat0_colors  
    pv_formula: URI  
  
  id_scheme:  
    description: Stoplight identification schemes  
    permissible_values:  
      boise:  
        description: Boise Identification scheme  
      tf:  
        description: Twin Falls Identification scheme  
      poc:  
        description: Pocatello Identification scheme
```

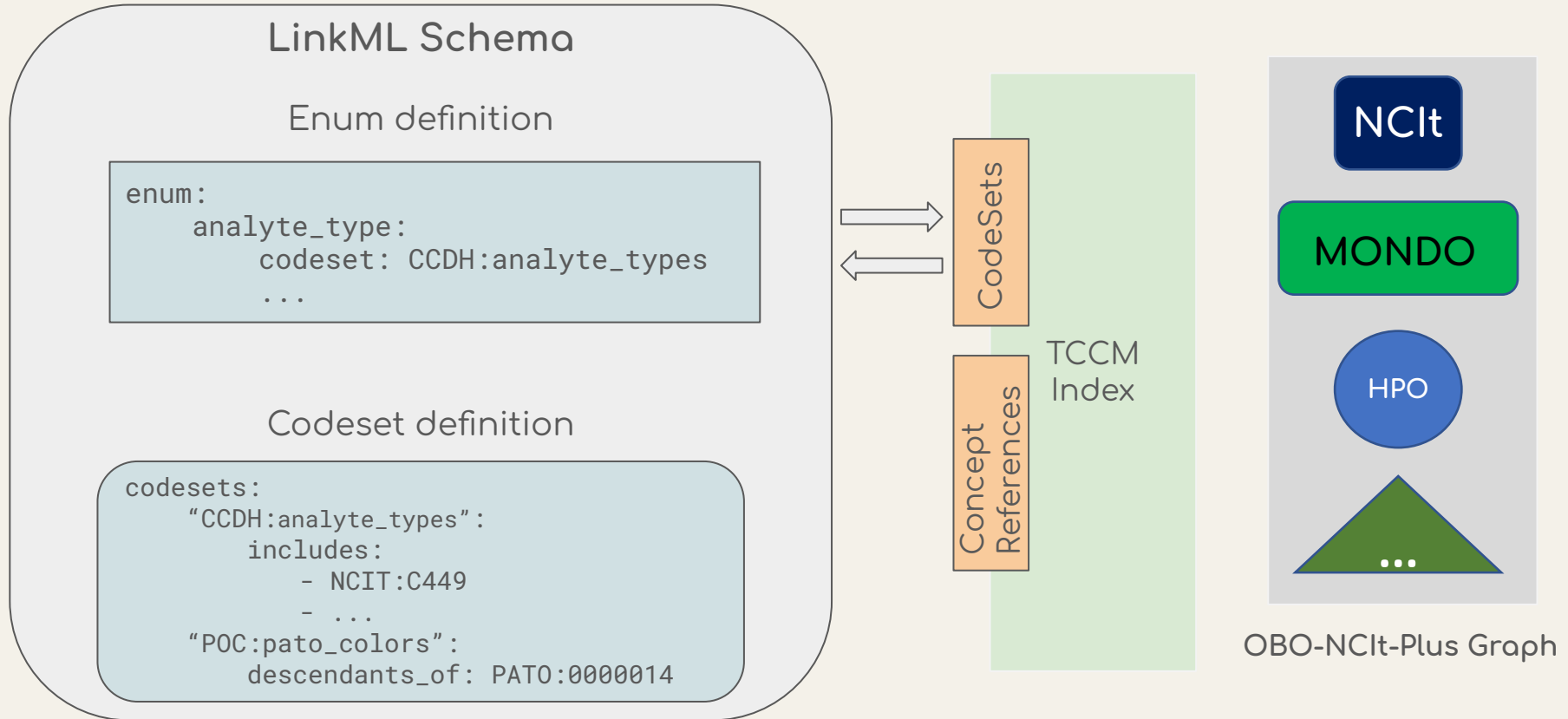
```
{ "members": [  
  {  
    "uri": "http://purl.obolibrary.org/obo/PATO_0000322"  
    "code": "PATO:0000322",  
    "defined_in": "PATO",  
    "designation": "red",  
    "definition": "A color hue with high wavelength ..."  
  },  
  {  
    "uri": "http://purl.obolibrary.org/obo/PATO_0000323"  
    "code": "PATO:0000323",  
    "defined_in": "PATO",  
    "designation": "white",  
    "definition": "An achromatic color of maximum brightness; ..."  
  },  
  ...  
]
```

Equivalent to

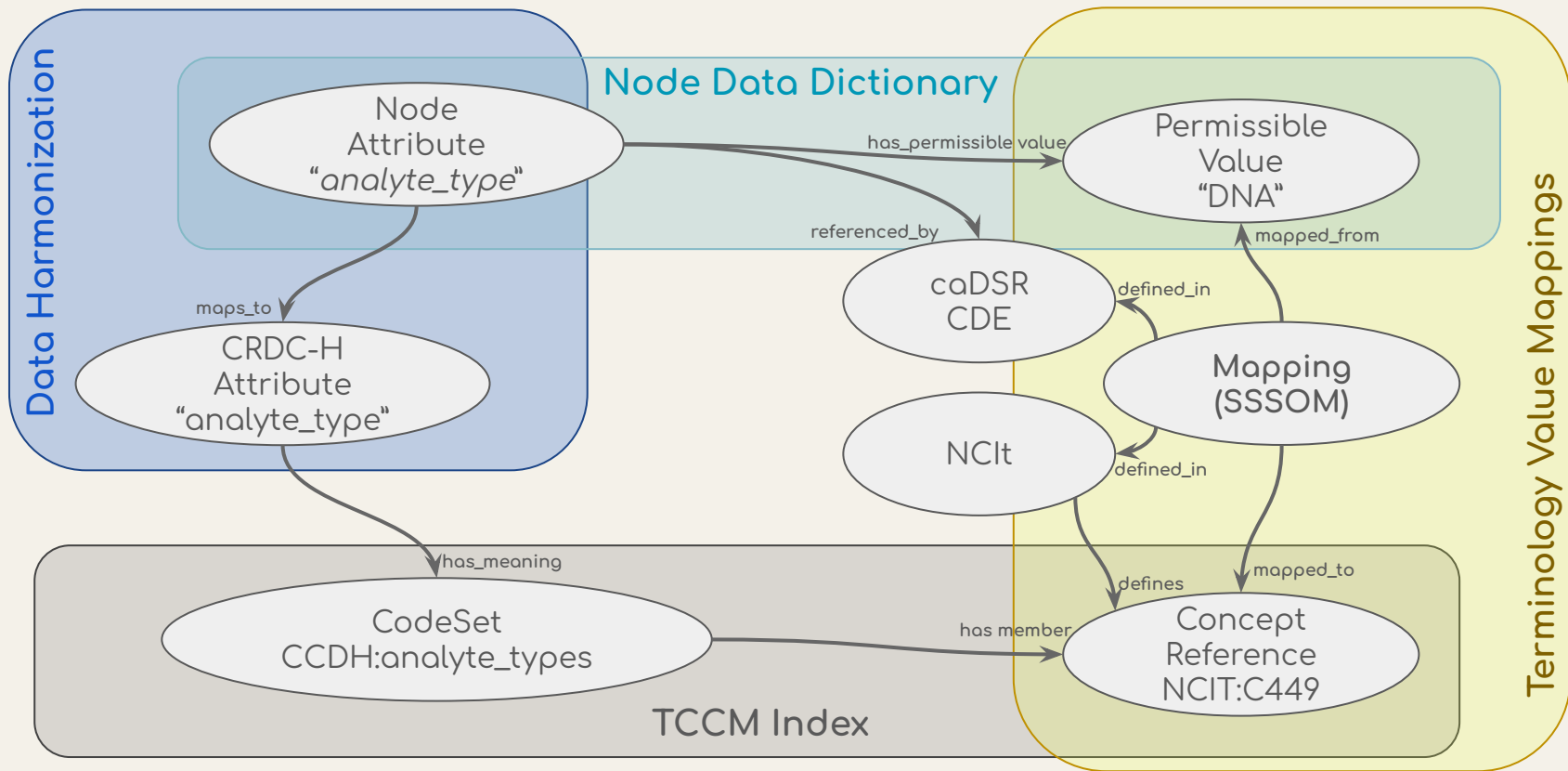
```
lens_color_v3:  
  description: Harmonized stoplight lens color URI  
  code_set: ORS:pat0_colors  
  permissible_values:  
    "http://purl.obolibrary.org/obo/PATO_0000322":  
      meaning: PATO:0000322  
    "http://purl.obolibrary.org/obo/PATO_0000323":  
      meaning: PATO:0000323
```

Codeset based
enumerations:
flexible but
semantically
defined

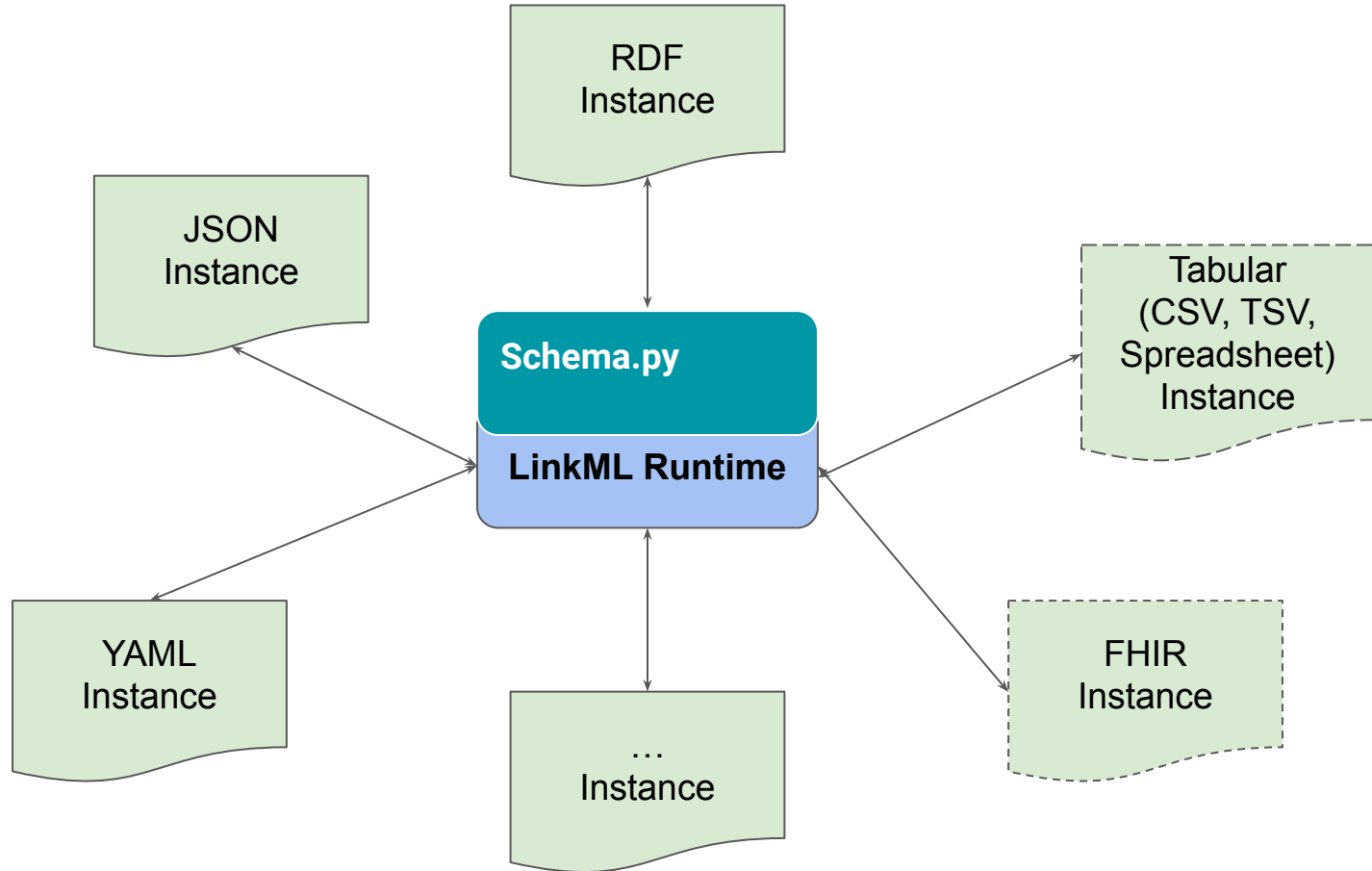
Terminology Services – TCCM (Terminology Common Core Model)



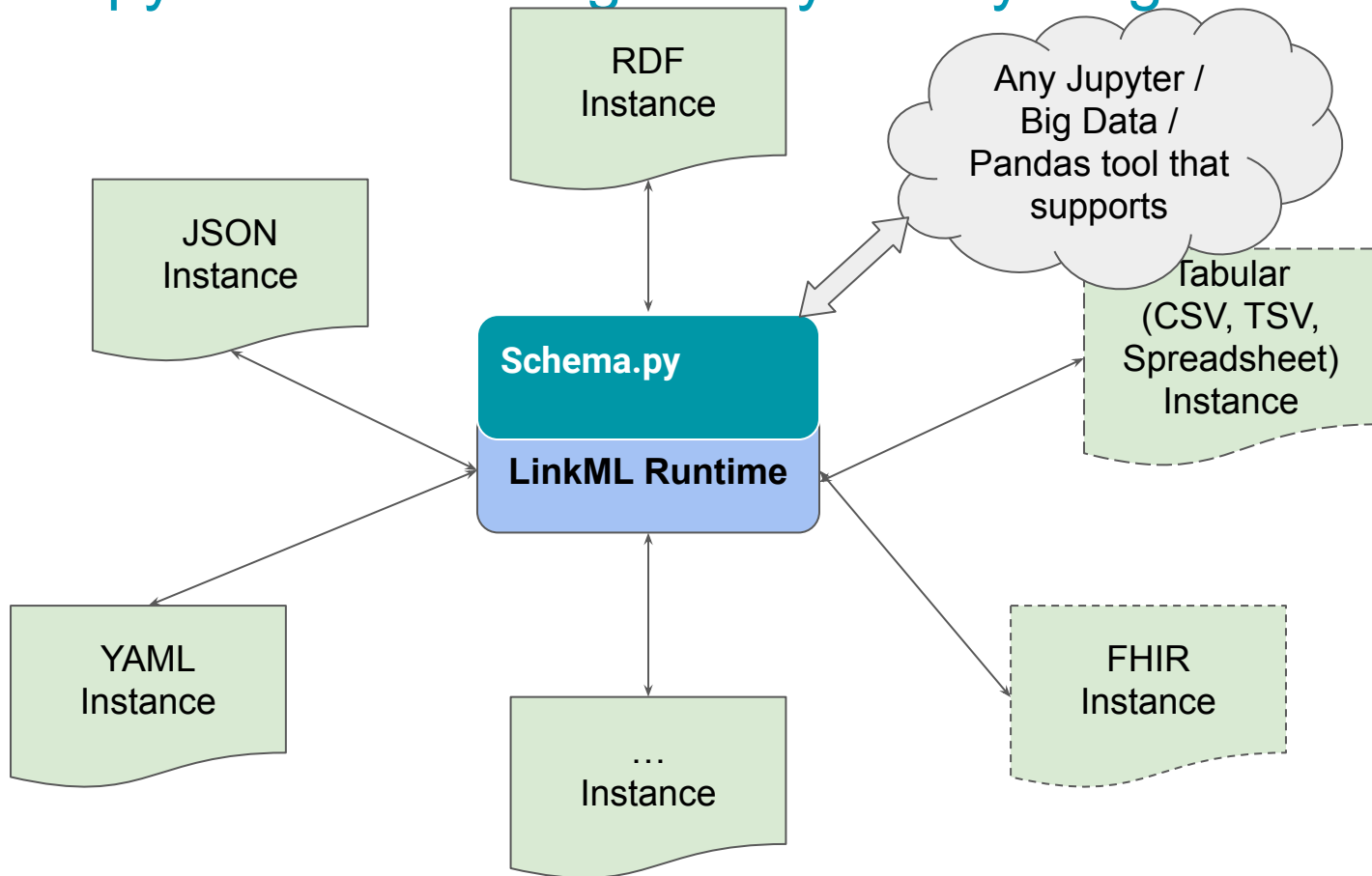
Value Mappings Graph Model



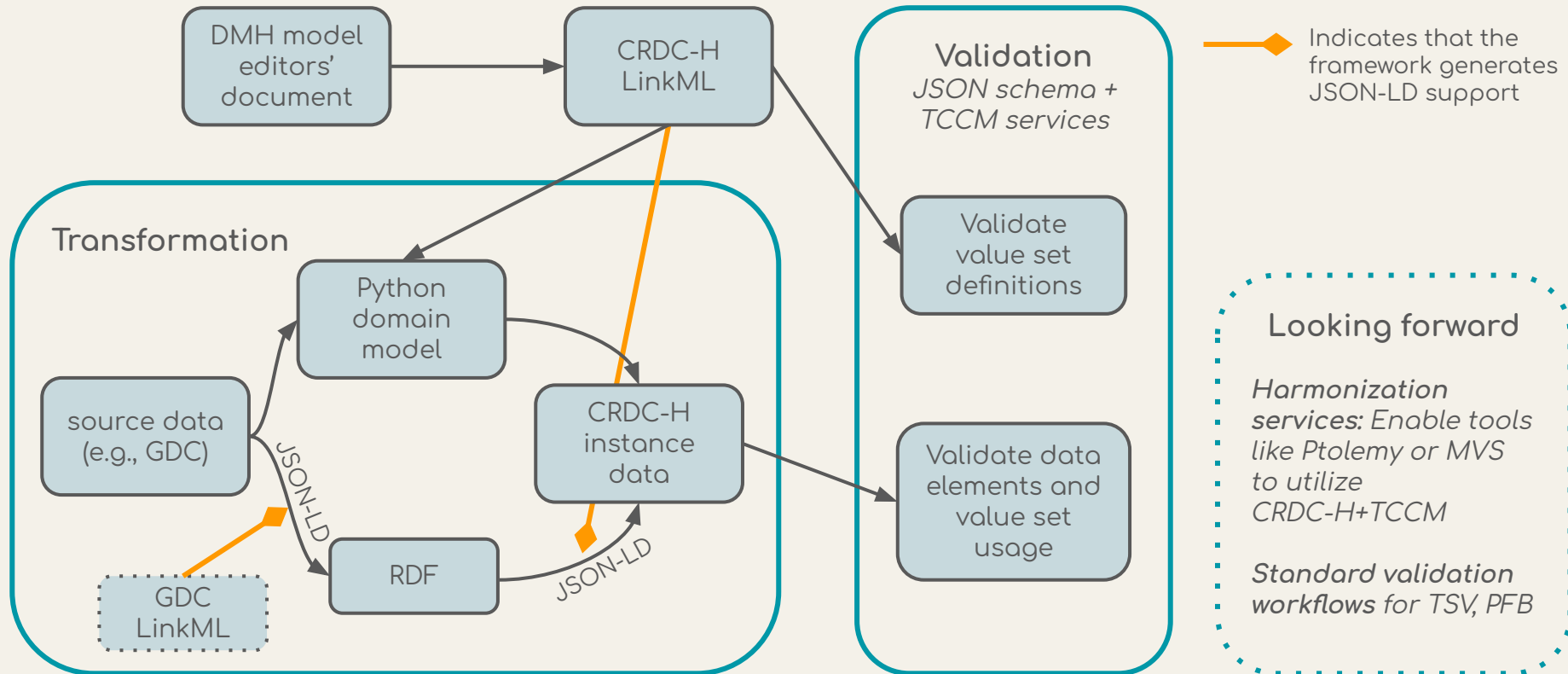
The LinkML runtime can consume and create...



Generated python can be a gateway to anything...



Transformation and validation tools



CCDH takeaways

- Creation of a common data model across data commons necessary to support cross-commons search and analytics
- Building data models using an implementation-independent language affords flexibility across platforms and contexts
- Terminology services and bindings to the model can be managed separately in a fit-for-purpose manner
- Leveraging existing resources such as caDSR for CDE value sets creates semantic interoperability
- The same data harmonization strategies and tools implemented by CCDH and CDA for CRDC could similarly be implemented within NCPI

With many thanks to the CCDH team



30 Minute Break #1

We will resume at 1:00 pm EDT

Announcements

- Fall 2021 Workshop poll: **tinyurl.com/NCPIfallpoll**
- If you have not registered, please do: **tinyurl.com/NCPIregistration**
- The NIH Office of Data Science Strategy recently announced four Notices of Special Interest for supplemental funding: **tinyurl.com/ODSSfunding**

Group Discussion on Community Interoperability Talks

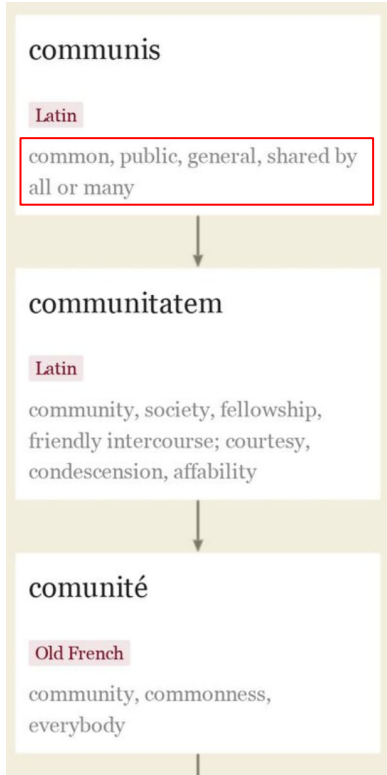
Adam Resnick (CHOP)



Community Interoperability (Discussion)

May 4, 2021

“Community”

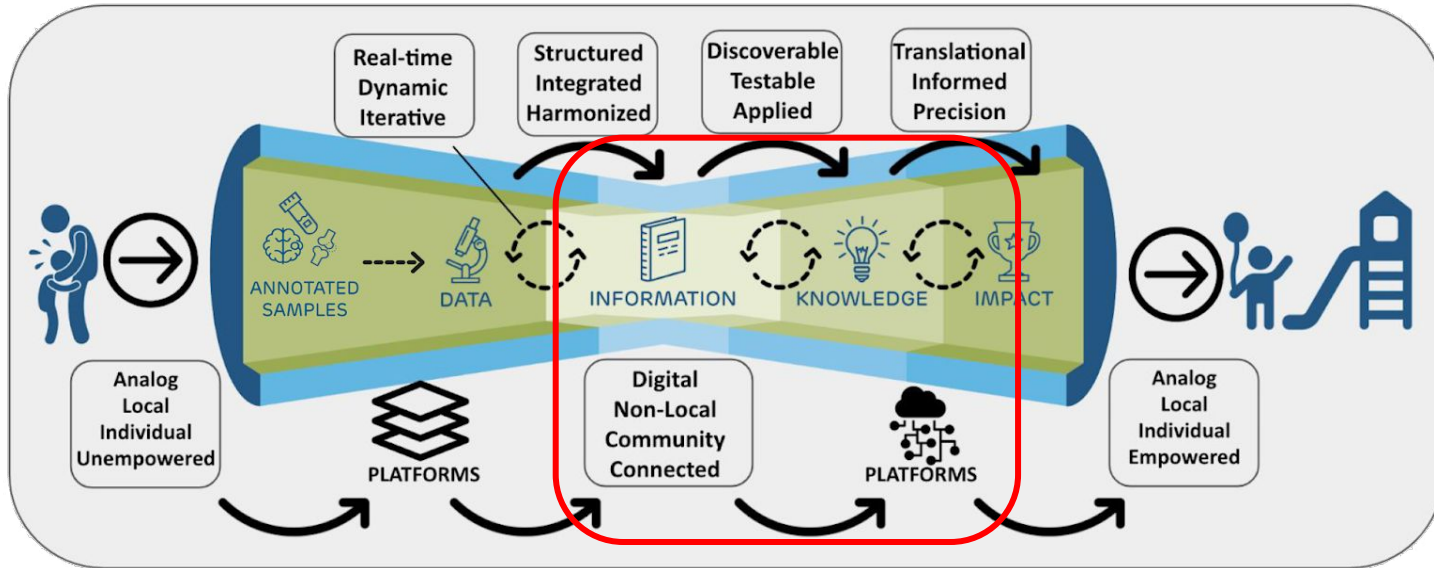


community

late 14c.

a number of people associated together by the fact of residence in the same locality; the common people

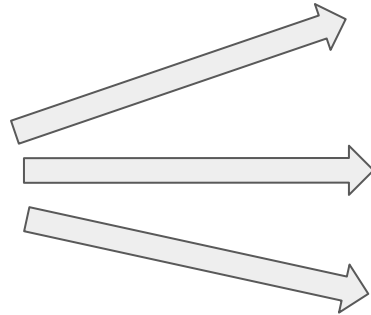
Focus of Last Six Months



Some Common “Themes”

Source Datasets

- “Scattered”
- Asynchronous



Combine datasets

Controls/Annotation/My own data

My own Pipelines/Workflows and combinations

Some Common “Themes”

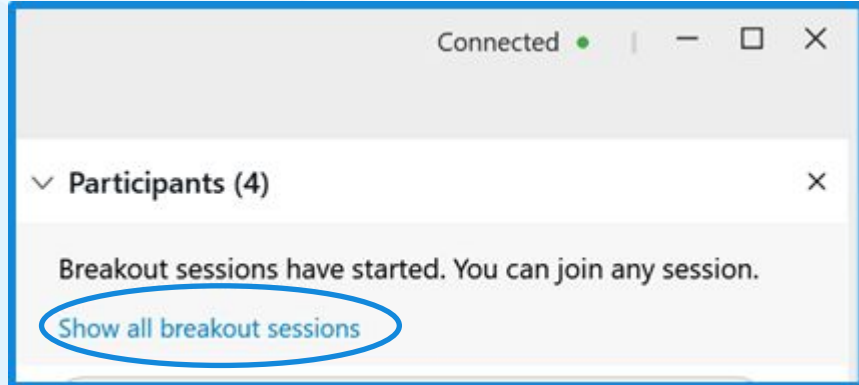
Source Data → Processed data → “New” Matrices

- 1) Can I use that workflow for my data
- 2) Is that workflow portable
- 3) “Dedicated” applications (Shiny Apps, notebooks)

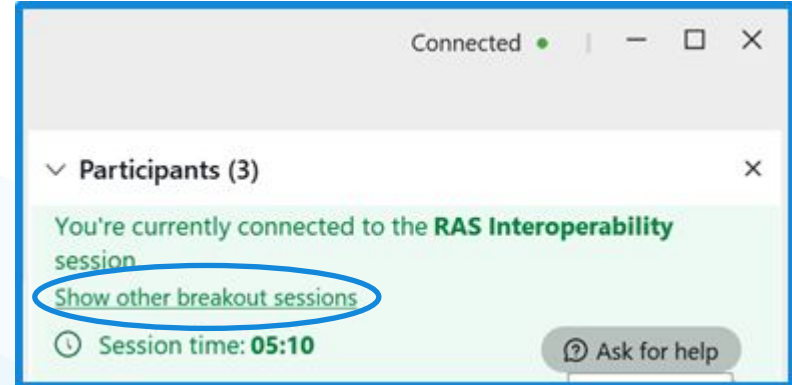
Other modes of “Search”

Breakout Groups: 1:20-2:30pm EDT

Please choose a Breakout Group: You must use the WebEx application



From the main session



From within another breakout group

30 Minute Break #2

We will resume at 3:00 pm EDT

Announcements

- **Last Chance!** Fall 2021 Workshop poll: tinyurl.com/NCPIfallpoll
- Breakout leads have 50 minutes until the Report Backs that begin at 3:20
- The NIH Office of Data Science Strategy recently announced four Notices of Special Interest for supplemental funding: tinyurl.com/ODSSfunding

The Future of Interoperability

Brian O'Connor
Broad Institute



NIH NCPI Effort - Breaking Down Data Silos

*The **NIH Cloud Platform Interoperability (NCPI)** effort empowers end-users to analyze data across participating platforms.*

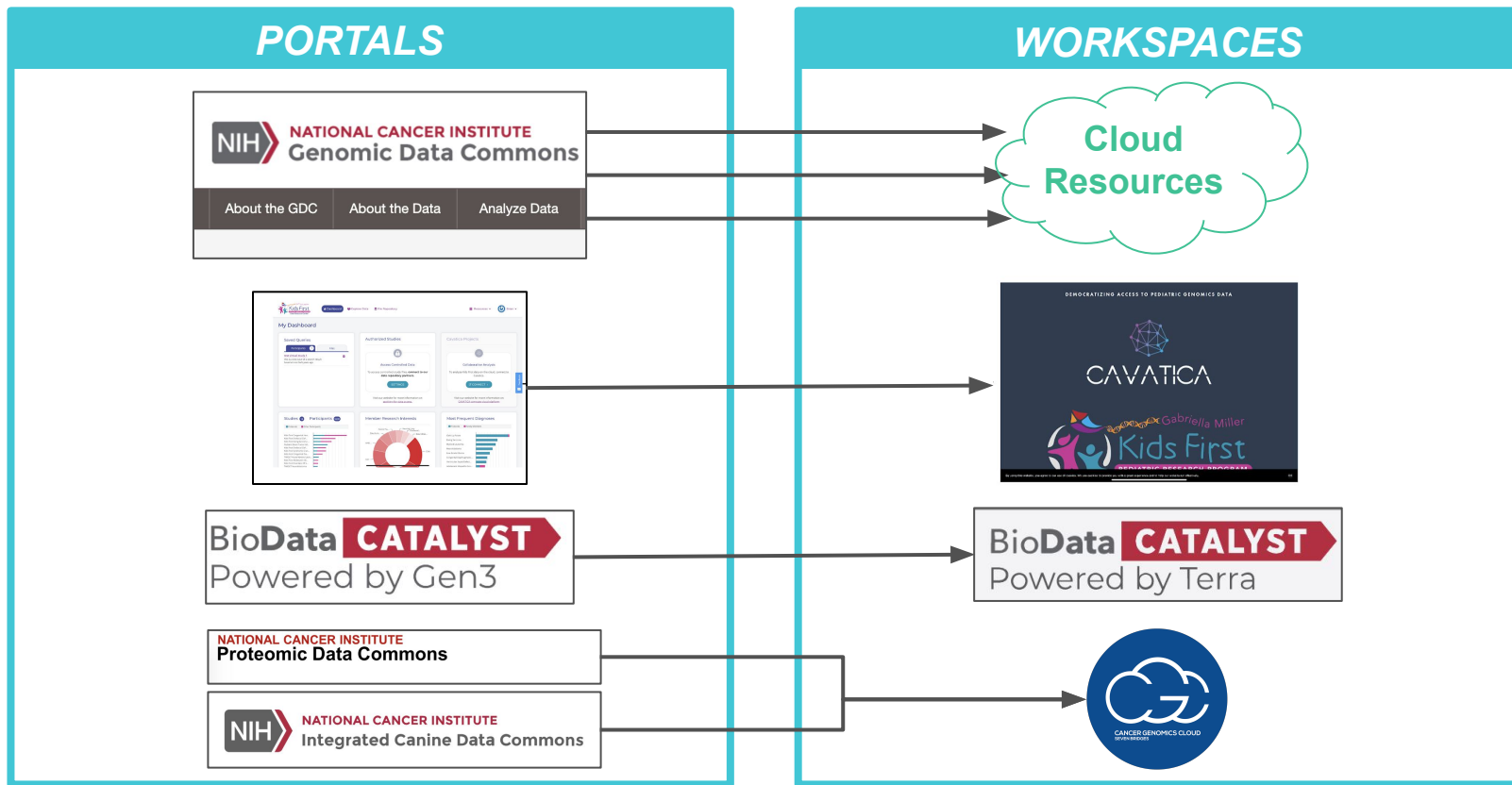
*It facilitates the realization of a **trans-NIH, federated data ecosystem** by establishing and implementing guidelines and technical standards.*



<https://anvilproject.org/ncpi>

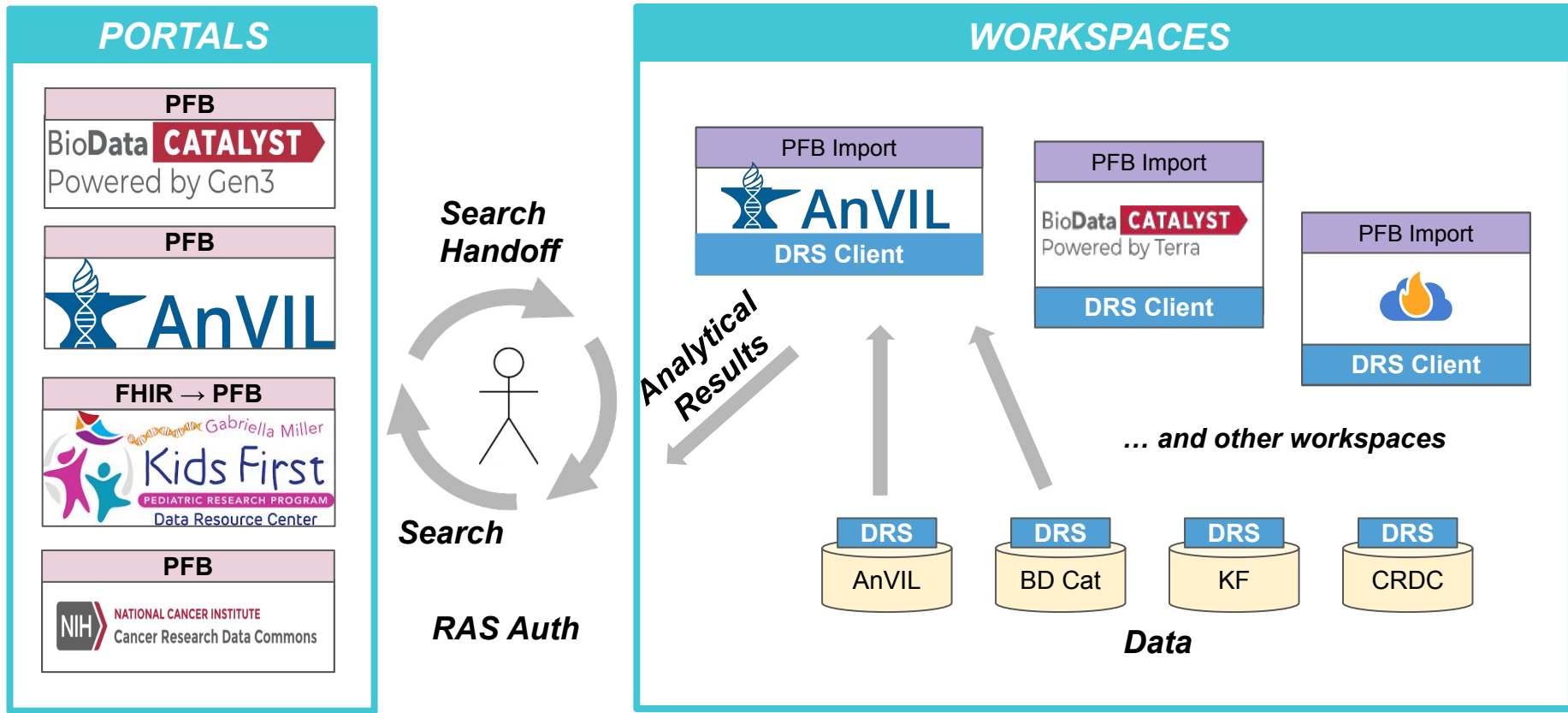
Starting Point in 2020 - NCPI Systems Interoperation

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



NCPI 2020 Vision for NIH Researchers

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



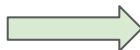
NCPI by the Numbers in 2020

Collectively, we have achieved improved interoperability in 2020 across multiple systems through **FHIR, PFB, GA4GH DRS, and GA4GH Passports (RAS)**.

2020 Results

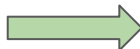
- **Search Result Handoff:** PFB

*2 portals,
~417K subjects*



- **Data Access:** DRS 1.1

4 DRS Servers
~7.6PB of data*



- **Auth:** RAS for AuthN

RAS login



Supported Platforms

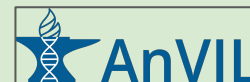
BioData **CATALYST**
Powered by Gen3



BioData **CATALYST**
Powered by Gen3

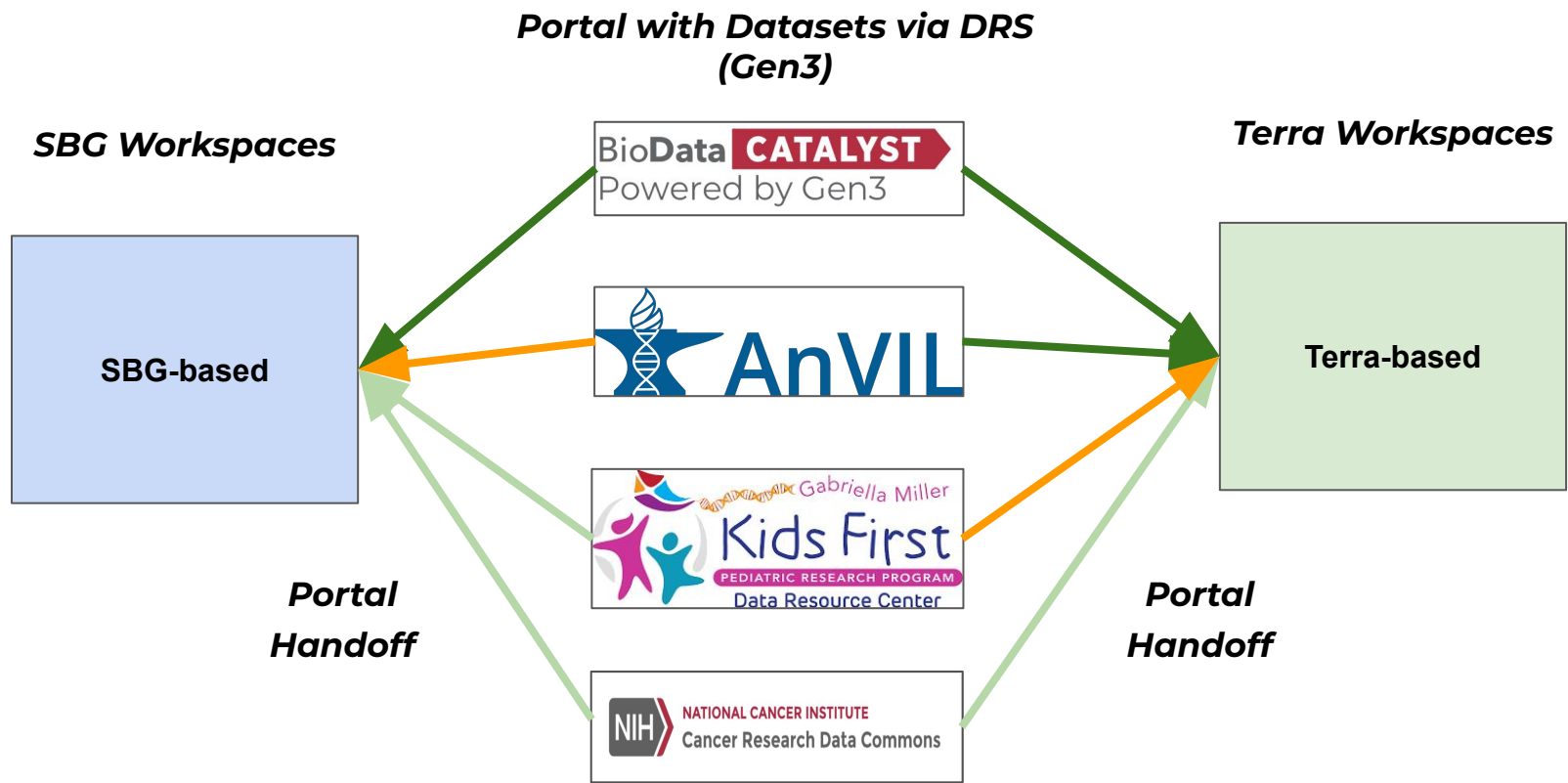


BioData **CATALYST**
Powered by Gen3



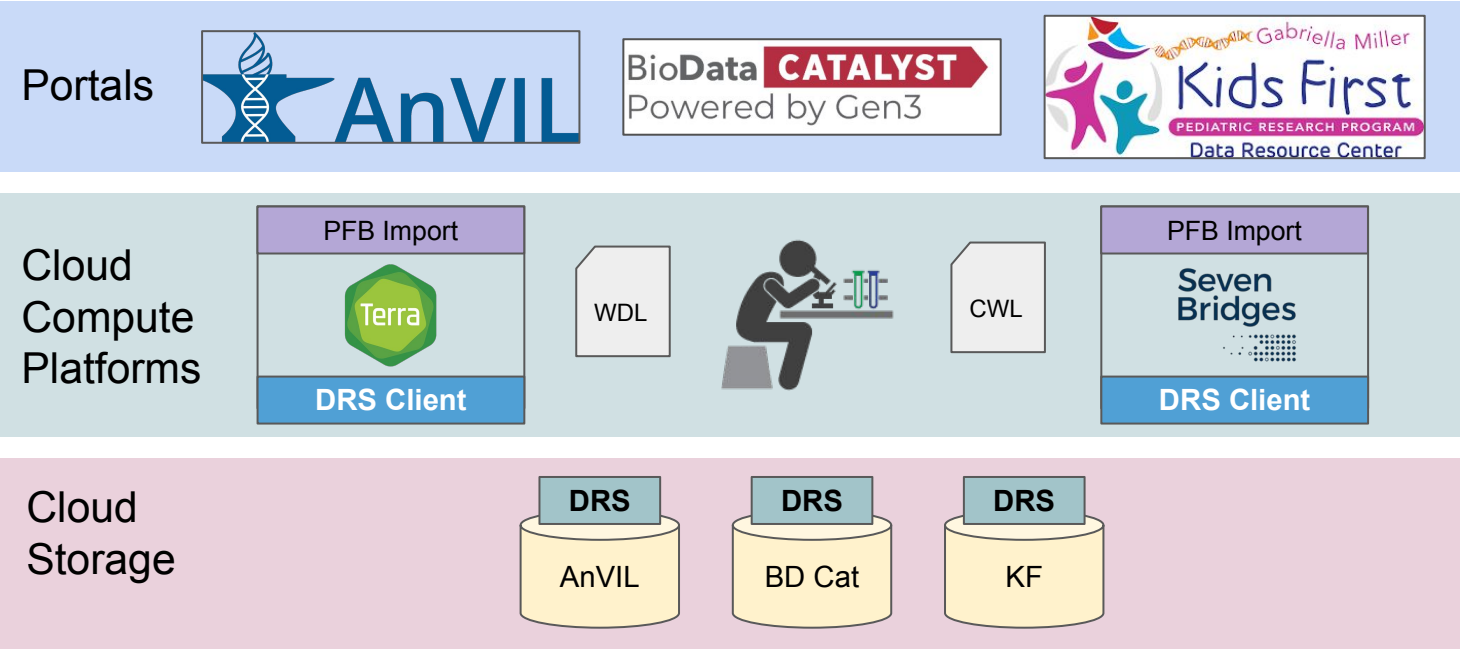
* NCBI DRS server to be added

Demonstrated handoff is now possible from all 4 portals to Terra & SBG workspaces



Supported Researcher Use Case

Use Case #7: Tim Majarian's cross dataset analysis



① Find Data



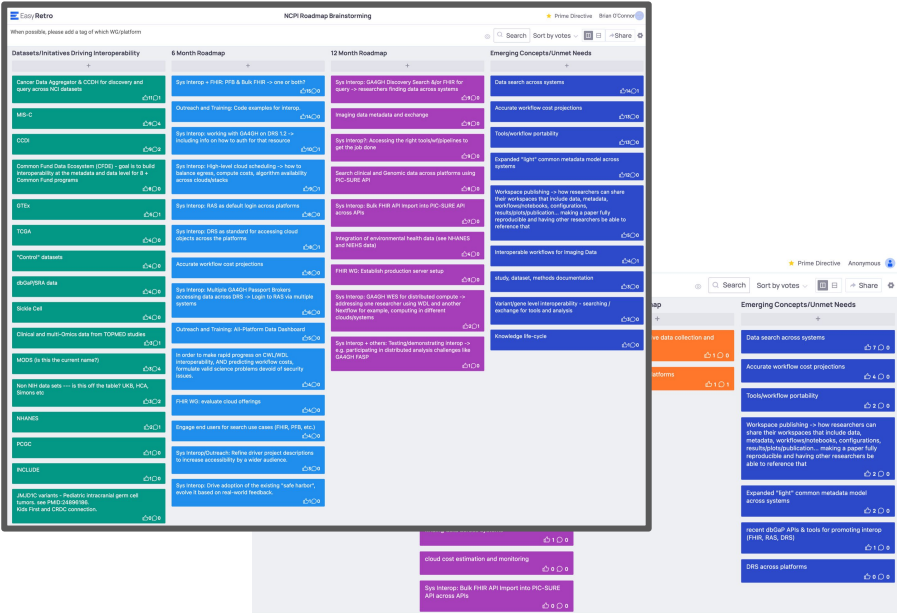
② Compute on Cloud Workspaces



③ Access Data on Cloud Storage

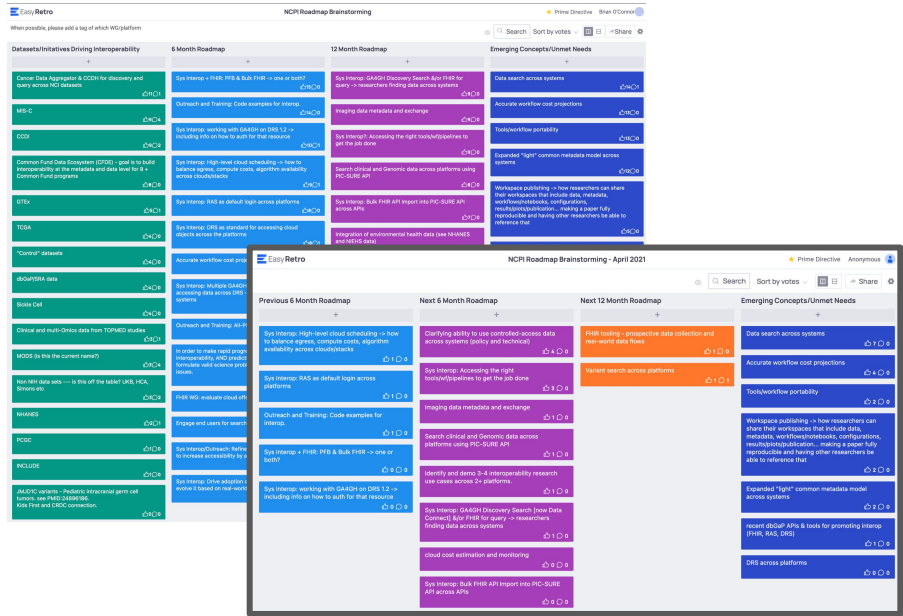
What's Left to be Done?

- In 2020, **FHIR, user authentication and standardized data access** between systems were major achievements!
- *Work in Progress*
 - **Authorization** - How are users authorized using "Passports"?
 - **Standards** - How are standards like FHIR, PFB, & DRS facilitate searching, handoff to workspaces, and data access?
 - **Policy** - What systems should be allowed to access data for a user?



What is the Focus for 2021?

- *What is the next goal post for NCPI Interoperability?*
- We surveyed the group ahead of this meeting using EasyRetro:
<https://bit.ly/3gVHmIN>
- Three major themes emerged



Improving Interoperability in 2021

What themes should we focus on in NCPI for the next 6-12 months?

1) Authorization & Policy - *A user should be able to log in to many NCPI systems using RAS and access data, and possibly other resources, they are authorized to use via their Passport+Visas. Clear policy on client trust and verification.*

28

Now

2) Search - *A user should be able to search across NCPI systems to find data through programmatic and web UIs. Common, standards-based interfaces for doing this.*

78

Next 6 months

3) Portable Compute - *A user should be able to move their algorithm between environments and enclaves, when data egress is not allowed or practical. Publish their workspaces.*

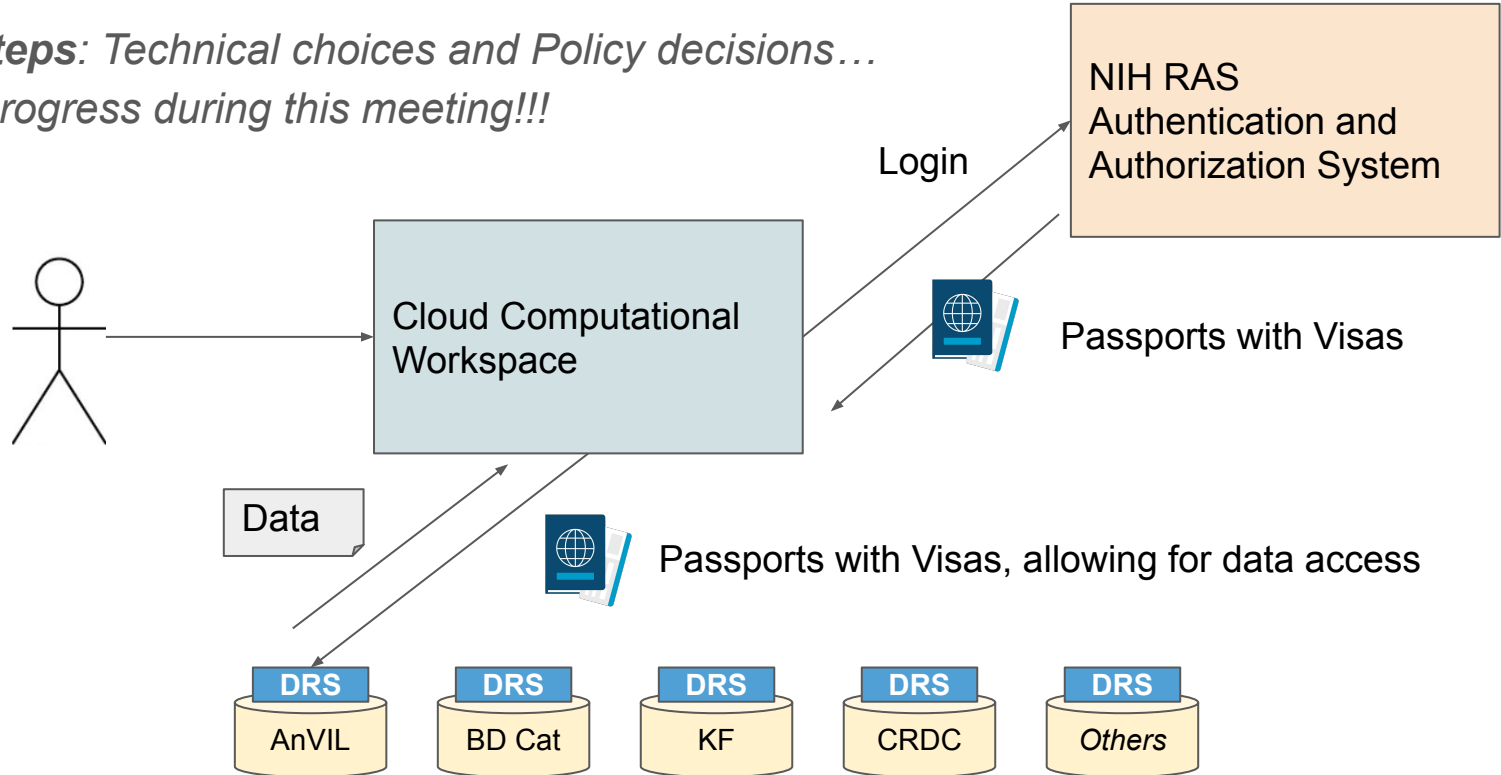
49

Next 12 months

1. Authorization

Problem: In 2020 we used RAS for login. In 2021, users should be able to use their RAS passport to access resources in a variety of systems using a consistent "flow".

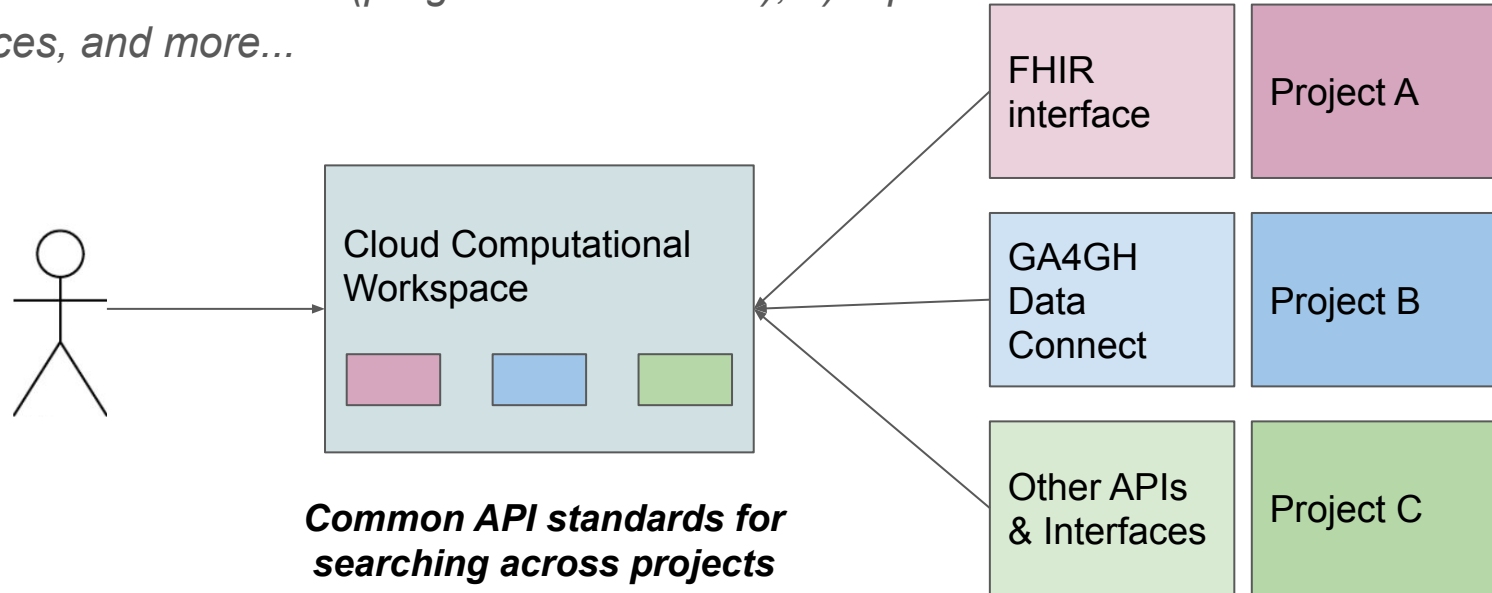
Next Steps: Technical choices and Policy decisions...
Made progress during this meeting!!!



2. Search

Problem: In 2020 focused on researchers finding data through individual portals and leveraging FHIR as a search API. In 2021 can we further empower researchers with standards for search?

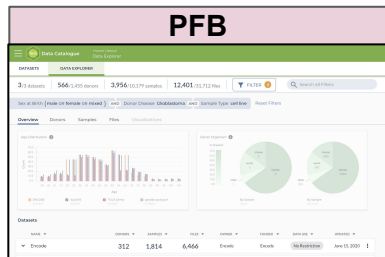
Scope: 1) context (dataset or subject-level), 2) common data model, 3) shared code lists, 4) consistent search interface (programmatic and UI), 5) representation and handoff of results to workspaces, and more...



3. Portable Compute

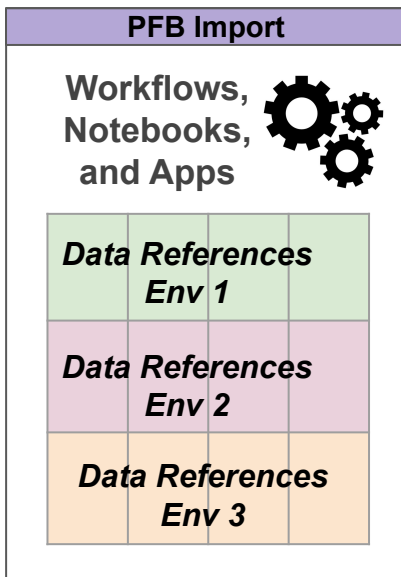
Problem: In 2020 we focused on data access across workspace systems. What about data enclaves where data cannot exit (or need to avoid egress?)

Data Portal

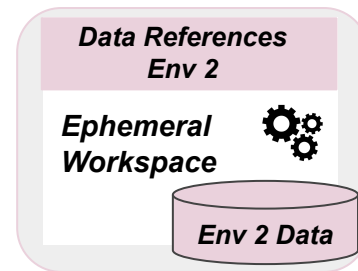
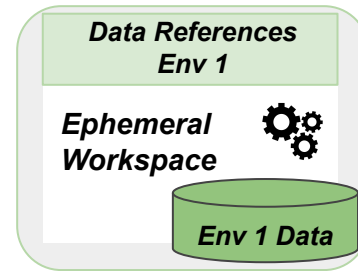


Search result handoff

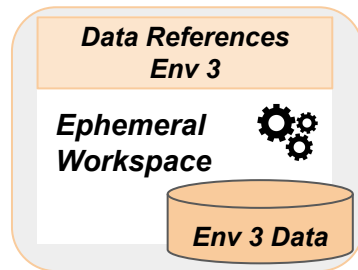
Virtual Workspace With Data Pointers



Workspace subset for analysis



Secure derived data returned



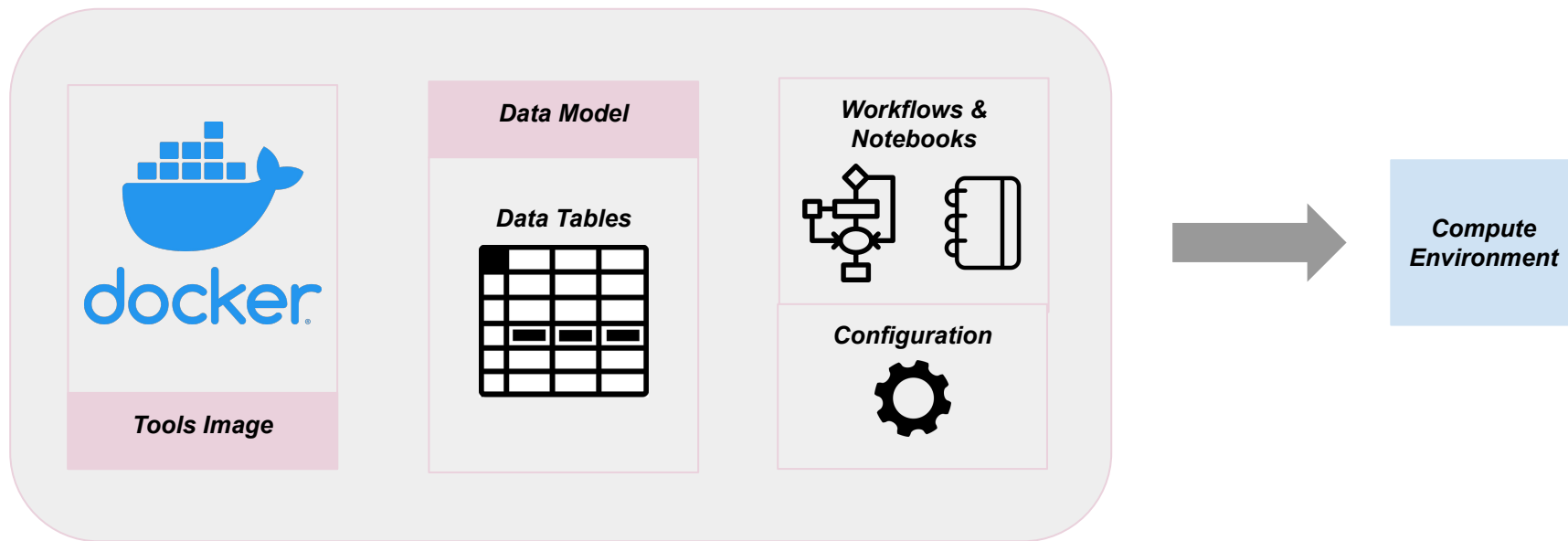
Next Steps: In 2021 can we enable sending algorithms to the data?

3. Portable Compute - Mobile Workspace

Problem: Can we make our workspaces mobile? This goes beyond just workflows.

Next Steps: Adopting mechanism to "package" workflows, notebooks, and apps along with settings, configurations, data models, etc. Making workspaces as FAIR as possible.

Mobile Workspace



Future of Interop in 2021 - Need for Drivers

In 2020 our researcher use cases helped drive our work forward...

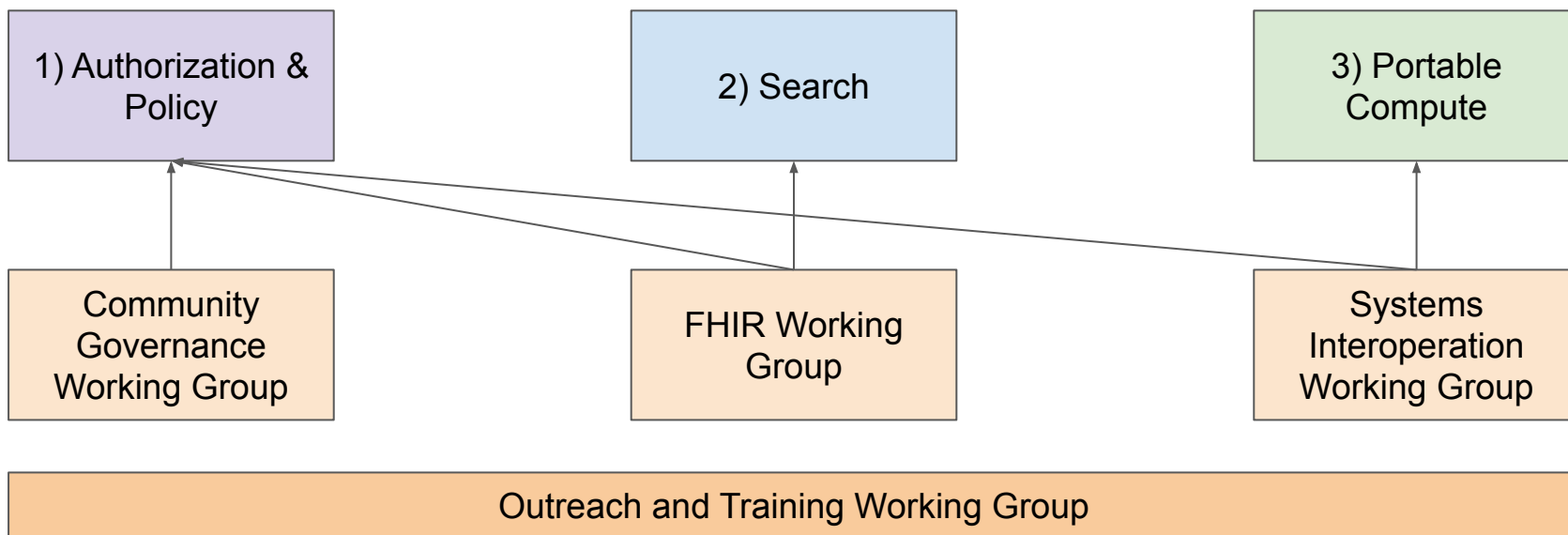
LEAD	ONE-LINE SUMMARY	STATUS
Gelb	PCGC (BDC, KF) <i>de novo mutations</i> with graph callers	Inactive
Grossman	PCGC (BDC, KF) & Vandy AFib joint calling, annotation, and GO enrichment; <i>interop/tech focus</i>	Active
Gharavi	GTEx (AnVIL, KF, BDC) find datasets as healthy controls	Active
Lyons	User journey from PICSURE-API to Platform (TOPMed) for variant level info	<i>In Prep</i>
Stranger	TCGA, GTEx (CRDC, AnVIL) sex-DE on normal & tumor	Inactive
Manning	PCGC, GTEx, F/JHS (BDC, KF, AnVIL) genetic factors in CHD	Active
Almeida	IDC (CRDC) tile server for autoML image analysis; bearer token auth	Active
Goldmuntz, Taylor, et al.	PCGC (BDC, KF) joint calling, harmonization, gene set analysis + ML	Active

In 2021 we need to expand use cases... both individual researcher as well as cross institute

Future of Interop in 2021 - Working Groups

From a practical perspective, how do we move these themes forward?

Key is to use our working groups to align, scope, and organize these efforts.



If We Are Successful

Researchers will be able to **safely and securely access data and resources** from a variety of platforms, **carrying their identity and authorizations via RAS Passport+Visas**

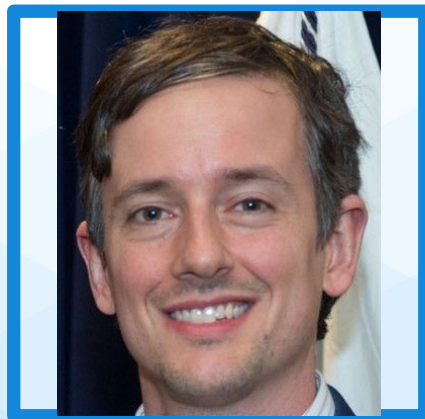
Researchers will be able **find data across a wide variety of systems** through consistent, standardized interfaces

Researchers will be able to **compute by pulling data into the platform of their choice or by sending their algorithms in a portable way to other platforms**

Breakout Session Report Back

Data/Tools/Workflow/Compute Interoperability and Functional Equivalence

Jack DiGiovanna
Seven Bridges



Michael Schatz
Johns Hopkins



Like a superhero movie, we are rebooting with a slightly different cast

Here we'll narrow focus with the **goal of actionable outcomes**

What solutions can we create in **the next 6 months?**

NIH Workshop on Cloud-Based Platforms Interoperability
October 30th and November 2nd, 2020

Genomic Analysis Use Cases & Working Groups



Jack DiGiovanna¹ & Michael Schatz^{2,3}

¹Program Director - Seven Bridges

²Program Director - AnVIL

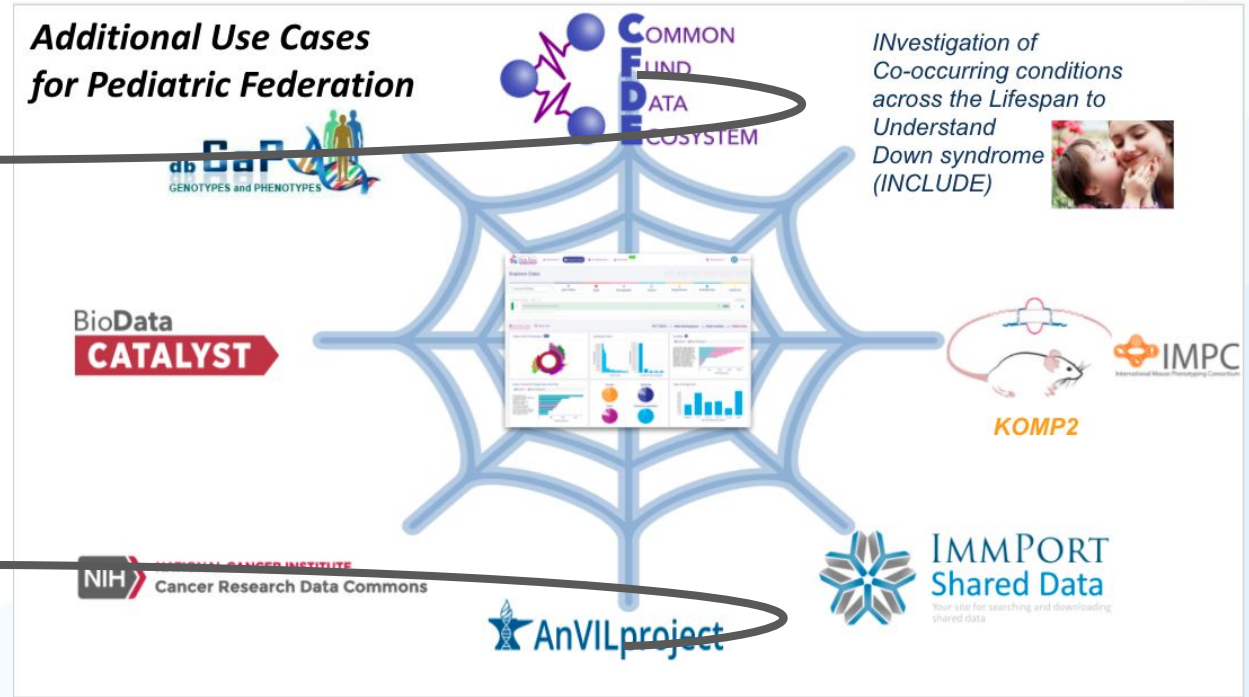
³Bloomberg Distinguished Associate Professor - Johns Hopkins

Representative use case

COMMON
WORKFLOW
LANGUAGE

{wdl}

**Additional Use Cases
for Pediatric Federation**



slide credit: Cotton & Resnick



Bring compute to the data!



As a researcher, likely with a sizeable development investment on HOME_PLATFORM, how should I interact with data on OTHER_PLATFORM?

- Bring me the data
 - can work great, even with pointers... what about egress, enclaves, and bears? Oh my!
- Rewrite my code in OTHER_DESCRIPTION_LANGUAGE
 - given infinite time, funding... is this the best investment?
- Don't use the data
 - Data value declines sharply with cleaning/logistics required to use it
- Send compute to OTHER_PLATFORM

This is NOT a bake-off

We are **NOT** asking for a bake-off here

That likely happened *before* the researcher reached this point

The research needs to run (two slightly different versions of) this **best tool** in two different places **today**





Functional Equivalence Challenges



- **Challenges**

- The **data** you need are spread across platforms
- The **workflows** you need are spread across platforms
- How do we compare and integrate workflows across platforms
- We don't want to make work items for folks who aren't funded to do this.
 - Focus on critical use cases for the NIH.

- **Range of outcomes**

- **Format conversions** - we expect/hope these will produce identical results but not always e.g. CIGAR strings have a 64kb limit in BAM files but not SAM
- **Primary analysis** (alignment/variant calling): we expect/hope these will produce similar results but random numbers, machine architecture, etc may slightly vary
 - Changes in reference genomes are challenging to adopt
- **Downstream analysis** - we expect different outcomes, hopefully small but could be substantial e.g. t-SNE is stochastic by design



Testing outcomes



- **Equivalence testing**
 - Focus on research outcomes: variant calls, associations are highly similar
 - CCDG/TopMed found 99.6% concordance with variant calls
 - Lessons learned from RNAseq - make sure the biological variability you report exceeds the technical variability observed from the replicates
 - Need many replicates as results may subtly change with time of day / ordering of data / reference sequences used
 - External databases / APIs can introduce unpredictable changes
- **What are the workflows to consider?**
 - SNVs are relatively stable, indels more challenging, SVs have highly variable

- **GA4GH WES/TES endpoints**
 - **Workflow Execution Service (WES): Abstract workflow descriptions**
 - “WES enables users to define workflows in a standard way, package them up, and then hand them to workflow engines that live in many different places”
 - **Task Execution Services (TES): Fully defined Input/outputs, command lines**
 - Orchestrate complex analyses across different compute environments. While the WES API orchestrates a series of steps in a workflow, the TES API can connect the workflow to a compute backend to execute specific steps without having to write new adaptors.
- **Cross-workflow engine**
 - Docker useful for packaging tools into a reproducible container, but hard to scale
 - Many workflow languages have support for K8s
 - AnVIL/Babble: Initial support for Snakemake
 - SB Considering similar technical developments

Call to action

- **We'd like a group** (Sys Interop, Tiger team, other?) **to investigate over next 6m**
- **Best Practice WF to compare**
 - Sex chromosome variant calling
 - Long Read Pipeline on Terra and SB (ONT WGS)
 - RNA-Seq
- **Work with truth-set test data**
 - 1000g
 - GIAB
 - GRU data discussed, but some sensitivities there
- **Working towards a SOP that's generalizable**
 - other WF are further down the horizon than 6 m



WE NEED YOU!

Breakout Session Report Back

Cloud Costs & Benchmarking

Alex Baumann
Broad Institute

David Pot
GDIT



Breakout process

- We had a series of possible topics in three main categories:
 - User Experience, User Education, Managing our Systems
- We had each person use 3 *'s to vote on their top 3 choices
- We landed on the following topics:
 - **Guiding/educating** our users on cloud costs & benchmarking
 - **Monitoring / alerting** for cloud costs
 - **Overcoming barriers to entry** for new users on the cloud

Guiding/educating our users on cloud costs & benchmarking

- Grant guidance and stock language
 - Cloud Resources should collaborate with NIH agencies on this
- Sharing of benchmarked results for standard analyses and having more than one datapoint to extrapolate
- Galaxy's approach: Use popular tools, look at historic data, run tests and build up a lookup table/API. Also use a tool called Polyester to generate synthetic data, try combinations of inputs

Guiding/educating our users on cloud costs & benchmarking

- Terra's approach: Test with open access data of different size (exome, WGS). Run a few times for average cost. Publish in featured work-spaces
- Warning people of what costs \$ and what to avoid: e.g. SSD left running, deduplication of data
- Benefits of whiteglove support and viral growth via super users in labs (educate the educator)

Monitoring / alerting for cloud costs

- Delayed reporting in costs from clouds (e.g. 24 hours later)
 - Need for near-real time reporting / no surprises for our users
- Setting up alerts and budgets for users to see burndown
- Providing whiteglove support on credit spending so those credits are closely monitored
- Compute tends to be the biggest concern for “runaway spend” - storage is a longer concern, but builds up over time
- Non-linearity for cost estimations can be problematic

Overcoming barriers to entry for new users on the cloud

- Comparing/convincing about on-prem vs cloud advantages despite new cost model
- Capital vs. operational expenses of institutions
- Groups need a motivation to convince them to change - top down, need that on-prem doesn't satisfy, etc - cloud is a disruptive change
- Cloud expertise is an issue - white glove service enormously helpful
- Having existing content and a community helps to reduce barriers

Went a bit off topic into convincing people to use cloud despite costs...

But that led to a set of possible recommendations:

- Keep up with free credits but make sure they are well tracked in near real time
- Clearly communicate costs and define what error margins people are comfortable with
- Increase NCPI training efforts / training on how to understand costs?
- Whiteglove help for viral growth to larger communities
- Create example benchmarks across platforms for standard pipelines
- Find ways to make cloud solutions as equivalent as on-prem functionality

continued ...

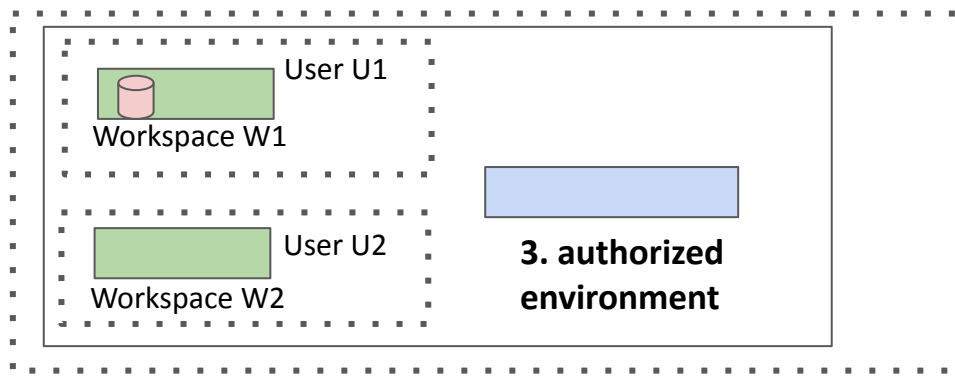
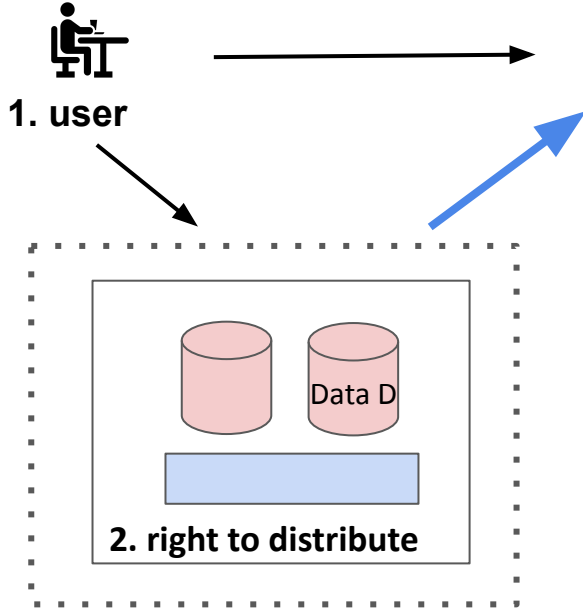
(Continued) Set of possible recommendations:

- Advertise availability of high value data (e.g. open access)
- Encourage top down incentivization in institutions
- Work on language to write in grants that use the cloud for research
- Provide additional costs if using cloud created surprise costs / cloud insurance / on-prem price matching
- Continue collaborative discussions across NCPI to share solutions/experiences ...

Breakout Report Back: Governance

Bob Grossman (UChicago) & Stan Ahalt (RENCI)

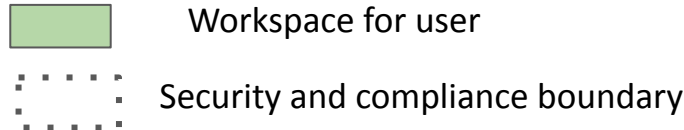
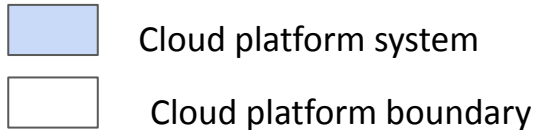
Four Key Concepts



Cloud Platform B boundary

1. A **user is authorized** to access a dataset
2. A cloud platform A has the **right to distribute** a particular dataset.
3. A cloud platform B is an **authorized environment** for a particular dataset.
4. Each dataset has a **data trustee** (aka **data steward**) that makes decisions about 1), 2) and 3)

We have **interoperability** when an authorized environment can access data from two or more cloud platforms..



Can we agree on these two considerations?

- Authorized environment Consideration (draft). Assume that the data steward responsible for a dataset D has approved a cloud platform B as an authorized environment for D. If a user in the cloud platform B is authorized to access the data, then the user can access the data within the authorized environment B.
- Right to distribute Consideration (draft). Assume that the data steward responsible for a dataset D has authorized a cloud platform A to distribute the dataset. Assume the data trustee has also approved cloud platform B as an authorized environment for the dataset D. If a user in a cloud platform B is authorized to access data D, then the user can access the data D from cloud platform A and analyze it in cloud platform B.



General Consensus



Question: Can we agree on these two considerations?

Response: Yes, we seem to have general agreement about these two considerations, but some wordsmithing is needed.



The Importance of Trust



1. POV of the considerations: The CISO (“data steward”) makes the decisions about which systems to “trust” and a cloud platform interoperates with systems that they trust.
2. **Trust is not a formula. It is a relationship that has been established between two platforms.**
3. We trust the other current NCPI systems, but what other systems?
4. If the levels of security vary between two systems, how should they interoperate?
5. Only “trust” systems (and thus interop with them) with the same or higher level of security (that is required for the data).
6. Ultimately, an IC has to decide if a relationship of trust exists, and the risk is reasonable.
7. RAS authorizes users, not systems. You can use, for example, SSL tunnels, to identify another system and decide whether to trust it.



dbGaP Agreements and the Considerations



1. Remember users have signed a legal DUC stating they, along with their SO, are responsible for their use.
2. SO-approval described in Data User Certifications:
https://osp.od.nih.gov/wp-content/uploads/Model_DUC.pdf
3. **These considerations do not require any changes to the current dbGaP agreements, but clarifications to them would be welcome** that highlight the compatibility with these considerations.
4. Some formal level of agreement necessary to constitute authorization for NCPI style interop should be spelled out in the dbGaP agreements. How would these be made visible so the Signing Officials



Opportunities available with these considerations



1. There is an interest to track data migration across systems and report back to the data stewards. In some cases, e.g., data that is downloaded, you can't easily (or actually) track where data travels.
2. how is user's use of controlled access data in the remote platform reported back to the data steward.



Requested Clarifications for Considerations



1. I would like to see an explicit definition of data steward
2. I think it's important to know how to outline what is needed in the current process to be able to satisfy these considerations
3. Are there consistent rules/principles/security considerations that could be NIH baseline for any platform, regardless of what it is?

NCPI Spring 2021 Workshop Day 2 Wrap Up

- Thank you for a fantastic meeting!
- Speakers please send us your presentations from today
- **Fall 2021 NCPI Workshop dates: Oct 5-6**
 - **To be hosted by NHLBI and RENCI**
- Feedback poll for this workshop:

tinyurl.com/NCPIfeedback