Gabriella Miller Kids First Pediatric Research Program

Public Webinar

October 6, 2020
2:00 pm EDT

NIH
The Common Fund

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM

Webinar Agenda

2:00pm - Introduction; *NIH Kids First Staff*
2:05pm - Progress on the Kids First Study on Novel Cancer Susceptibility in Families; *Dr. Sharon Plon, Baylor College of Medicine*
- Questions for Dr. Plon

2:35pm - Kids First Data Resource Center; *Kids First DRC Staff*
- New Portal Features
- Kids First Variant Workbench Demo

3:05pm - NIH Common Fund Data Ecosystem (CFDE) presentation; *Dr. C. Titus Brown, UC Davis*
3:25pm - NIH Program Updates; *NIH Kids First Staff*
3:40pm - Questions & Answers

NIH
The Common Fund

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

Valerie Cotton
Kids First Program Manager
Eunice Kennedy Shriver National Institute of
Child Health and Human Development (NICHD)

NIH Kids First Working Group

Kids First is an NIH Common Fund program coordinated by a **trans-NIH Working Group**, which is chaired by four institutes:

Eunice Kennedy Shriver National Institute of Child Health and Human Development (**NICHD**)

National Human Genome Research Institute (**NHGRI**)

National Heart, Lung, and Blood Institute (**NHLBI**)

National Cancer Institute (**NCI**)

The Common Fund

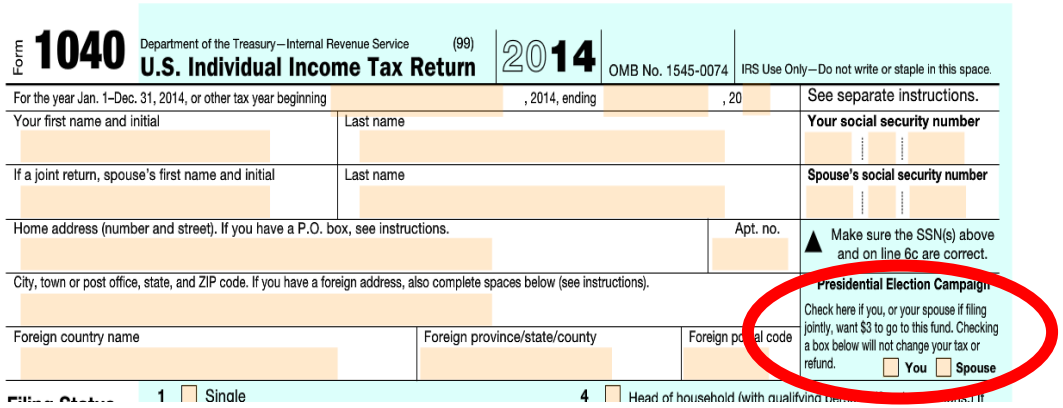
Other Working Group Representation:

NIDCR **NIAAA** **NIDDK** **NEI** **NIAID** **ORIP**

NIDA **NINDS** **NIEHS** **NIAMS** **NCATS** **CDC**

How did Kids First get started?

- Initiated in response to the 2014 Gabriella Miller Kids First Research Act:
  - Signed into law on April 3, 2014
  - Ended taxpayer contribution to presidential nominating conventions
  - Transferred $126 million into the Pediatric Research Initiative Fund
  - Authorized appropriation of $12.6 million per year for 10 years to the NIH Common Fund for pediatric research; first appropriation was for FY2015



Form **1040** Department of the Treasury—Internal Revenue Service (99) **U.S. Individual Income Tax Return** 2014 OMB No. 1545-0074 IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2014, or other tax year beginning , 2014, ending , 20 See separate instructions.

Your first name and initial | Last name | Your social security number

If a joint return, spouse's first name and initial | Last name | Spouse's social security number

Home address (number and street). If you have a P.O. box, see instructions. | Apt. no. | Make sure the SSN(s) above and on line 6c are correct.

City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below (see instructions). | **Presidential Election Campaign**

Check here if you, or your spouse if filing jointly, want $3 to go to this fund. Checking a box below will not change your tax or refund. ☐ You ☐ Spouse

Foreign country name | Foreign province/state/county | Foreign postal code

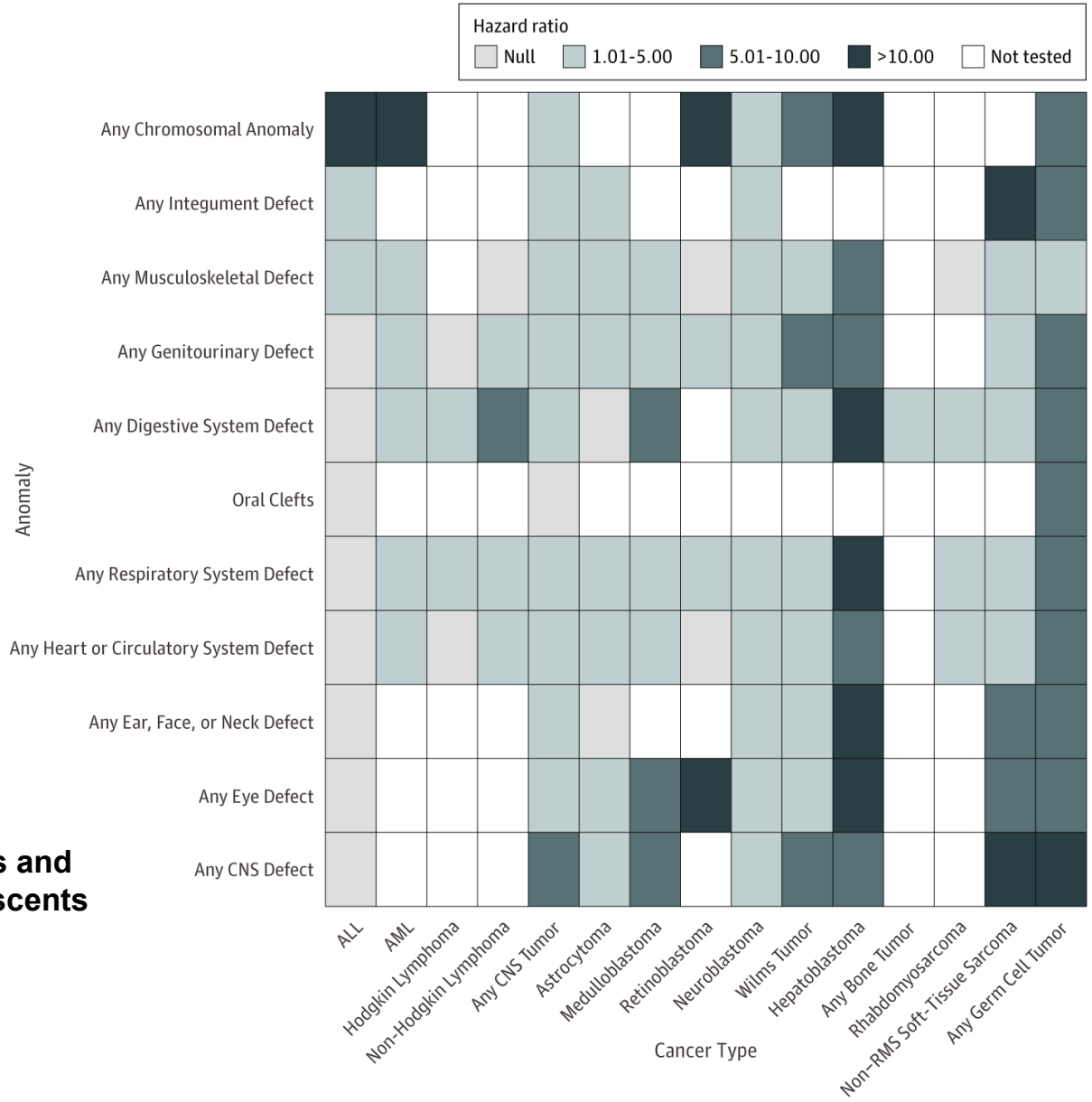
Filing Status 1 ☐ Single 4 ☐ Head of household (with qualifying

Vision

Alleviate suffering from childhood cancer and structural birth defects by fostering **collaborative research** to uncover the etiology of these diseases and supporting **data sharing** within the pediatric research community.

Why study childhood cancer & structural birth defects together?

- ***Both are leading causes of childhood mortality***
- ***Birth defects are associated with increased risk of cancer among children… suggesting shared genetic pathways***



From: **Association Between Birth Defects and Cancer Risk Among Children and Adolescents in a Population-Based Assessment of 10 Million Live Births**

Lupo et al, JAMA Oncol. 2019;5(8):1150-1158. doi:10.1001/jamaoncol.2019.1215

Kids First Major Initiatives

Through 2021:

1. Identify & sequence cohorts of children with **childhood cancer and/or structural birth defects**.
2. Build the **Gabriella Miller Kids First Data Resource** to empower discovery

Year | 15 ✓ | 16 ✓ | 17 ✓ | 18 ✓ | 19 ✓ | 20 ✓ | 21 | 22 | 23 | 24

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

Phase 2

The Kids First Dataset is Growing!

40 projects | 40,000 genomes | 16,000 cases | 14 released datasets

BROAD INSTITUTE | HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY | St. Jude Children's Research Hospital

- Disorders of Sex Development
- Congenital Diaphragmatic Hernia
- Ewing Sarcoma
- Structural Heart & Other Defects
- Syndromic Cranial Dysinnervation Disorders
- Cancer Susceptibility
- Adolescent Idiopathic Scoliosis
- Neuroblastomas
- Enchondromatoses
- Orofacial Clefts in Caucasian, Latin American, Asian & African, Filipino populations
- Osteosarcoma
- Familial Leukemia
- Hemangiomas, Vascular Anomalies & Overgrowth
- Craniofacial Microsomia
- Intersection of childhood cancer & birth defects
- Microtia
- Esophageal Atresia and Tracheoesophageal Fistulas
- Kidney and Urinary Tract Defects
- Nonsyndromic Craniosynostosis
- Bladder Exstrophy
- Hearing Loss
- Cornelia de Lange Syndrome
- Intracranial & Extracranial Germ Cell Tumors
- Fetal Alcohol Spectrum Disorders
- Myeloid Malignancies + overlap with Down syndrome
- Congenital Heart Defects & Acute Lymphoblastic Leukemia in Children with Down Syndrome
- Structural Brain Defects
- Structural Defects of the Neural Tube (Spina Bifida: Myelomeningocele)
- CHARGE Syndrome
- Laterality Birth Defects
- T-cell Acute Lymphoblastic Leukemia
- Pediatric Rhabdomyosarcoma
- Valvar Pulmonary Stenosis

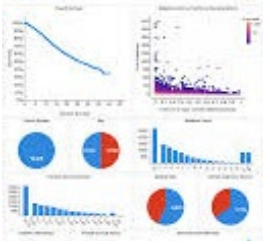
The Kids First Data Resource for Collaborative Discovery

Data Resource Portal

Entry point. Query, search, discover, build & visualize synthetic cohorts

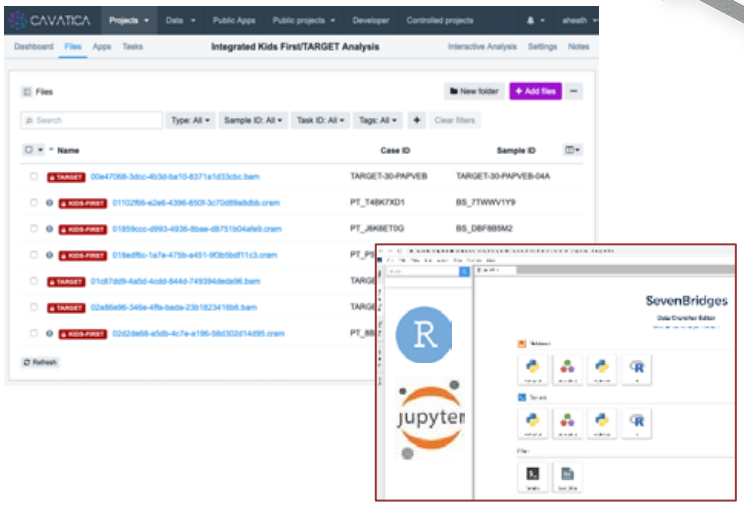
Knowledge Base Integrations (PedcBioPortal)

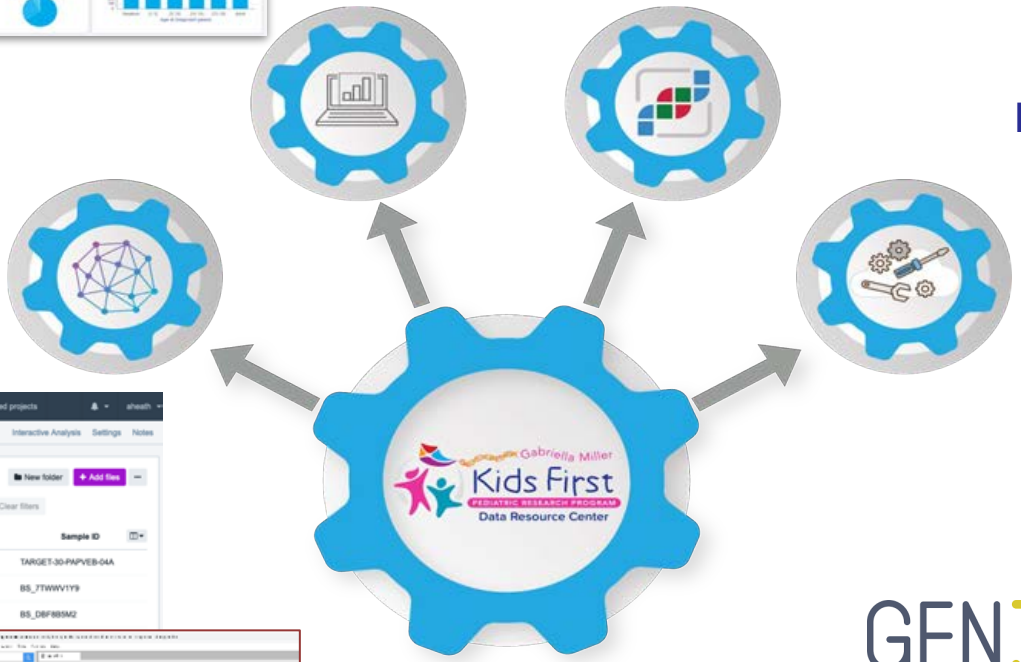
Integrations with existing curated/published data visualizations



Cavatica

Pull data from multiple sources into one workspace.
Use notebooks, bring-your-own or use available workflows.





Data Services

Model clinical data in FHIR-based data services for semantic interoperability and coordination

Framework Services

GEN3 DATA COMMONS

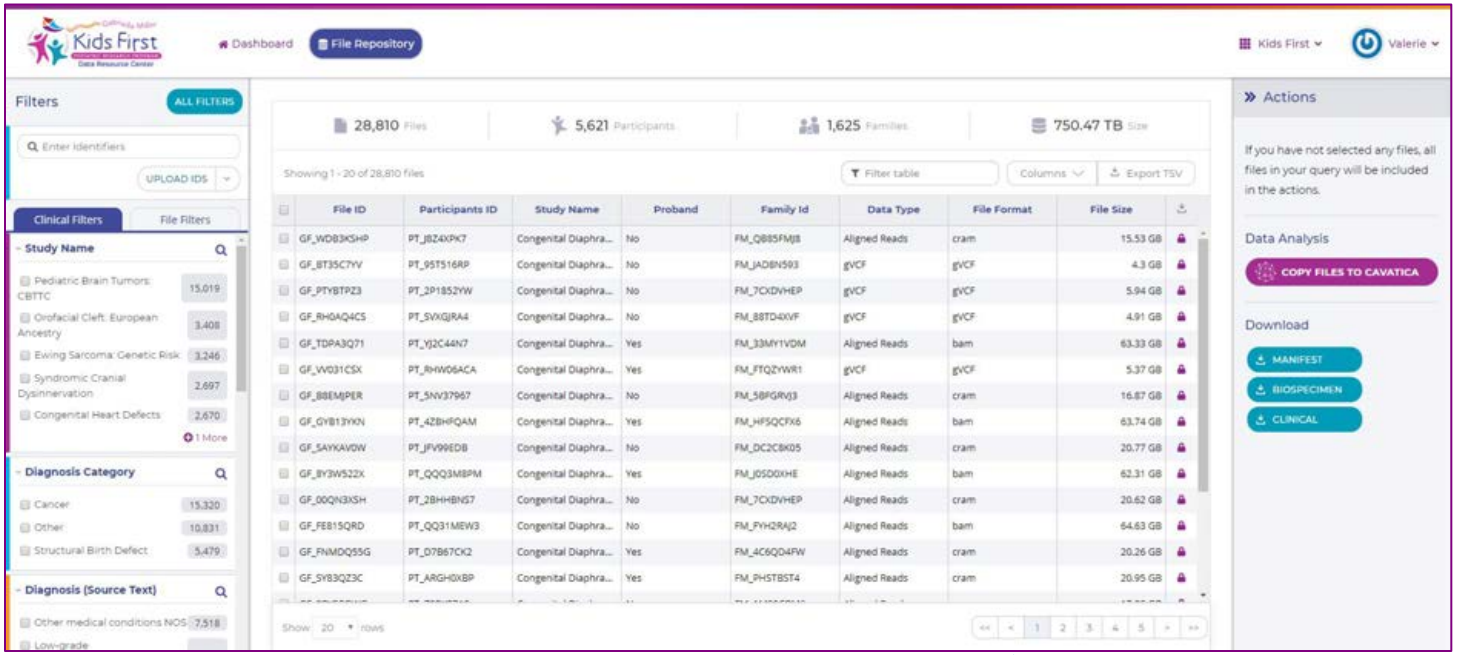
Index and point to files in the cloud (for approved users)

aws

STRIDES

More researchers are accessing Kids First data!

>1700 registered users since 2018 launch



Individual-level sequence data

<u>>150</u> Data Access Requests approved by the Kids First Data Access Committee across **14** Kids First genomic datasets available

NIH Kids First Data Access Committee

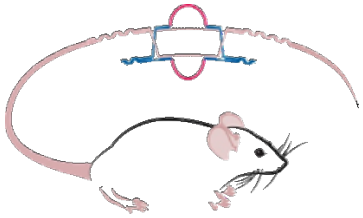
dbGaP GENOTYPES and PHENOTYPES

Researchers are using Kids First data to answer new scientific questions

- ***14 awards*** *for R03 for analyses of Kids First data* (PAR-16-348 ; PAR-18-733; PAR-19-069, PAR-19-375)
- ***2 awards for NIDCR R03*** (PAR-16-070)
- ***2 awards for INCLUDE R03*** (RFA-OD-20-006)

- ***3 awards for R01s*** (PA-13-302, PAR-17-236)

- Spurred **new collaborations** with KOMP2 & INCLUDE



Knockout Mouse Phenotyping Project (KOMP2)

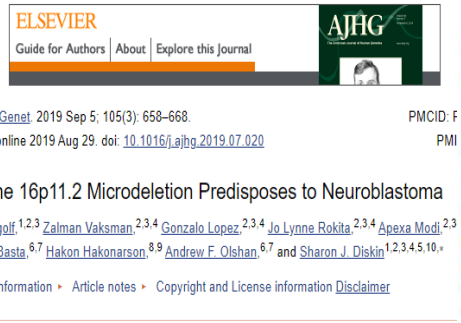
INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE)

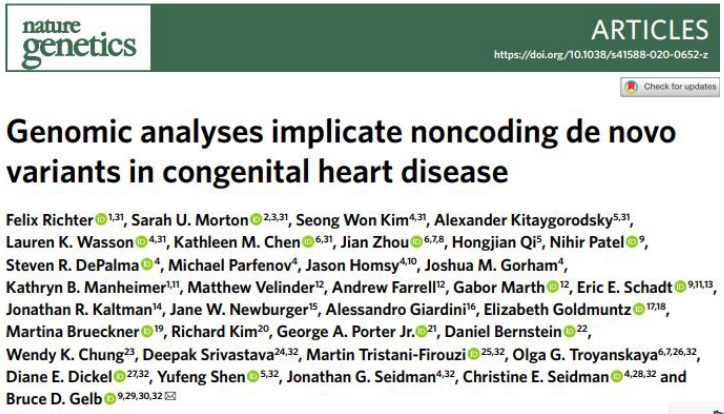
Researchers Are Making Discoveries

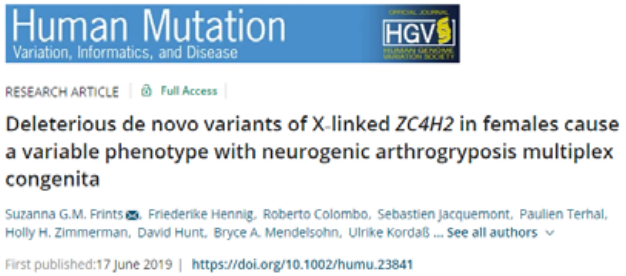
19 Journal Publications To-Date

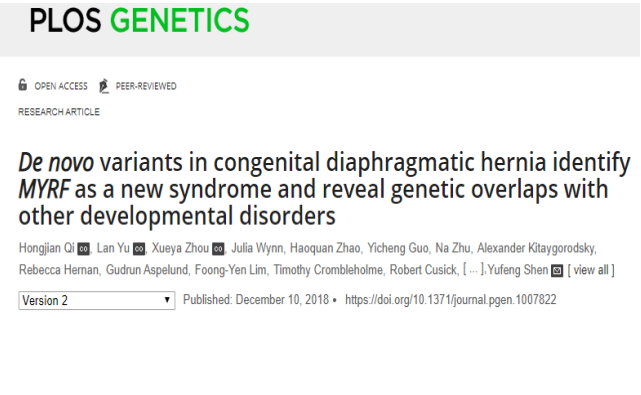
Average RCR: 1.36

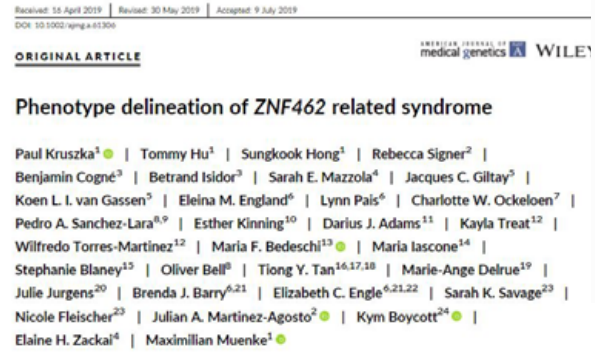
https://commonfund.nih.gov/publications?pid=40

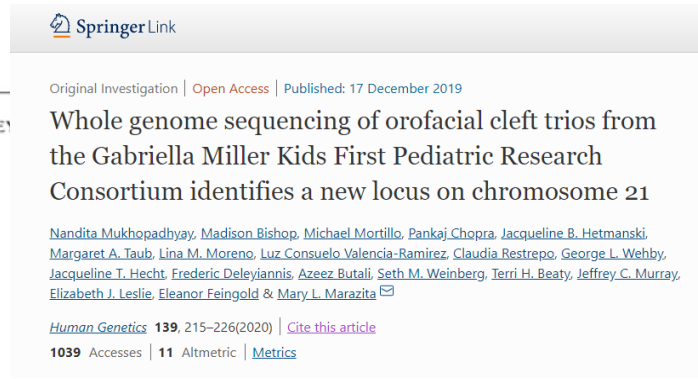












Kids First X01: Identifying novel cancer susceptibility mutations from unselected childhood cancer patient and parent trios

Sharon E. Plon, MD, PhD, FACMG
Baylor College of Medicine

Owen Hirschi
Baylor College of Medicine

1 X01 HL136994-01 - Identifying novel cancer susceptibility mutations from unselected childhood cancer patient and parent trios (BASIC3)

Sharon E. Plon, MD, PhD, FACMG

Departments of Pediatrics/Hematology-Oncology and Molecular and Human Genetics

Human Genome Sequencing Center

Baylor College of Medicine

Baylor College of Medicine

Texas Children's Cancer and Hematology Centers

Disclosures – Sharon E. Plon, MD, PhD

- I have the following financial relationships to disclose:
  - I am a member of the Baylor Genetics Laboratory Scientific Advisory Board

Three different approaches to understanding genetic basis of childhood cancer

- Several studies of clinical genome-scale testing of diverse cohorts of childhood cancer patients (BASIC3 and Texas KidsCanSeq study)

- Sequencing of large cohorts of specific pediatric cancer patients

- Precision oncology treatment trials of matched tumor/normal sequencing for patients with relapsed or recurrent tumors (NCI/COG Pediatric MATCH Trial)

Clinical Sequencing Exploratory Research

*Moving the **Genome** Into the **Clinic***

377 Researchers
21 Institutions
1 Consortium

cser
Clinical Sequencing Exploratory Research

www.genome.gov/CSER
www.cser-consortium.org

Baylor College of Medicine BASIC3 Key Team Members

Will Parsons
Pediatric Oncology

Murali
Chintagumpala

Stacey
Berg

Richard
Gibbs

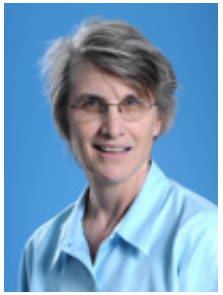
Christine
Eng

David
Wheeler

Amy
McGuire

Laurence
McCullough

Richard
Street



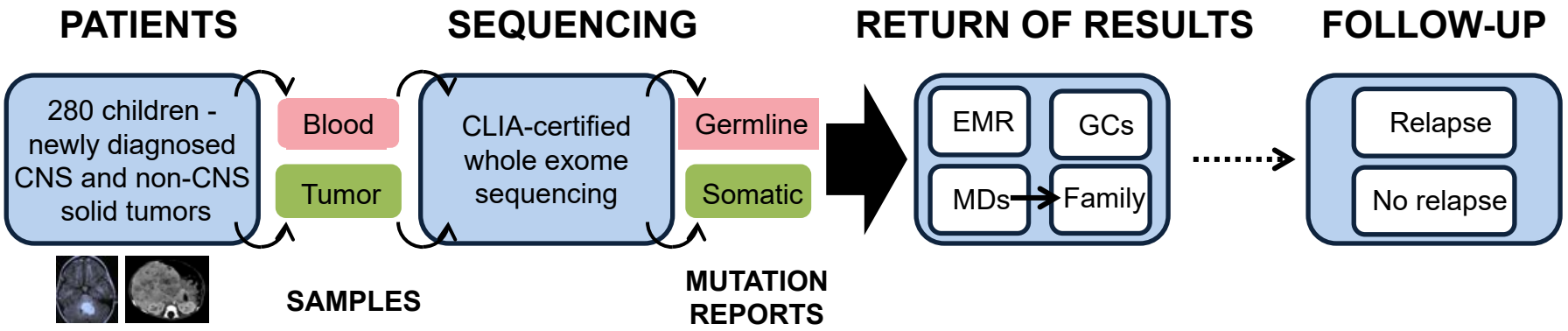
Sue
Hilsenbeck

Angshumoy
Roy

Dolores
Lopez-Terrada

BASIC^3

Baylor College of Medicine **A**dvancing **S**equencing **I**nto **C**hildhood **C**ancer **C**are



Study objectives:

- To integrate information from CLIA-certified germline and tumor exome sequencing into the care of newly diagnosed solid and brain tumor patients at Texas Children's Cancer Center
- To perform parallel evaluation of the impact of tumor and germline exomes on families and physicians

Will Parsons
Pediatric Oncology

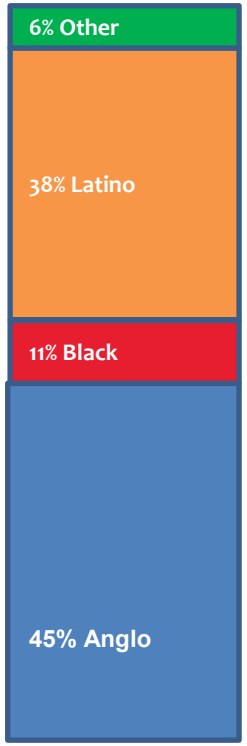
Original Investigation

Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors

D. Williams Parsons, MD, PhD; Angshumoy Roy, MD, PhD; Yaping Yang, PhD; Tao Wang, PhD; Sarah Scollon, MS, CGC; Katie Bergstrom, MS, CGC; Robin A. Kerstein, BS, MT; Stephanie Gutierrez, BS; Andrea K. Petersen, MD; Abhishek Bavle, MD; Frank Y. Lin, MD; Dolores H. López-Terrada, MD, PhD; Federico A. Monzon, MD; M. John Hicks, MD, PhD, DDS; Karen W. Eldin, MD; Norma M. Quintanilla, MD; Adekunle M. Adesina, MD, PhD; Carrie A. Mohila, MD, PhD; William Whitehead, MD; Andrew Jea, MD; Sanjeev A. Vasudevan, MD; Jed G. Nuchtern, MD; Uma Ramamurthy, PhD; Amy L. McGuire, JD, PhD; Susan G. Hilsenbeck, PhD; Jeffrey G. Reid, PhD; Donna M. Muzny, MSc; David A. Wheeler, PhD; Stacey L. Berg, MD; Murali M. Chintagumpala, MD; Christine M. Eng, MD; Richard A. Gibbs, PhD; Sharon E. Plon, MD, PhD

JAMA Oncol. doi:10.1001/jamaoncol.2015.5699
Published online January 28, 2016.

Race/Ethnicity of BASIC3 Subjects are Representative of Houston Population



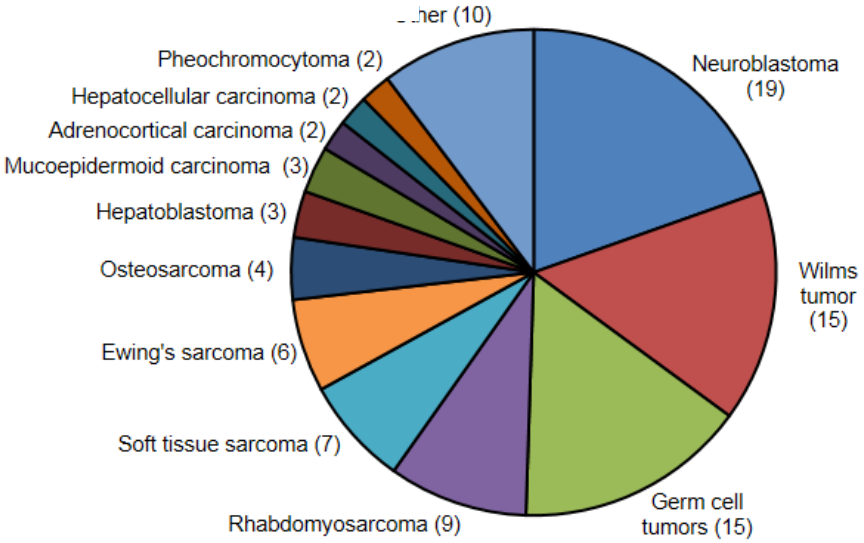
Characteristics of patients enrolled and not enrolled on study - updated

Characteristic - no. (%)	Enrolled (n=239)	Declined (n=103)	P Value
Ethnicity			0.54
Hispanic	111 (46%)	41 (40%)	
Non-Hispanic	119 (50%)	52 (50%)	
Not reported	10 (4%)	10 (10%)	
Race			0.11
White	141 (59%)	74 (72%)	
Black or African American	25 (10%)	12 (12%)	
Asian	7 (3%)	4 (4%)	
American Indian or Alaska Native	10 (4%)	2 (2%)	
Multiple	14 (6%)	---	
Not reported	42 (18%)		

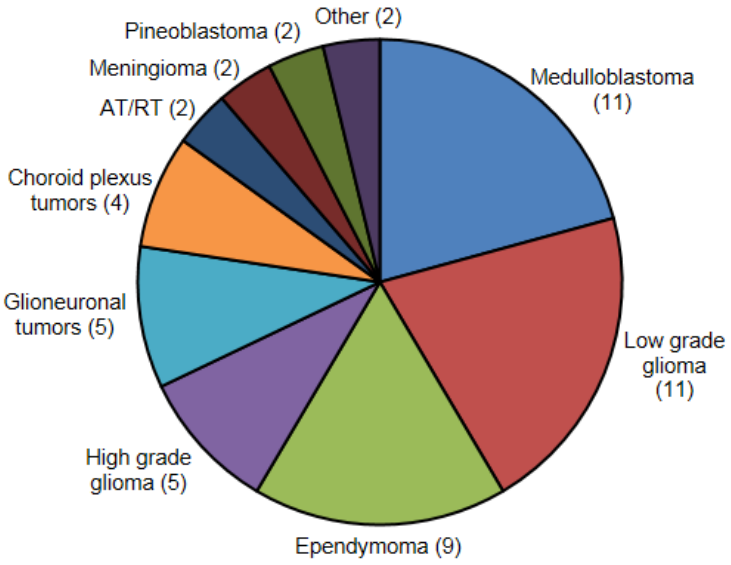
Updated from Scollon et al., *Genome Medicine* 2014

BASIC3 DIVERSE PEDIATRIC TUMOR DIAGNOSES

NON-CNS



CNS

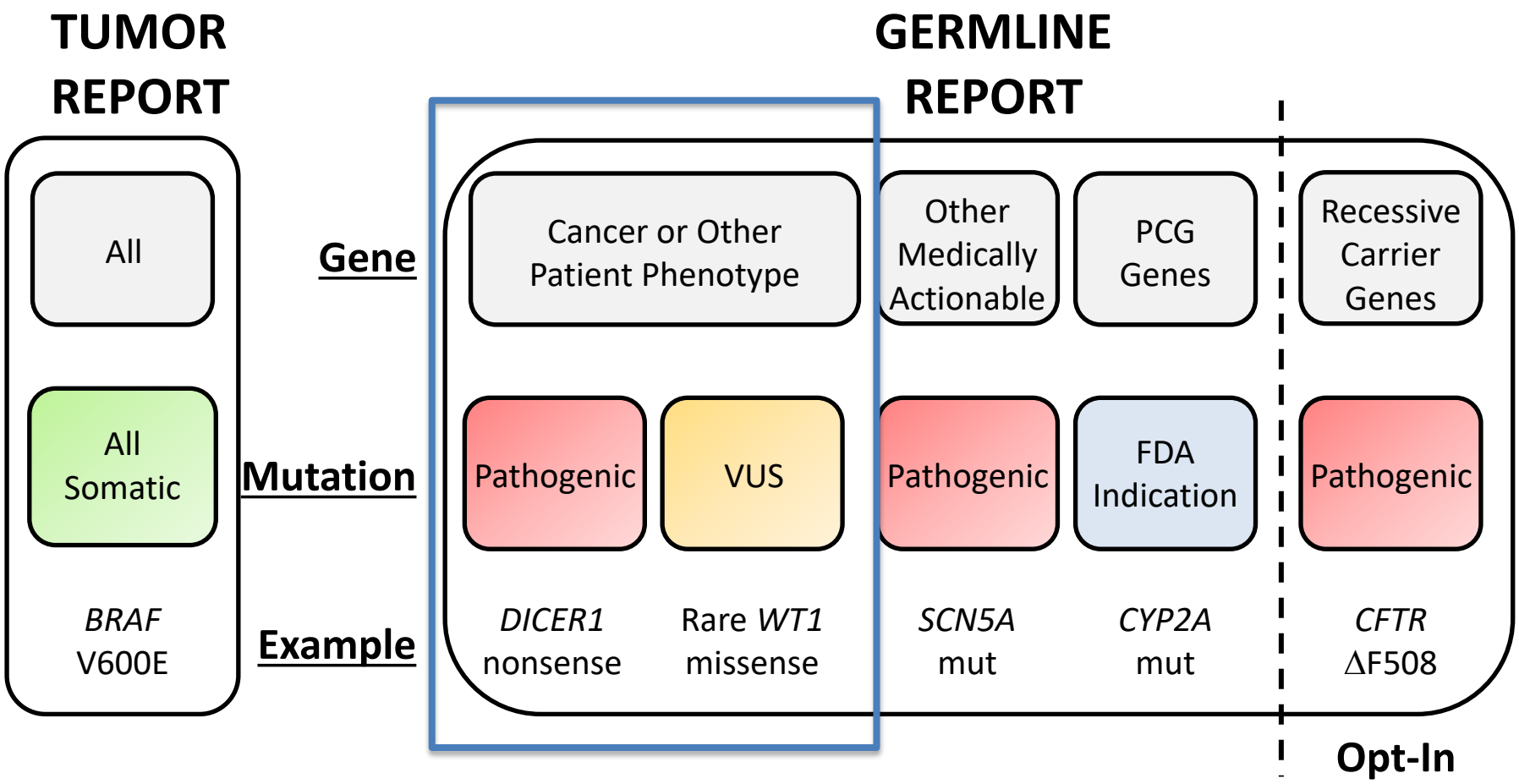


81/94 (86%)

40/56 (71%)

Tumor available for WES

Diversity of clinical exome results returned

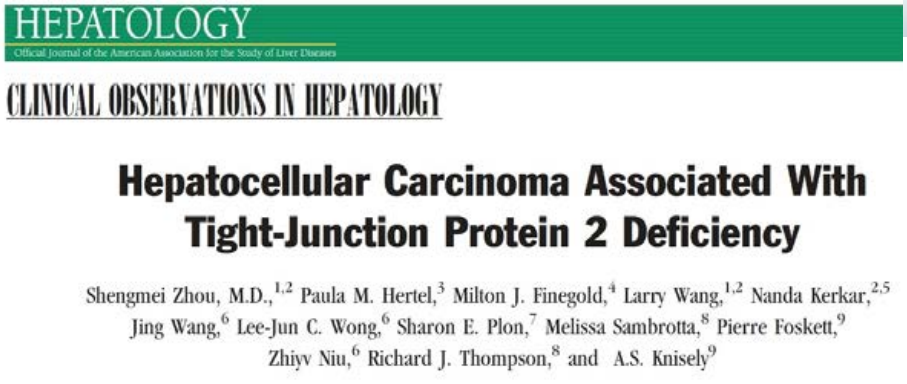


Cancer susceptibility molecular diagnosis in 9.8% (27/278) pediatric cancer patients

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

Rationale for WGS from Heterogeneous Dataset

- Several examples from BASIC3 exome data of identifying a rare variant in a single individual that was then enlarged to identify other rare cases outside the study resulting in new germline or somatic drivers.
  - *TJP2* deficiency and risk of hepatocellular carcinoma
  - Internal tandem duplication of *BCOR* as major somatic driver of clear cell sarcoma of the kidney



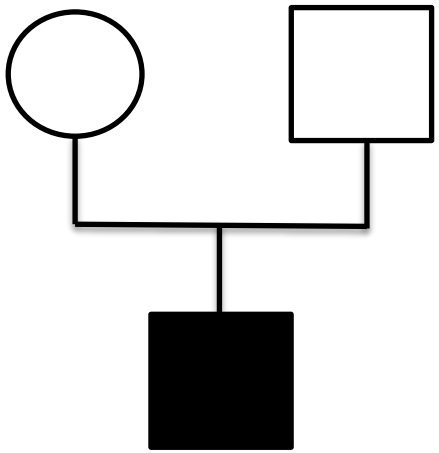
Goal of KidsFirst Project

- Utilize the rich existing dataset from BASIC3 to make further discoveries using whole genome sequencing of proband and parent trios.
- Existing data:
  - Germline exomes of probands only
  - Tumor exomes from probands
  - RNAseq for a subset of tumors
  - DNA from patient and parents isolated from blood samples
  - Lymphoblastoid cell lines available from almost all patients and parents for subsequent functional or splicing studies as needed.

Approach to KidsFirst Analysis

- Perform sequencing of patient/parent BASIC3 trios with following priority:
  - Complete trios with adequate DNA
  - Patients with unusual histories (multiple malignancies, birth defects)
  - Some solved cases to look for evidence of genomic instability
    - For example, number of de novo variants in child of *TP53* carrier
- Analysis has focused on both single nucleotide variants (SNV) and structural variants (SV)
  - Prioritized *de novo* variants in both situations
  - Also analyzed known cancer susceptibility genes for "missed" variants from exome analysis
  - Performed spliceAI analysis of rare variants for cryptic splicing variants

De novo SNV calling in BASIC3 WGS trios (n=54 complete trios)



Platypus

Published: 13 July 2014

Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications

Andy Rimmer[1 na1], Hang Phan[1 na1], Iain Mathieson[1], Zamin Iqbal[1], Stephen R F Twigg[2], WGS500 Consortium, Andrew O M Wilkie[2], Gil McVean[1,3 na1] & Gerton Lunter[1]

Nature Genetics **46**, 912–918(2014) | Cite this article

Prioritization:

Prioritization of *de novo* variants:

De novo coding variants

↓

Determine functional effect

↓

Rank based on pathogenicity calculations

↓

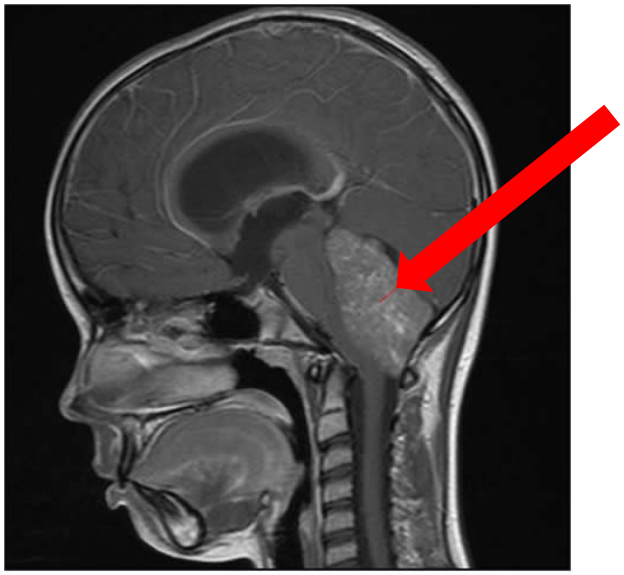
Literature and database queries

Outcome:

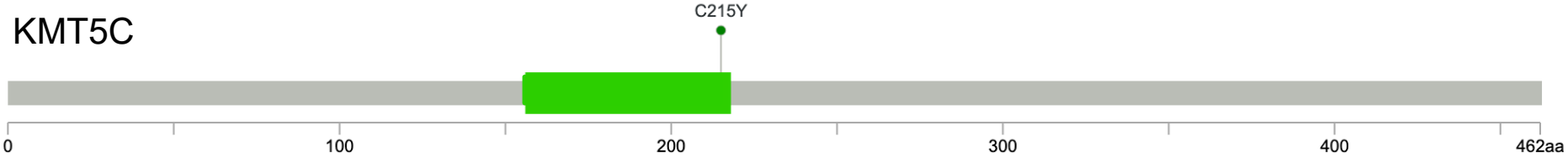
- SNV analysis completed on 54 proband-parent trios
- The pipeline resulted in an expected number of variants per trio

Variant Type	Frequency per trio
Genome-wide *de novo*	60 to 190
Coding *de novo*	0 to 4

De novo heterozygous missense variant in *KMT5C* in pediatric ependymoma



- Mutation is rare, occurs in highly conserved functional domain, and has a high CADD (28.3) and REVEL (.662) score

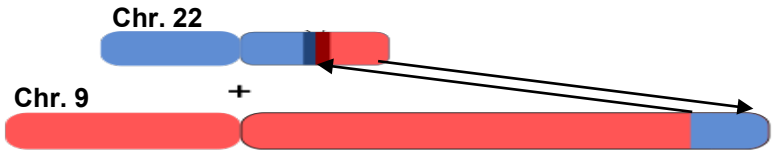


STRUCTURAL VARIANT CALLING IN BASIC3/KIDSFIRST WGS DATA

Owen Hirschi
Graduate Student
Genetics and Genomics Program
Baylor College of Medicine

Identification of germline structural variation in childhood cancer is limited

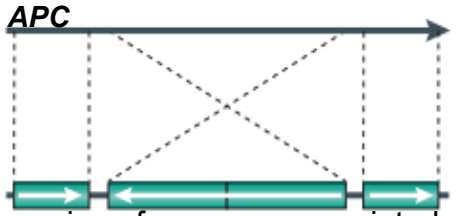
Somatic structural variants play a critical role in cancers



Ex. Philadelphia chromosome translocation in leukemia

Germline structural variants are implicated in pediatric cancers

Discovered through RNA-seq



Ex. Inversion of a cancer-associated gene in a child with liver cancer

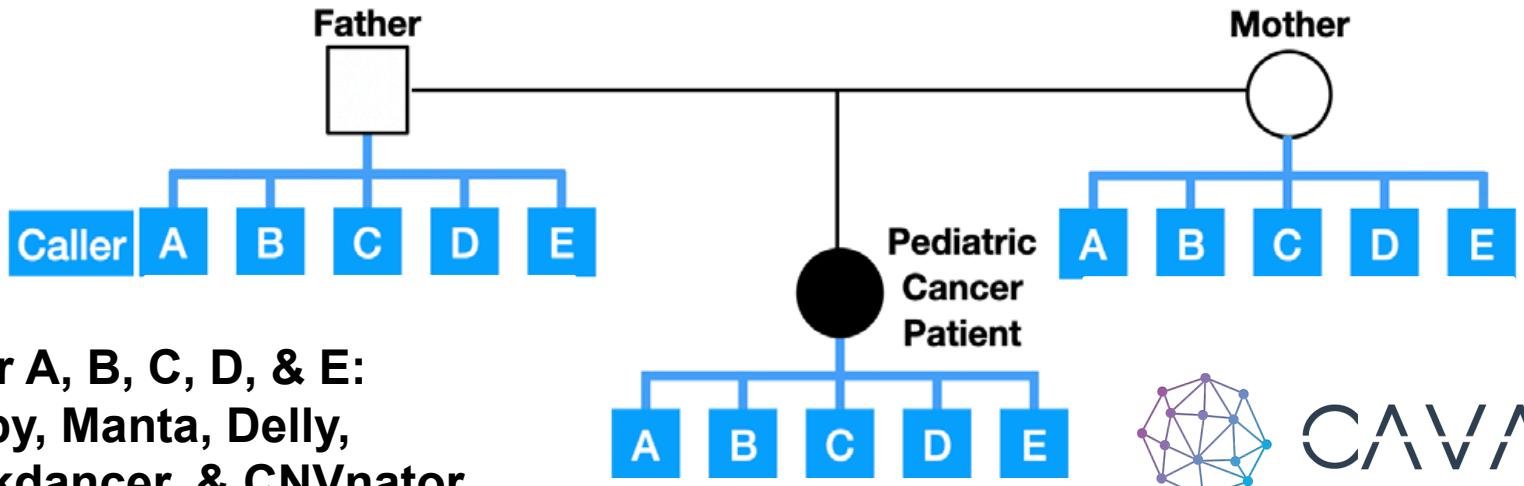
Haines K, et al. 2019

Since 2010 **over 40** programs have been created to identify SV from short-read WGS - primarily built for identification in tumors

- Sensitivity ranges from **10-70**% and false positive rate **up to 89**%

Alkan C, et al. 2011
Mahmoud M, et al. 2019

Developed new pipeline for more effective germline structural variant calling



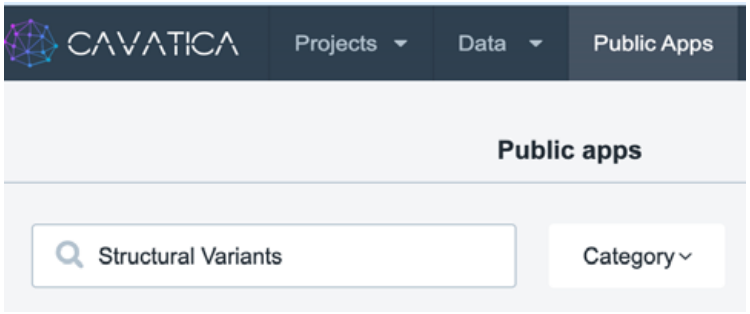
Caller A, B, C, D, & E: Lumpy, Manta, Delly, Breakdancer, & CNVnator

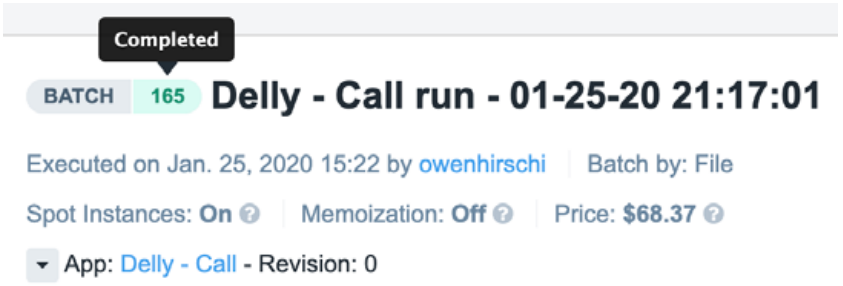
CAVATICA

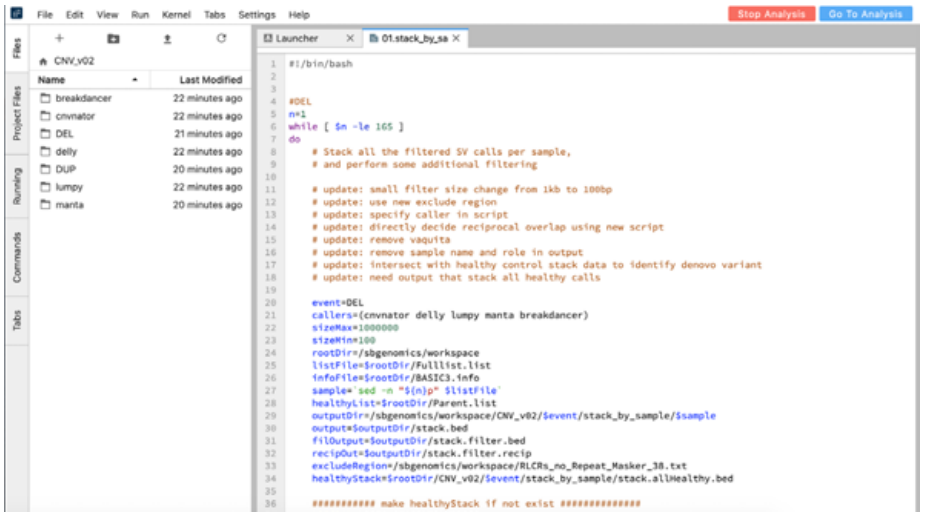
Using multiple structural variant callers allowed for more specificity and sensitivity in variant identification

The pipeline is built for identification of *de novo* structural variants and is being tuned for inherited variants

Analysis of structural variant on Cavatica requires multiple features of the platform

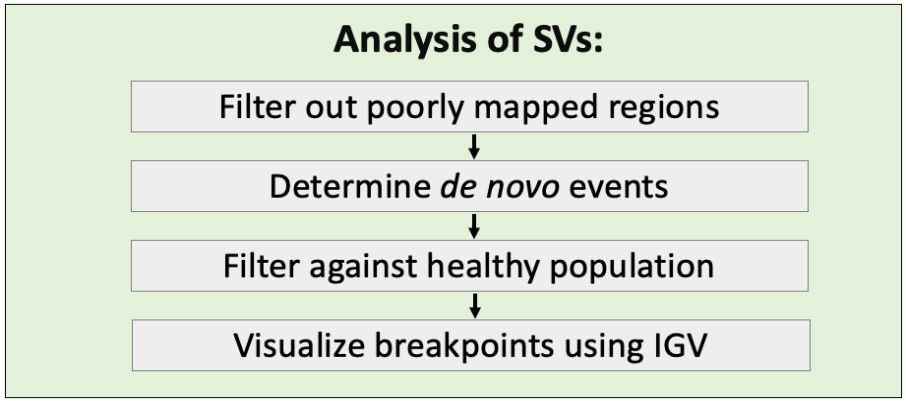






De novo structural variant calling in BASIC3 WGS trios (n=54 trios)

Analysis:



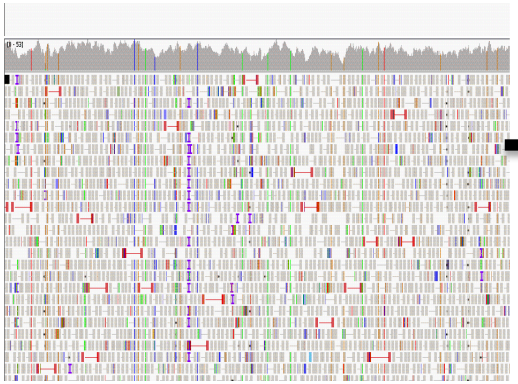
Outcome:

- SV analysis completed on 54 proband-parent trios
- The pipeline identified between 0 to 1 SVs per trio

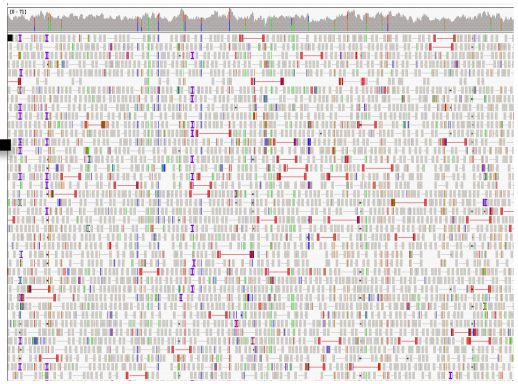
Structural Variant Type	# of SVs identified
Deletions	10
Duplications	7
Inversions	0

Germline *de novo* heterozygous 20kbp deletion of *NF2* in a meningioma patient

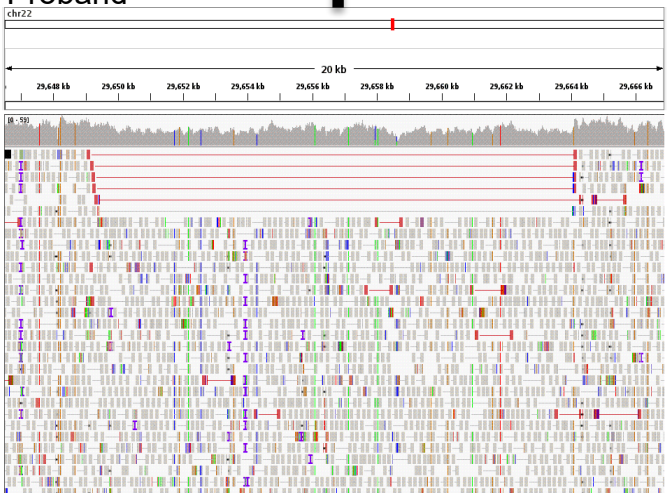
Mother



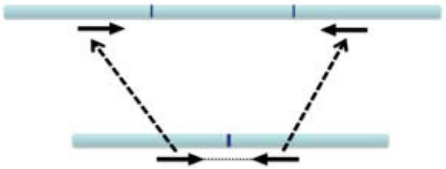
Father



Proband



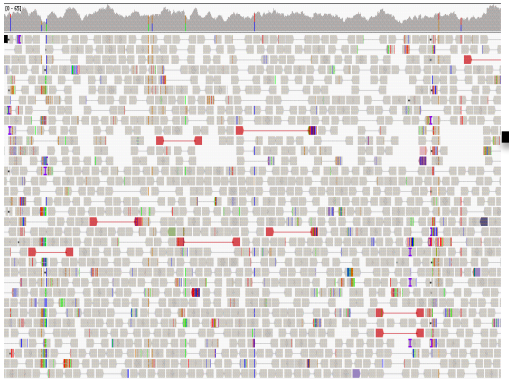
NF2 exon:5-9



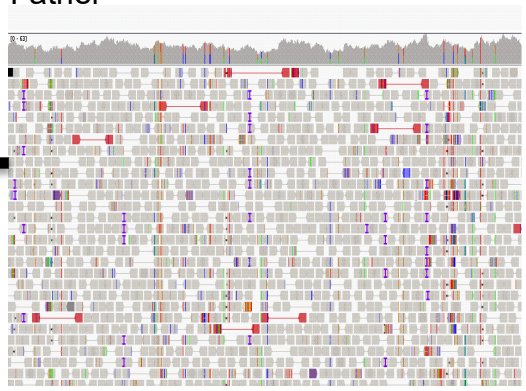
Potentially Somatic
Mosaic
6/40 reads

Germline *de novo* heterozygous 8kbp duplication of *PTGR2* in an ependymoma patient

Mother

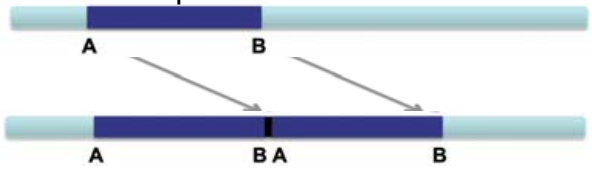


Father



Proband

PTGR2 promoter and exon 1



Head-to-tail insertion

PTGR2 duplication includes 5 predicted promoter-like regions

	DNase max-Z	H3K4me3 max-Z	H3K27ac max-Z	CTCF max-Z	chr	start	length	nearest genes: protein-coding / all	predicted function
P	5.50	4.63	4.48	3.11	chr14	73,851,666	347	***PTGR2***	**Promoter**
P	5.15	4.75	4.64	2.76	chr14	73,852,048	168	***PTGR2***	**Promoter**
P	4.37	4.38	4.35	1.67	chr14	73,852,538	348	***PTGR2***	**Promoter**
P	4.31	3.86	4.25	1.81	chr14	73,850,993	238	***PTGR2***	**Promoter**
P	3.18	1.96	2.63	1.29	chr14	73,850,000	218	***PTGR2***	**Promoter**

High H3K4me3
High H3K27ac
High CTCF
High DNase
Z-score < 1.64
P/D Proximal/Distal to a Transcription Start Site

Whole genome sequencing revealed novel germline structural variants

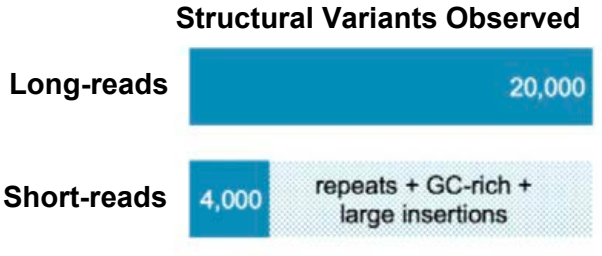
We have developed a pipeline with the aim to better identify germline *de novo* structural variation in pediatric cancer

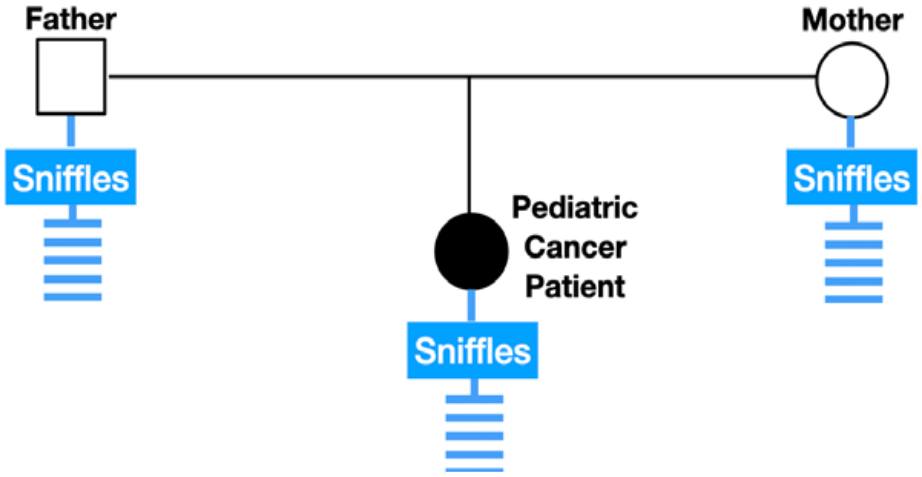
We have identified a germline *de novo* heterozygous duplication in a pediatric ependymoma that we are beginning to functionally assess

We have benefited from additional research studies involving both tumor and germline sequencing of BASIC3

Long-read sequencing of BASIC3 through Kids First

Long-read sequencing allows for greater detection of SV





Goals of this sequencing study:

- We will compare the SV results of the long-read sequencing to those of the short-read sequencing using our pipeline

Enabling us to identify:

- Limitations and accuracy of short-read structural variant in trio studies
- Causative SVs only found via long-read sequencing

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

Merker JD, et al. 2017

QUESTIONS?

Gabriella Miller Kids First Data Resource Center

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

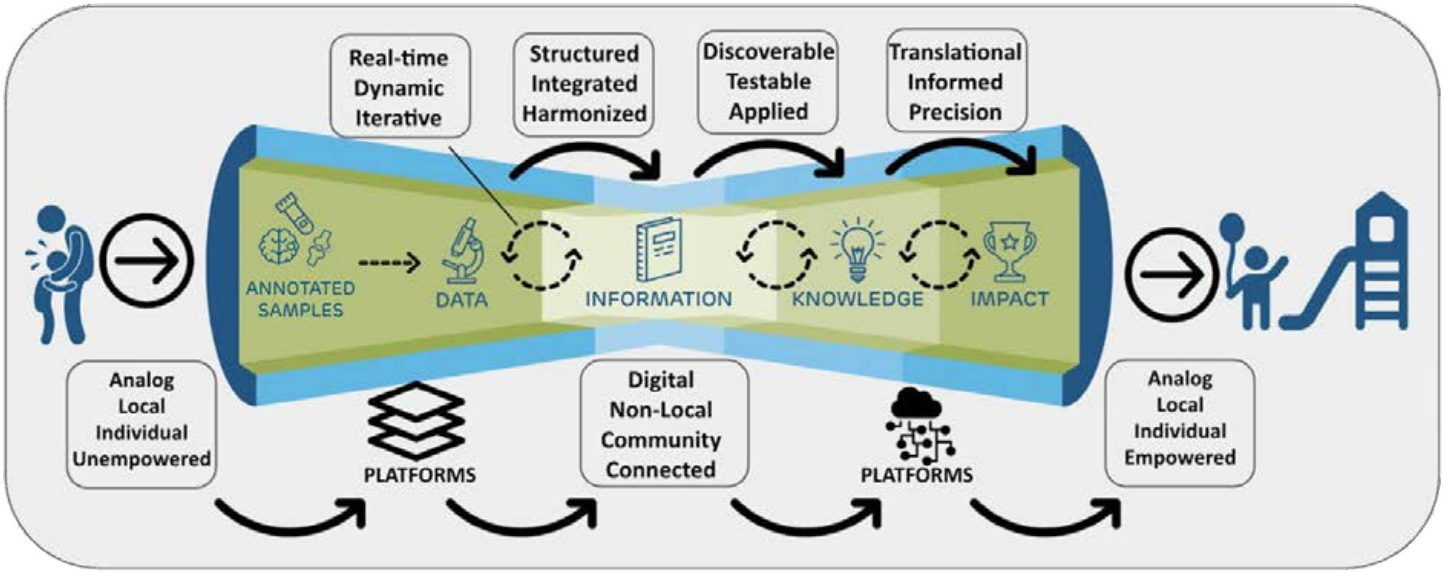
Adam Resnick, PhD
Children's Hospital of Philadelphia
Principal Investigator,
Gabriella Miller Kids First
Data Resource Center

Vincent Ferretti, PhD
Sainte-Justine University Hospital
Principal Investigator,
Gabriella Miller Kids First
Data Resource Portal

KFDRC Portal Update

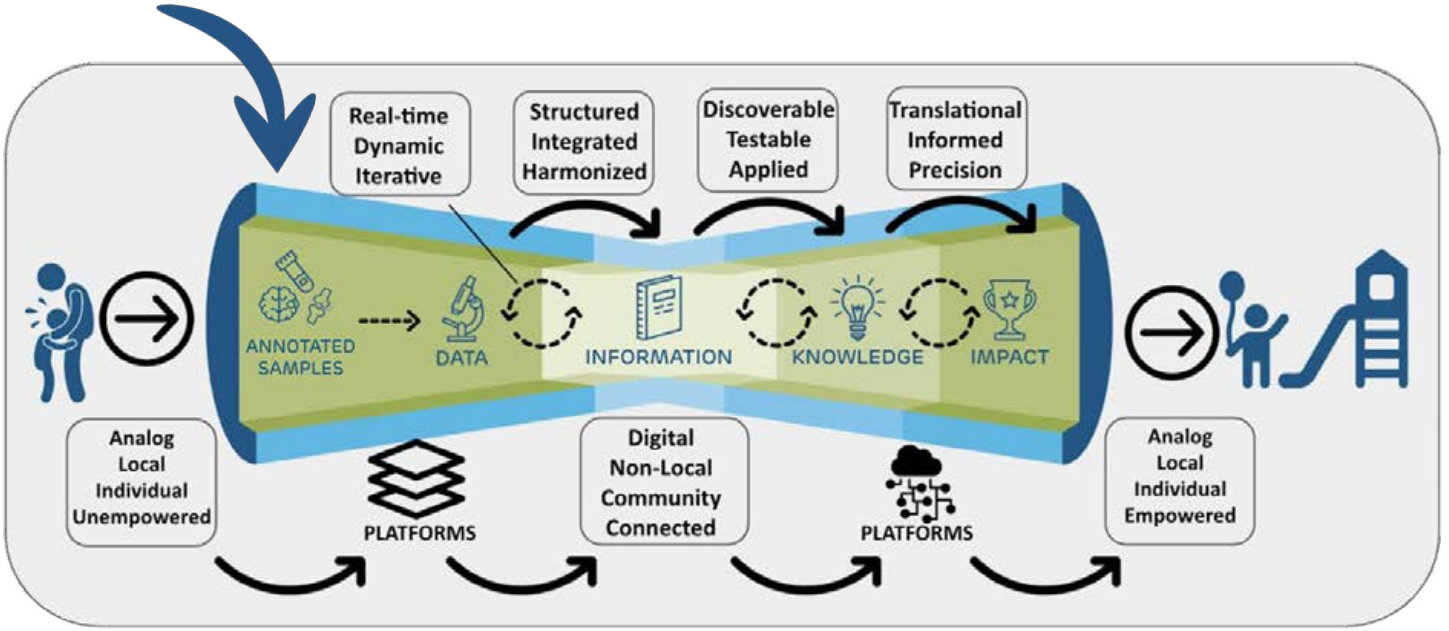
Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

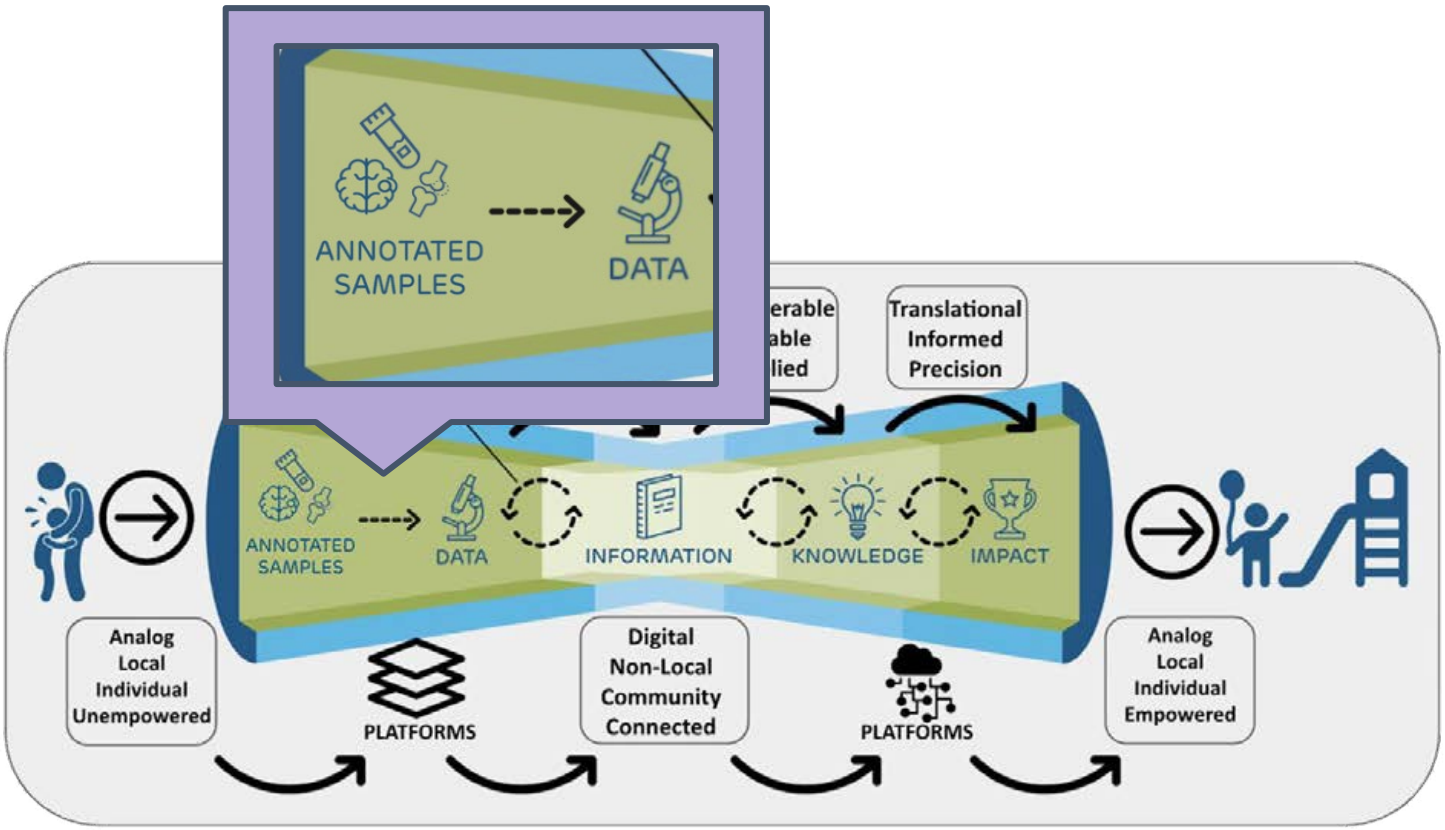
Methodology



Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Double "funnel" model




ANNOTATED SAMPLES
DATA
erable
able
lied
Translational
Informed
Precision
ANNOTATED SAMPLES
DATA
INFORMATION
KNOWLEDGE
IMPACT
Analog
Local
Individual
Unempowered
PLATFORMS
Digital
Non-Local
Community
Connected
PLATFORMS
Analog
Local
Individual
Empowered

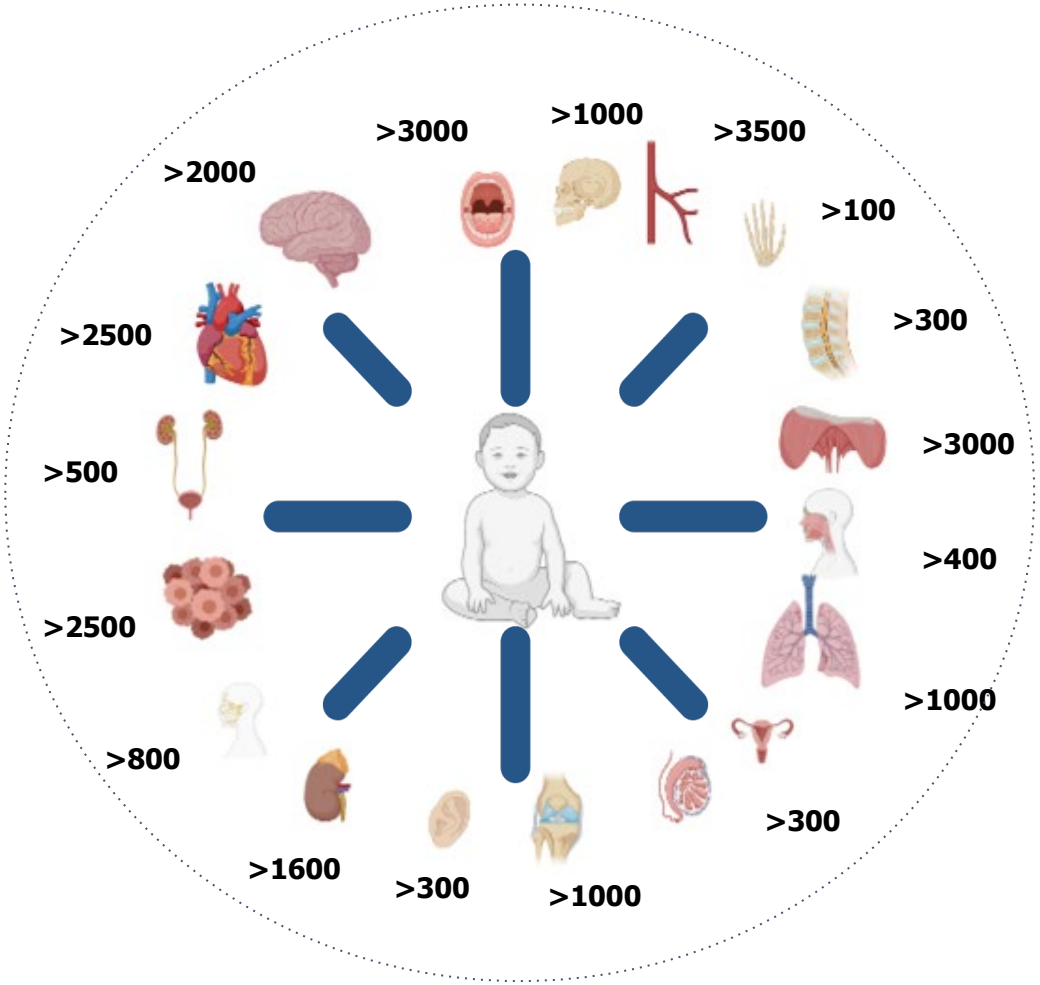
Released Studies

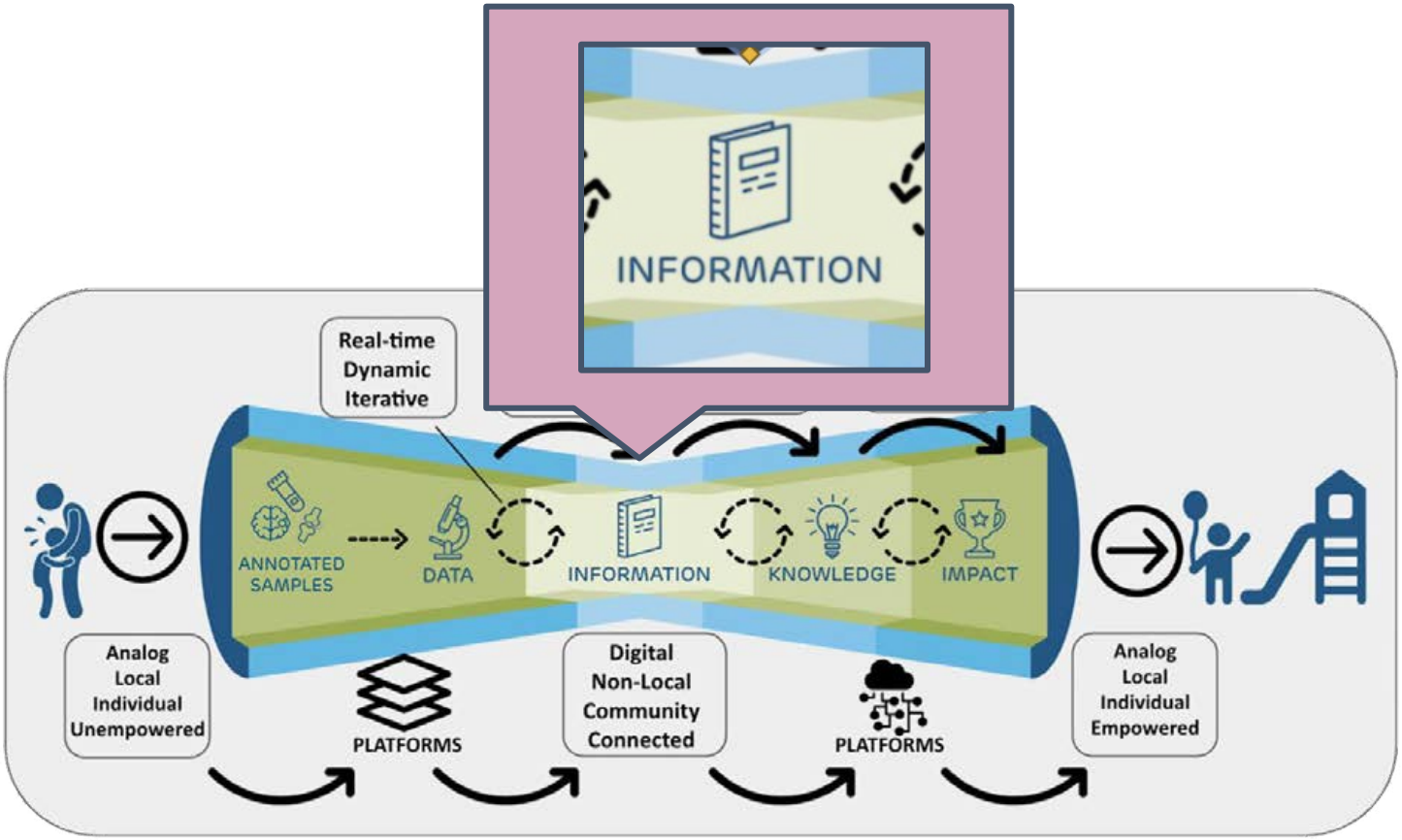
Study Name	Funding Year	Category	Participants	Trios	Other Families
Neuroblastoma	FY17	Cancer	1612	463	121
Congenital Diaphragmatic Hernia	FY15 & FY16 & FY17	Structural Birth Defects	2245	714	60
Congenital Heart Defects (PCGC)	FY15 & FY16	Structural Birth Defects	2133	711	n/a
Orofacial Clefts - European Ancestry	FY15	Structural Birth Defects	1295	380	90
Orofiacial Clefts - Asian & African Ancestry	FY17	Structural Birth Defects	725	238	6
Novel Cancer Susceptability (from BASIC3)	FY16	Cancer	291	66	48

Released Studies (Cont'd)

Study Name	Funding Year	Category	Participants	Trios	Other Families
Ewing Sarcoma	FY15	Cancer	1047	290	153
Syndromic Cranial Dysinnervation	FY15	Structural Birth Defects	801	248	n/a
Orofacial Cleft - Latin American Ancestry	FY16	Structural Birth Defects	804	262	9
Adolescent Idiopathic Scoliosis	FY16	Structural Birth Defects	262	65	7
Familial Leukemia	FY16	Cancer	365	53	3
Enchondromatoses	FY17	Structural Birth Defects	79	25	2
Disorders of Sex Development	FY15	Structural Birth Defects	300	91	3
Osteosarcoma	FY15	Cancer	84	proband-only study	n/a

Approaching 30,000 samples released and in the que




INFORMATION
Real-time
Dynamic
Iterative
ANNOTATED
SAMPLES
DATA
INFORMATION
KNOWLEDGE
IMPACT
Analog
Local
Individual
Unempowered
PLATFORMS
Digital
Non-Local
Community
Connected
PLATFORMS
Analog
Local
Individual
Empowered

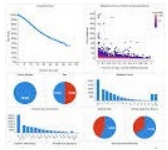
The Kids First Data Resource for Collaborative Discovery

Data Resource Portal

Entry point. Query, search, discover, build & visualize synthetic cohorts

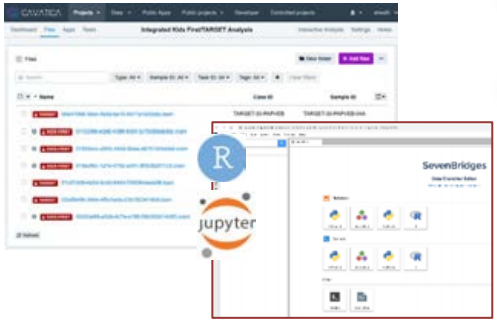
Knowledge Base Integrations (PedcBioPortal)

Integrations with existing curated/published data visualizations



Cavatica

Pull data from multiple sources into one workspace. Use notebooks, bring-your-own or use available workflows.



Data Services

Model clinical data in FHIR-based data services for semantic interoperability and coordination

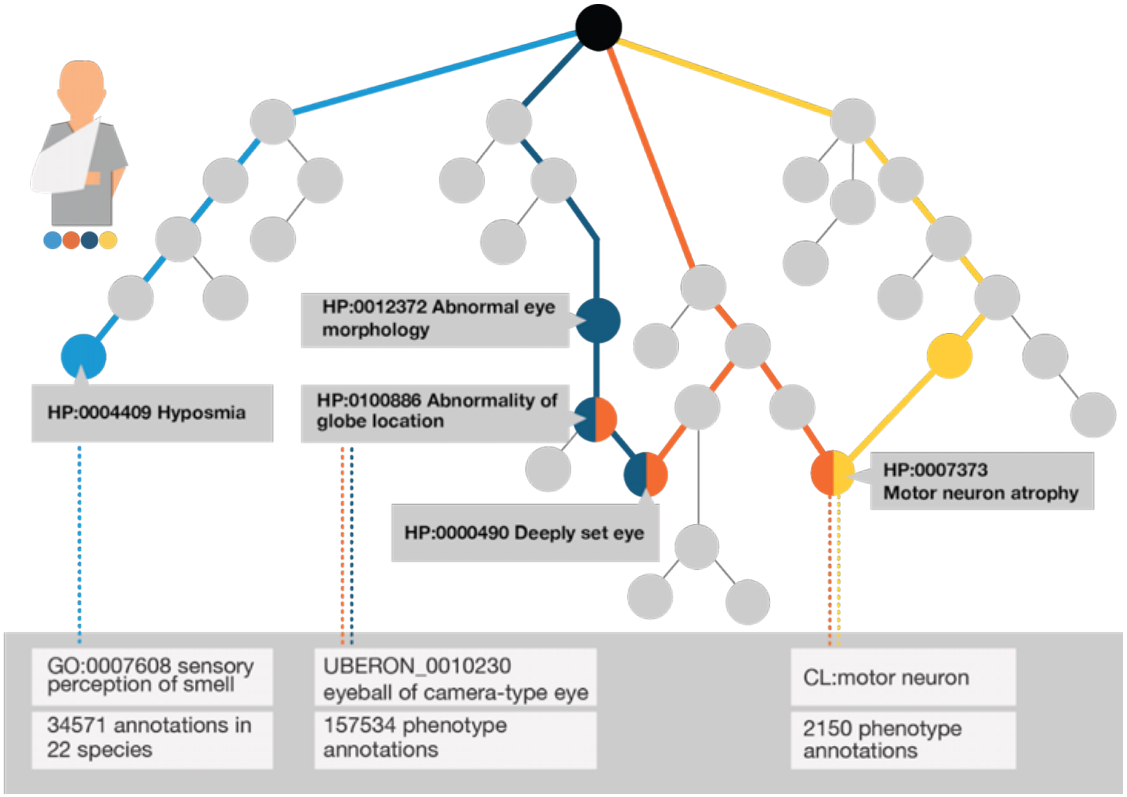
Framework Services

Index and point to files in the cloud (for approved users)

aws

STRIDES

Ontologies Enable Computable Models



human phenotype ontology

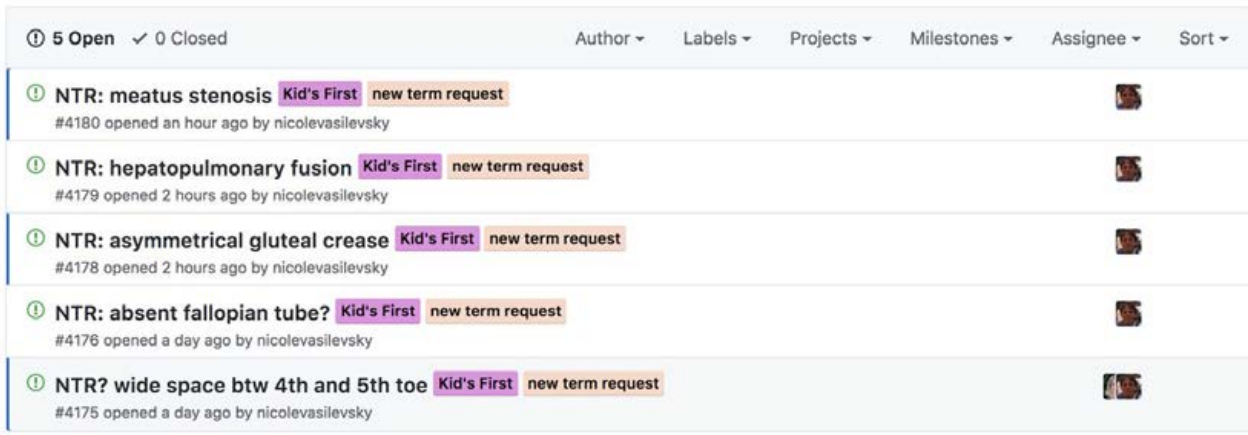
monarch INITIATIVE

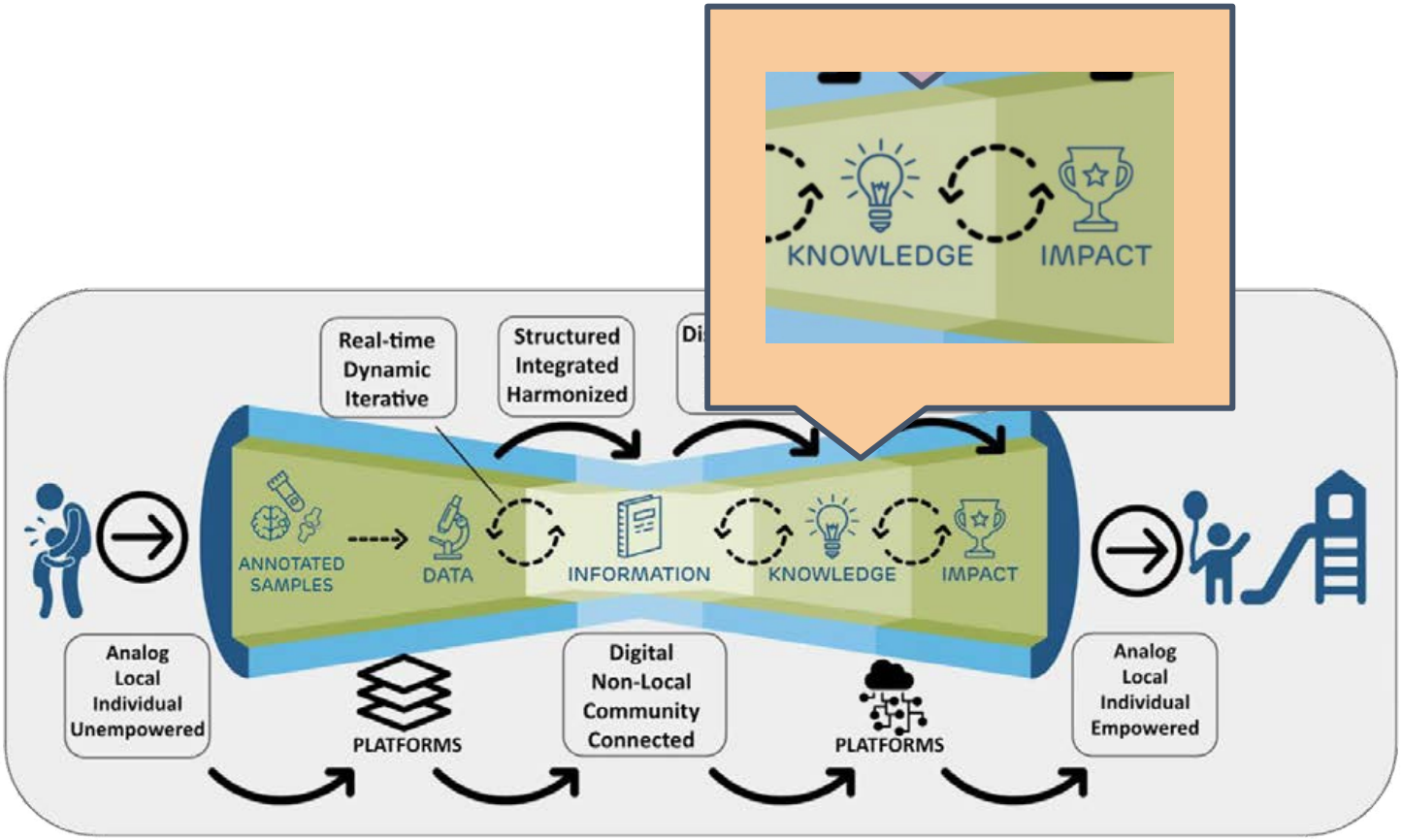
hpo.jax.org

Curation of Kids First data sets

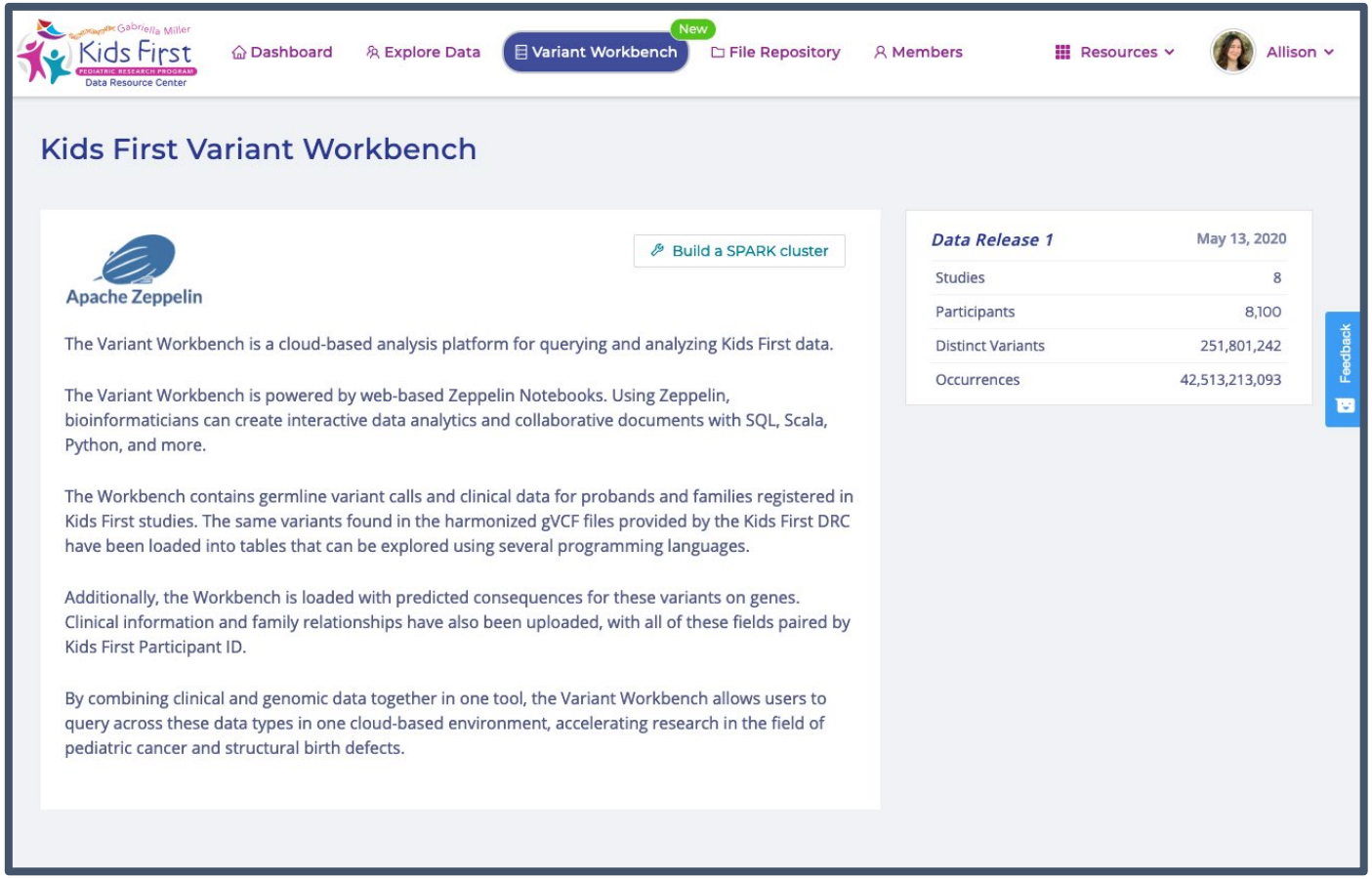
2	source_text	hpo_id	hpo_official_text	HPO_ID	HPO_Label	HPO_Modifier_ID	HPO_Modifier_Label
50	Anteriorly Placed Anus	HP:0001545	Anteriorly placed anus				
51	Asplenia	HP:0001746	Asplenia				
52	Assymetric crying facies	HP:0011333	Asymmetric crying face				
53	Asthma	HP:0002099	Asthma				
54	Asymmetric mouth movement; hypoplastic left jaw	HP:0009118	Aplasia/Hypoplasia of the mandible	HP_0000347	Micrognathia	HP_0012835	Left
55	Atresia of left external ear	HP:0000413	Atresia of the external auditory canal			HP_0012835	Left
56	Atresia of the large intestine	HP:0010448	Colonic atresia				
57	Atrial Septal Defect	HP:0001631	Atrial septal defect				
58	Autistic spectrum disorder	HP:0000729	Autistic behavior				
59	Behavior/Mood disorder	None	None		Behavioral		
60	Belpharoptosis	HP:0000508	Ptosis				
61	Bicornuate Uterus	HP:0000813	Bicornuate uterus				

Engage with community for additions and/or recommendations



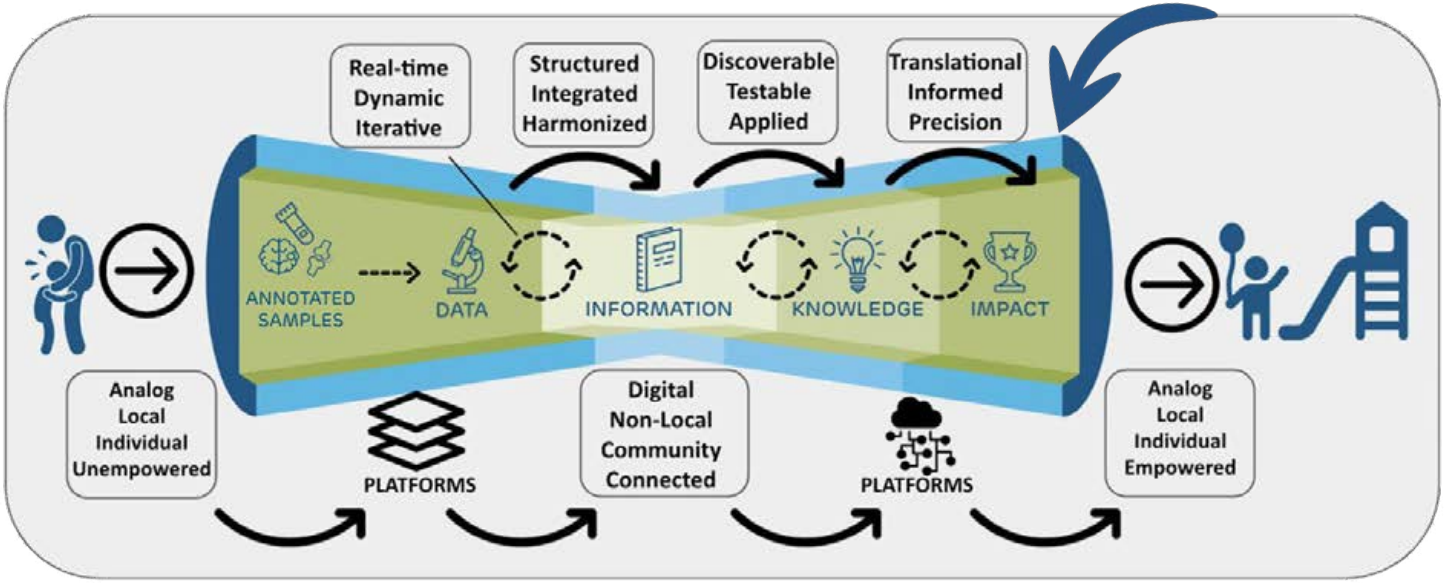

KNOWLEDGE
IMPACT
Real-time
Dynamic
Iterative
Structured
Integrated
Harmonized
ANNOTATED
SAMPLES
DATA
INFORMATION
KNOWLEDGE
IMPACT
Analog
Local
Individual
Unempowered
PLATFORMS
Digital
Non-Local
Community
Connected
PLATFORMS
Analog
Local
Individual
Empowered

Variant Workbench



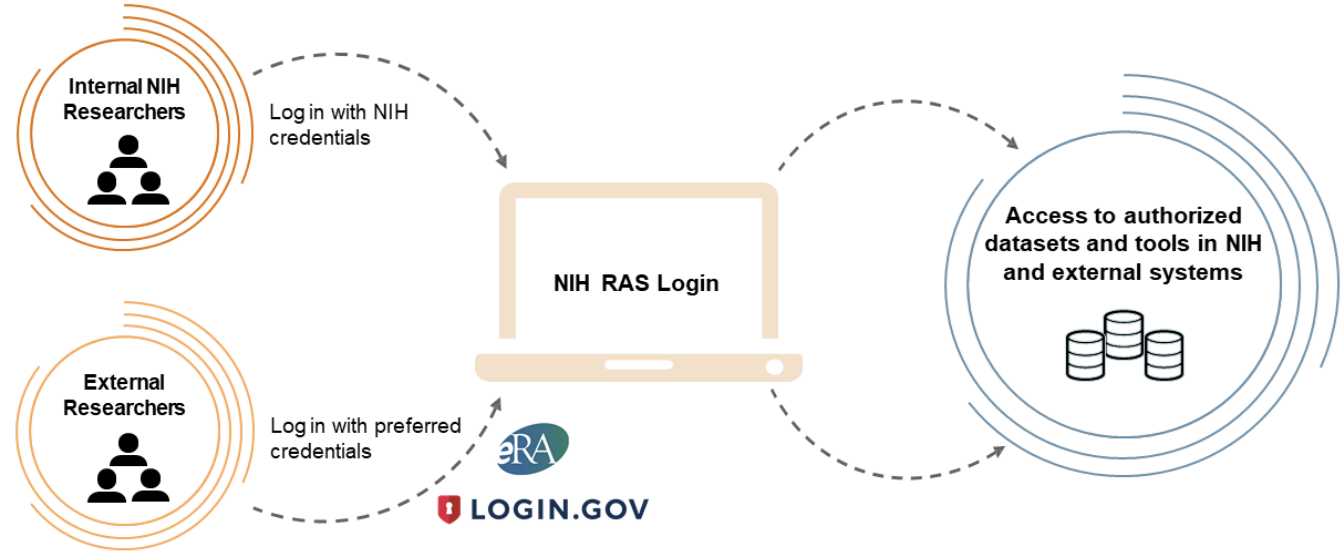
Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

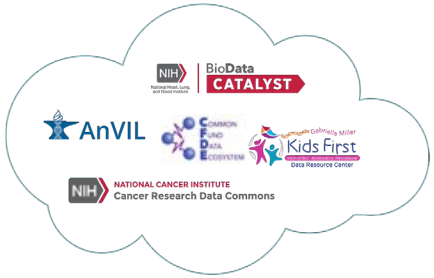
Double "funnel" model



Interoperability Across NIH Resources

NIH RAS is a unified, efficient, and secure authentication and authorization service deploying in late 2020 provided by NIH CIT that enables streamlined researcher access to NIH-funded data assets and data repositories across multiple systems and provides standardized methods of logging and auditing such access.





GEN3 DATA COMMONS

Framework Services

Index and point to files in the cloud (for approved users)

Kids First DRC Team

Children's Hospital of Philadelphia

SevenBridges

CHU Sainte-Justine Research Center
Mother and Child University Hospital Center

THE UNIVERSITY OF CHICAGO
CENTER FOR DATA INTENSIVE SCIENCE

Children's National Health System

OREGON HEALTH & SCIENCE UNIVERSITY
OHSU

Kids First DRC Team

Children's Hospital of Philadelphia

SevenBridges

CHU Sainte-Justine Research Center
Mother and Child University Hospital Center

THE UNIVERSITY OF CHICAGO
CENTER FOR DATA INTENSIVE SCIENCE

Children's National Health System

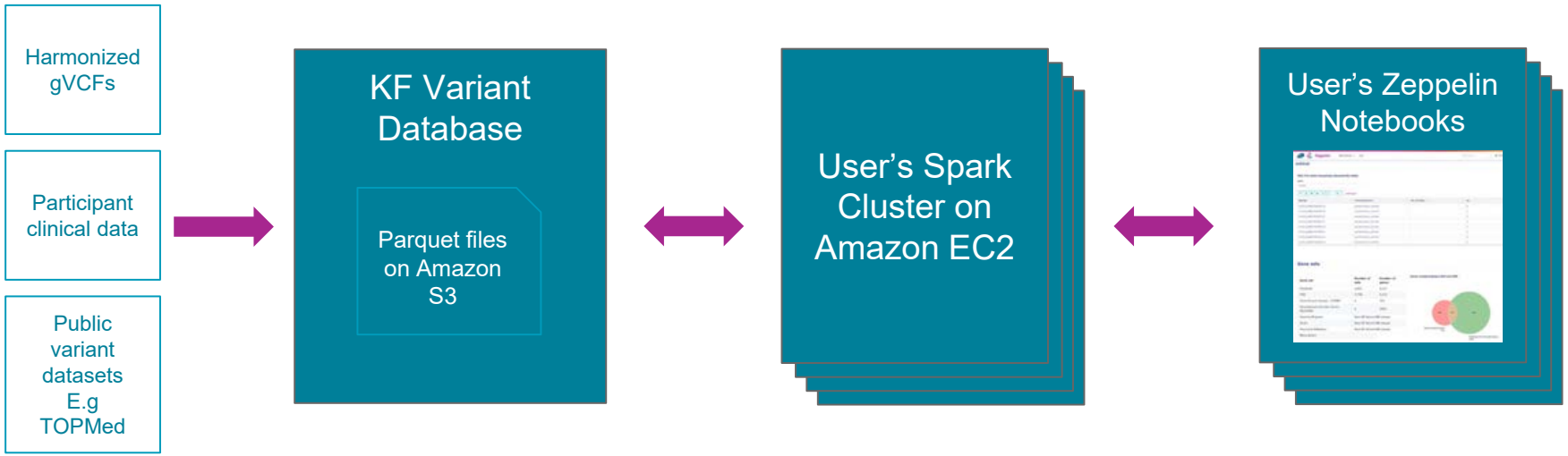
OREGON HEALTH & SCIENCE UNIVERSITY
OHSU

The KF Germline Variant Workbench Project

- Prototype presented at the last webinar in May
- Productionizing software and infrastructure
  - Security, data quality, robustness, bioinformatics, cluster management and software deployment tools
- Objective
  - Extract, annotate and import variant data from harmonized gVCF files to a *queryable* database
- Challenges
  - Complex data, big data, confidentiality and security
- Production version to be released
  - Phase I: X01 investigator's teams, Oct-Nov 2020
  - Phase II: All Kids First Portal members, Jan 2021

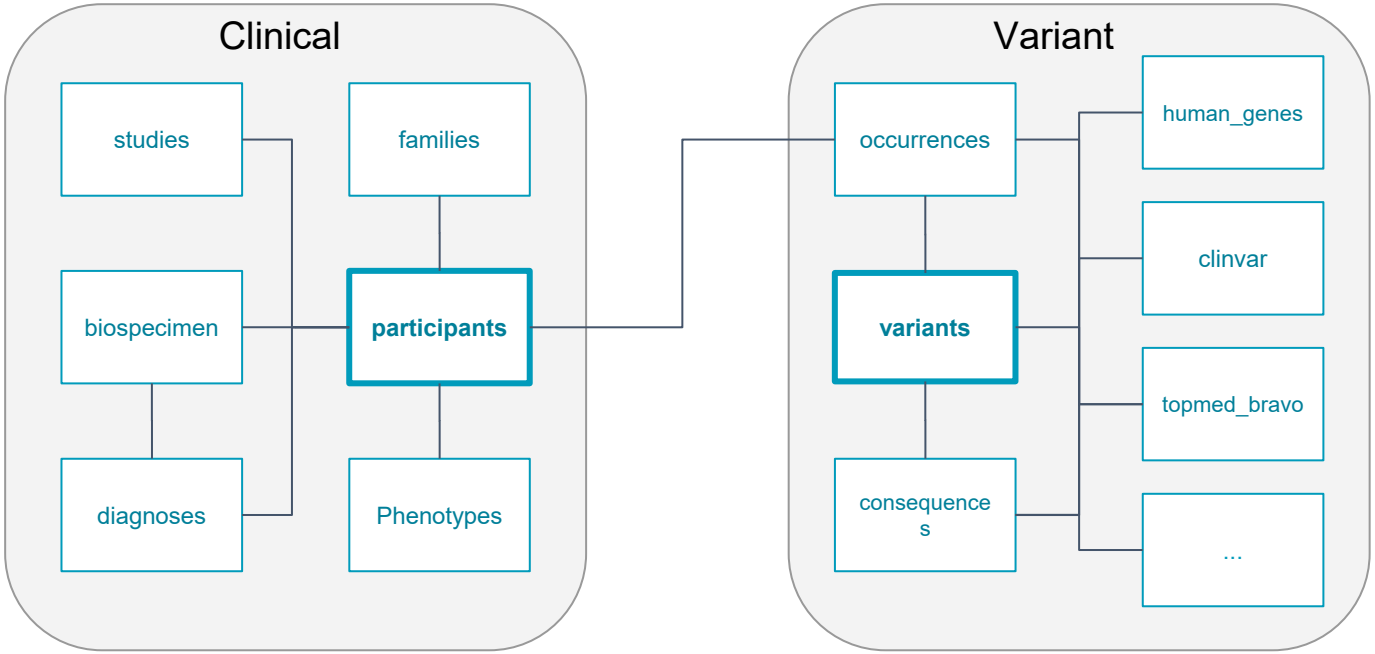
Two main components

- An annotated variant database
- User private compute environment with Spark and Zeppelin



Variant Database

- Data is organized into tables like in a traditional relational database
- 31 tables in current version



Current Annotations

- From VEP and dbsnfp
  - AA change, genes, transcripts,...
  - > 360 different annotations (e.g. predictive functional scores, gtex)
- Gene ids and alias from the NCBI
- Clinvar, dbSnp
- Gene panels
  - Cosmic, HPO, Omim, Orphanet
- Populational databases
  - TOPMed, gnomAD 2.1 and 3.0, 1000 genomes
- KF annotations
  - Allele frequencies, parent's genotypes, zygosity, QC metrics

Example: Find all **de-novo rare missense** variants with **high functional impact** in **low grade glioma** patients affected by any **cardiovascular abnormalities**

Data Currently Loaded for Initial Beta Phase

11 studies

11 thousand participants

297 million unique variants

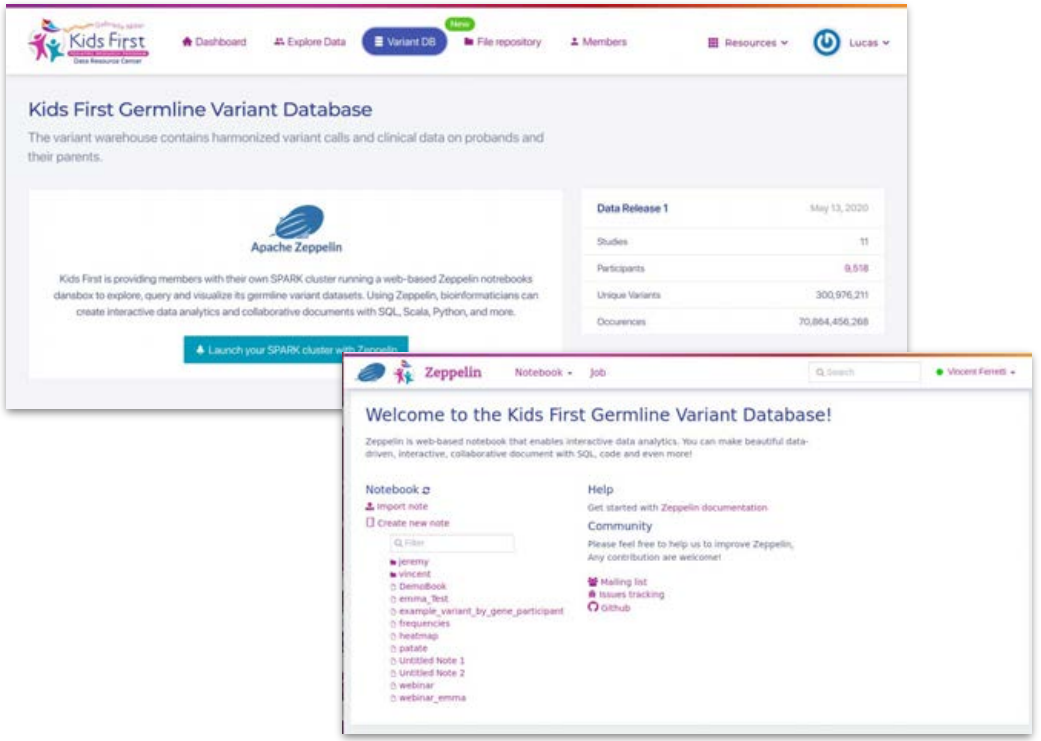
53 billion occurrences

- Each variant occurs on average in 178 participants
- Each participant has on average 4.8 million variants
- 52% (153M) of the KF variants are in TOPMed and/or gnomAD

The Zeppelin Data Analytic Environment

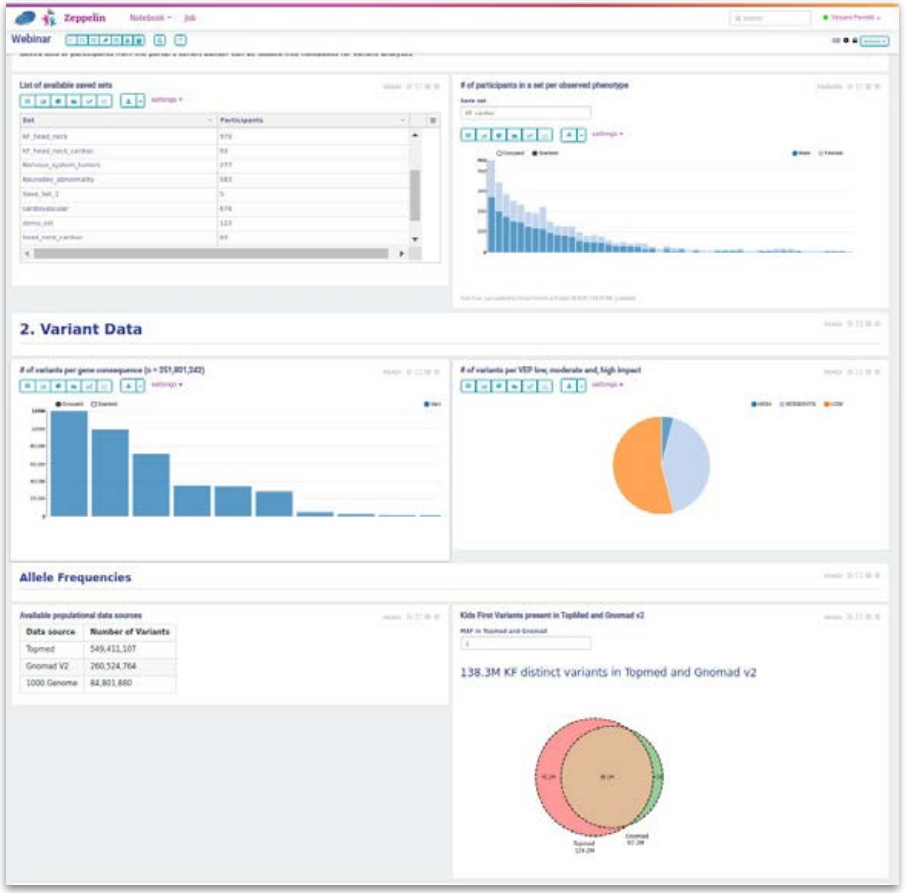
Provides programmatic access to the variant database from web browsers

- Accessible from the Portal
- Powered by individual Spark clusters on AWS
- User notebook workspace for creating and managing notebooks
- Users can only access variant datasets they have been authorized to



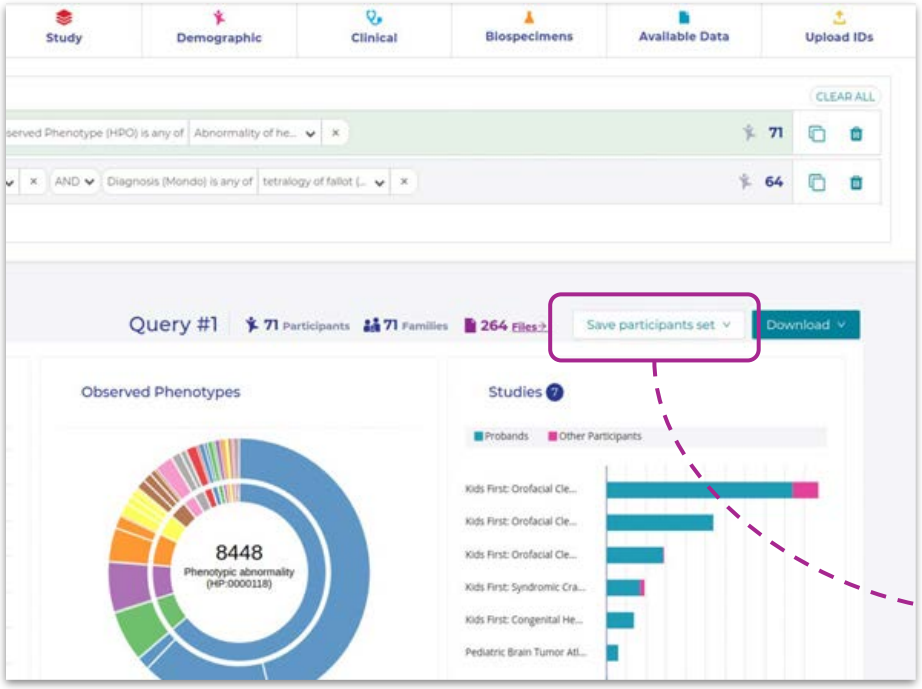
Zeppelin Notebooks

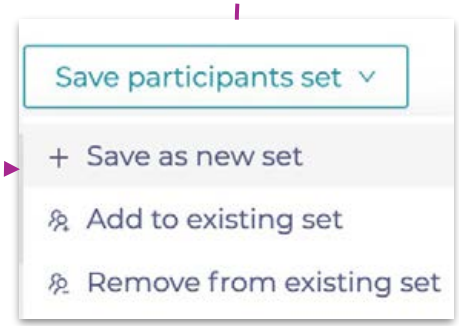
- Support for various programmatic languages (SQL, Python, R, Scala, Markdown)
- Can use different languages in the same Notebook
- Built-in chart plotting capability
- Interactive forms
- Multi-user capability
- Pre-installed libraries
  - **Hail**, Pandas, Numpy, Matplotlib, Plotly, Seaborn



Cohort builder & variant database integration

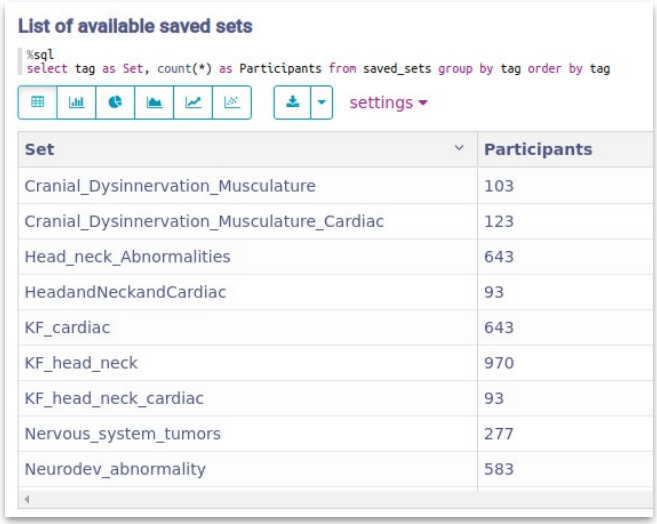
Virtual cohorts built using the Portal cohort builder can be saved as *Participant sets*





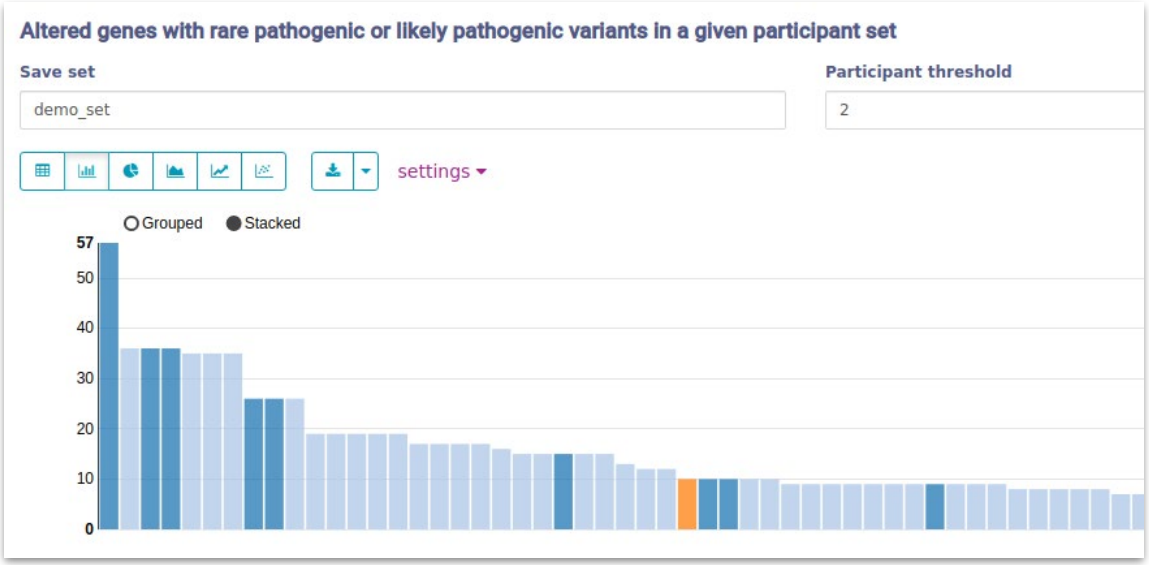
User dashboard

User's saved sets as part of the variant database



Variant workbench

Saved sets can be joined with other variant tables for genomic analyses of virtual cohorts



Who is the variant workbench for?

- Initially:
  - KF DRC for data quality control
  - Bioinformaticians who know programming in SQL, Python, R, etc.

- **Next Aim:** Increase the user base by offering a series of interactive and comprehensive notebooks directly accessible from the portal
  - e.g. a gene-centric analysis notebook
- Users will be able to
  - input analysis parameters (Zeppelin forms)
  - import their own data to their notebooks (e.g. gene panels, variant datasets)
  - modify notebooks as they need and save them in their workspace
- A good tradeoff between standard data portal static analyses and the complexity of programming notebooks from scratch

Further work

- Indexing the variant database with Elasticsearch
- Building data querying interfaces within the portal (integrating notebook use cases)
    - GA4GH-like Beacon service that returns yes/no answers on variant occurrences
    - GnomAD-like interface that returns aggregations e.g. allele frequencies
    - Integration with the Cohort Builder enabling complex queries with *both* clinical and genomic criteria
- Adding more annotations supporting variant prioritization

Special Thanks To

CHU Ste-Justine

- **Jeremy Costanza**, Lead software architect and developer
- Developers
  - Adrian Paul, Evans Girard,Francis Lavoie
- UX
  - Lucas Lemonnier

CHOP

- DevOps
  - **Alex Lubneuski**
- Bioinformatics
  - **Yuankun Zhu**, Yiran Guo, Miguel Brown, David Higgins

COMMON FUND DATA ECOSYSTEM

AN OVERVIEW

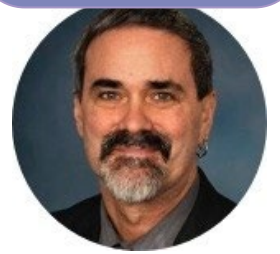
OCTOBER 6TH, 2020

C. TITUS BROWN, UC DAVIS

nih-cfde.org

CFDE - CC
Owen White, PI

UMB
Owen White, PI



USC/ISI
Carl Kesselman, PI

UCHI
Ian Foster, PI

RTI
Becky Boyles, PI

UMB
Bob Carter, PM

Oxford
Susanna Sansone, PI

ICAHN/Mt. Sinai
Avi Ma'ayan, PI

UCD
Titus Brown, PI

UCD
Amanda Charbonneau, PM

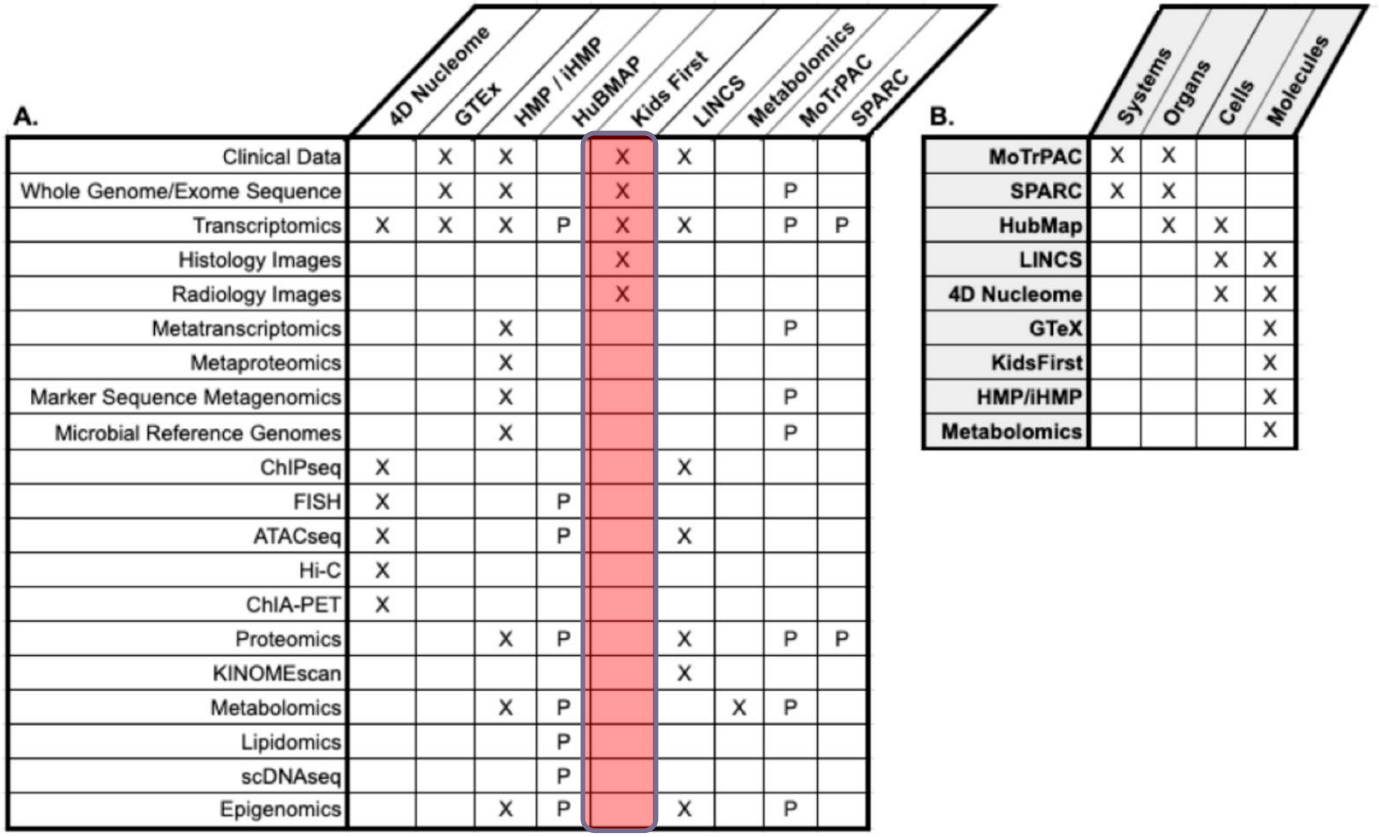
MISSION: INCREASING DATA REUSE

- Enhance
  - Reuse, and comparison to NIH biomedical resources
  - Access to current and sunsetted programs
  - Application of tools / systems across (CF) programs
  - Ability to ask scientific questions across (CF) data sets

- Increase the on-ramp for more researchers with training

WHY: COMMON FUND DATA ECOSYSTEM

DIVERSE PORTFOLIO



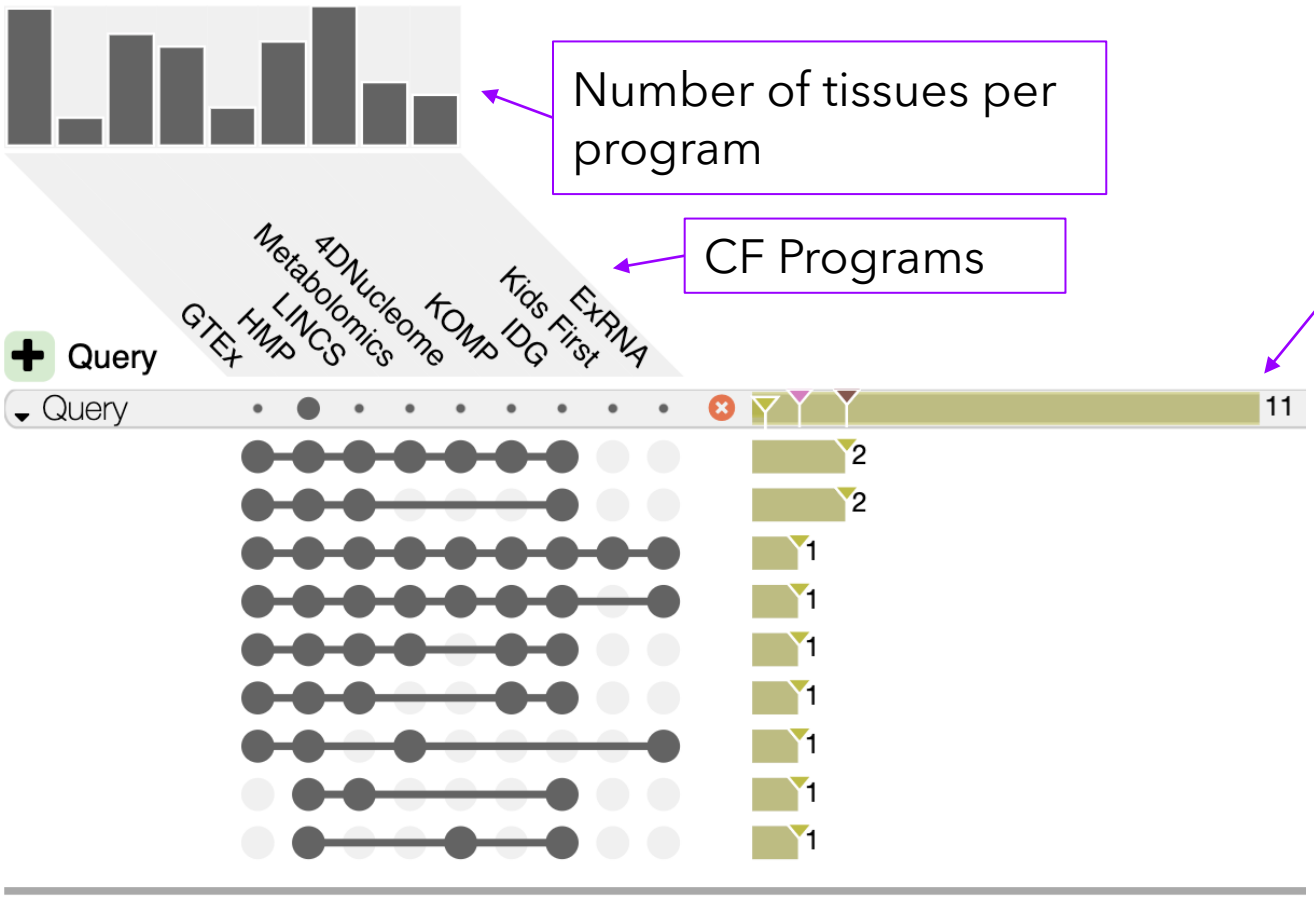
A.

	4D Nucleome	GTEx	HMP / iHMP	HuBMAP	Kids First	LINCS	Metabolomics	MoTrPAC	SPARC
Clinical Data		X	X		X	X			
Whole Genome/Exome Sequence		X	X		X			P	
Transcriptomics	X	X	X	P	X	X		P	P
Histology Images					X				
Radiology Images					X				
Metatranscriptomics			X					P	
Metaproteomics			X						
Marker Sequence Metagenomics			X					P	
Microbial Reference Genomes			X					P	
ChIPseq	X					X			
FISH	X			P					
ATACseq	X			P		X			
Hi-C	X								
ChIA-PET	X								
Proteomics			X	P		X		P	P
KINOMEscan						X			
Metabolomics			X	P			X	P	
Lipidomics				P					
scDNAseq				P					
Epigenomics			X	P		X		P	

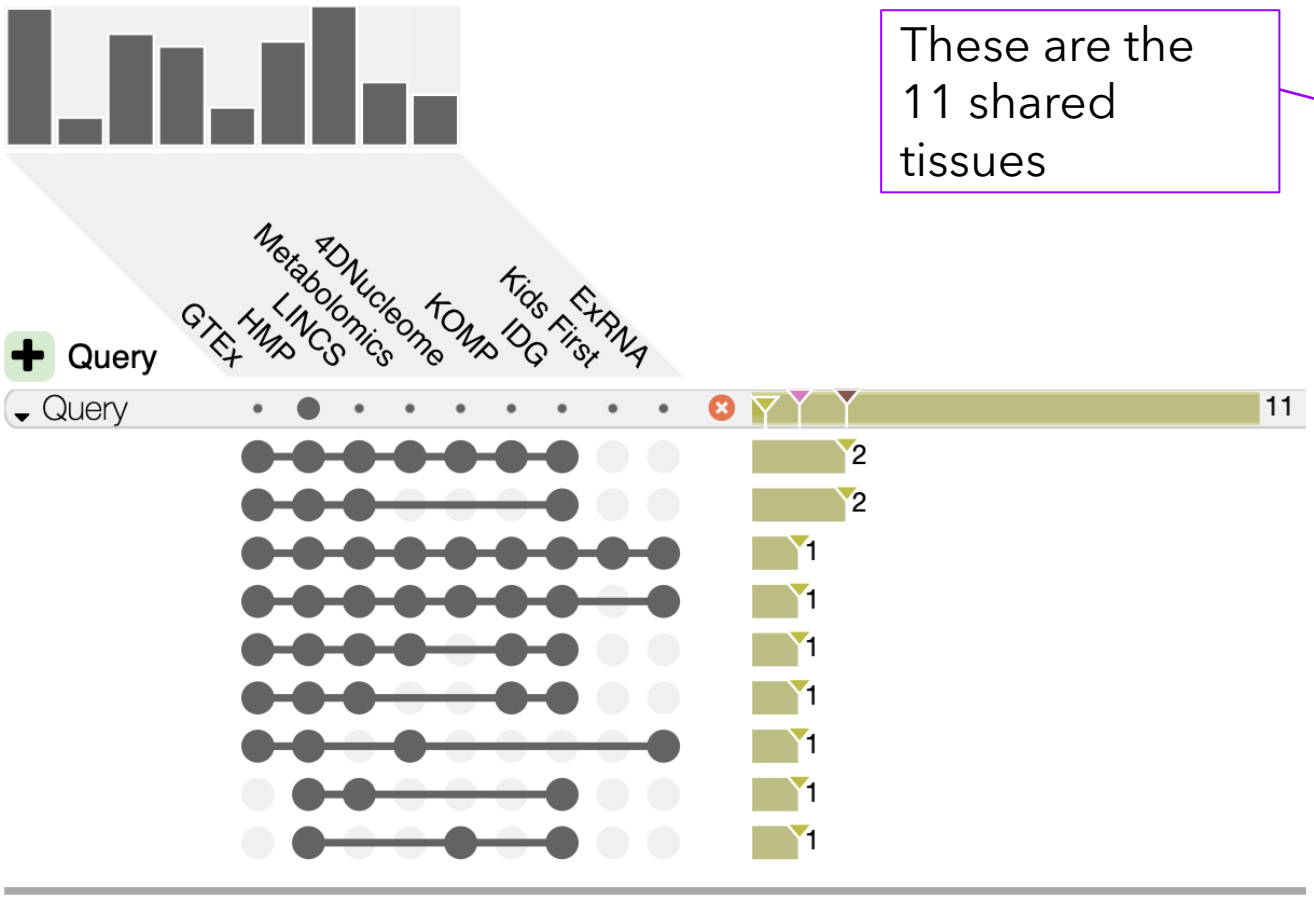
B.

	Systems	Organs	Cells	Molecules
MoTrPAC	X	X		
SPARC	X	X		
HubMap		X	X	
LINCS			X	X
4D Nucleome			X	X
GTeX				X
KidsFirst				X
HMP/iHMP				X
Metabolomics				X

TISSUE TYPES SHARED BY HMP AND OTHER CF PROGRAMS



TISSUE TYPES SHARED BY HMP AND OTHER CF PROGRAMS

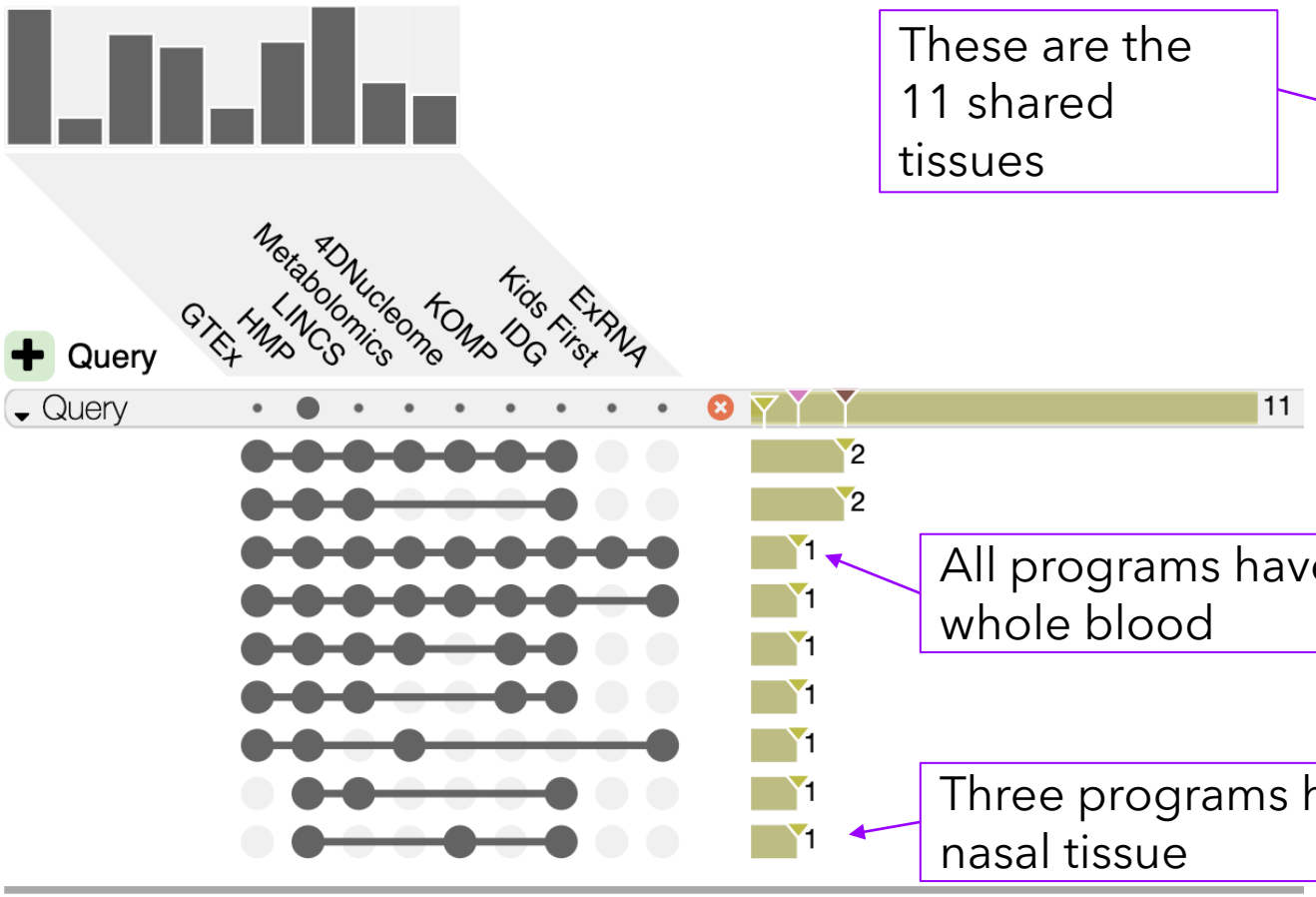


These are the 11 shared tissues

Query Results

Tissue	Set Count
Cervix - Ectocervix	4
Colon - Sigmoid	9
Colon - Transverse	10
Esophagus - Mucosa	5
Minor Salivary Gland	5
Nasal	3
Rectum	3
Skin - Not Sun Exposed (Suprapubic)	7
Skin - Sun Exposed (Lower leg)	6
Vagina	4
Whole Blood	11

TISSUE TYPES SHARED BY HMP AND OTHER CF PROGRAMS



These are the 11 shared tissues

Query Results

Tissue	Set Count
Cervix - Ectocervix	4
Colon - Sigmoid	9
Colon - Transverse	10
Esophagus - Mucosa	5
Minor Salivary Gland	5
Nasal	3
Rectum	3
Skin - Not Sun Exposed (Suprapubic)	7
Skin - Sun Exposed (Lower leg)	6
Vagina	4
Whole Blood	11

All programs have whole blood

Three programs have nasal tissue

CFDE INTEGRATION: HOW?

THE CROSSCUT METADATA MODEL (C2M2)

Goal: DCCs to share structured, detailed metadata about their experimental resources across the ecosystem.

Not a warehouse

No data replication

Users directed to DCCs as primary resource

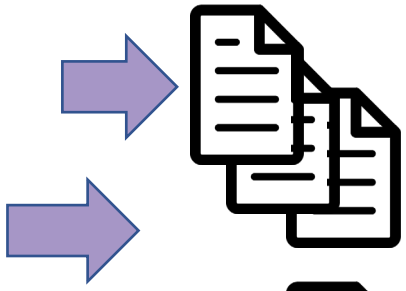
Gabriella Miller Kids First Pediatric Research Program Data Resource Center

GTEx Portal

NIH Human Microbiome Project

4D Nucleome

RNAseq, Variant files



Metagenomic, Electrophysiology

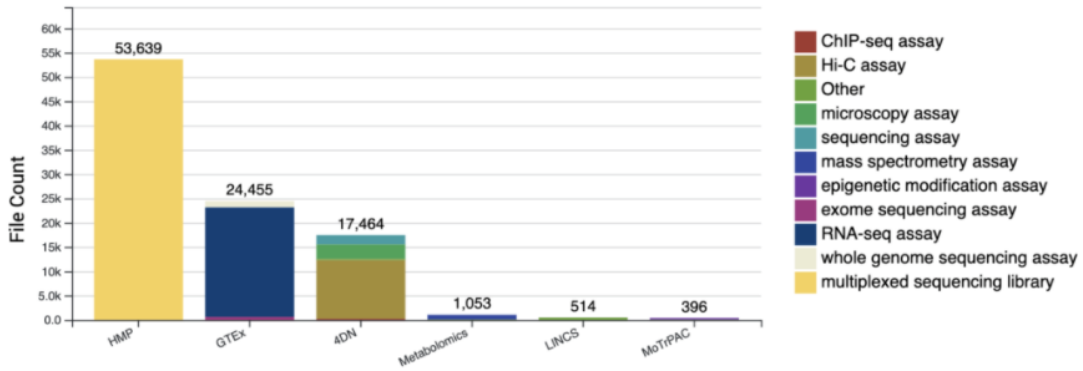
Metadata ONLY

- File type
- Organism
- Assay
- Patient information

CFDE Catalog

Tools Engagement About

Common Fund Data Ecosystem Web Portal



Technical Documentation

Use Case Library

Training

Tools

Publications

FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources

An Open Ecosystem for Pervasive Use of Persistent Identifiers

News & Events

Fall 2020 Cross Pollination Series starts 9/29/2020. Click for more info..

Data portal website coming in November 2020.

The NIH Common Fund has switched to rolling submission of Engagement Plans, effective May 2020. More information available on the Engagement page.

Tweets by @CfdeNih

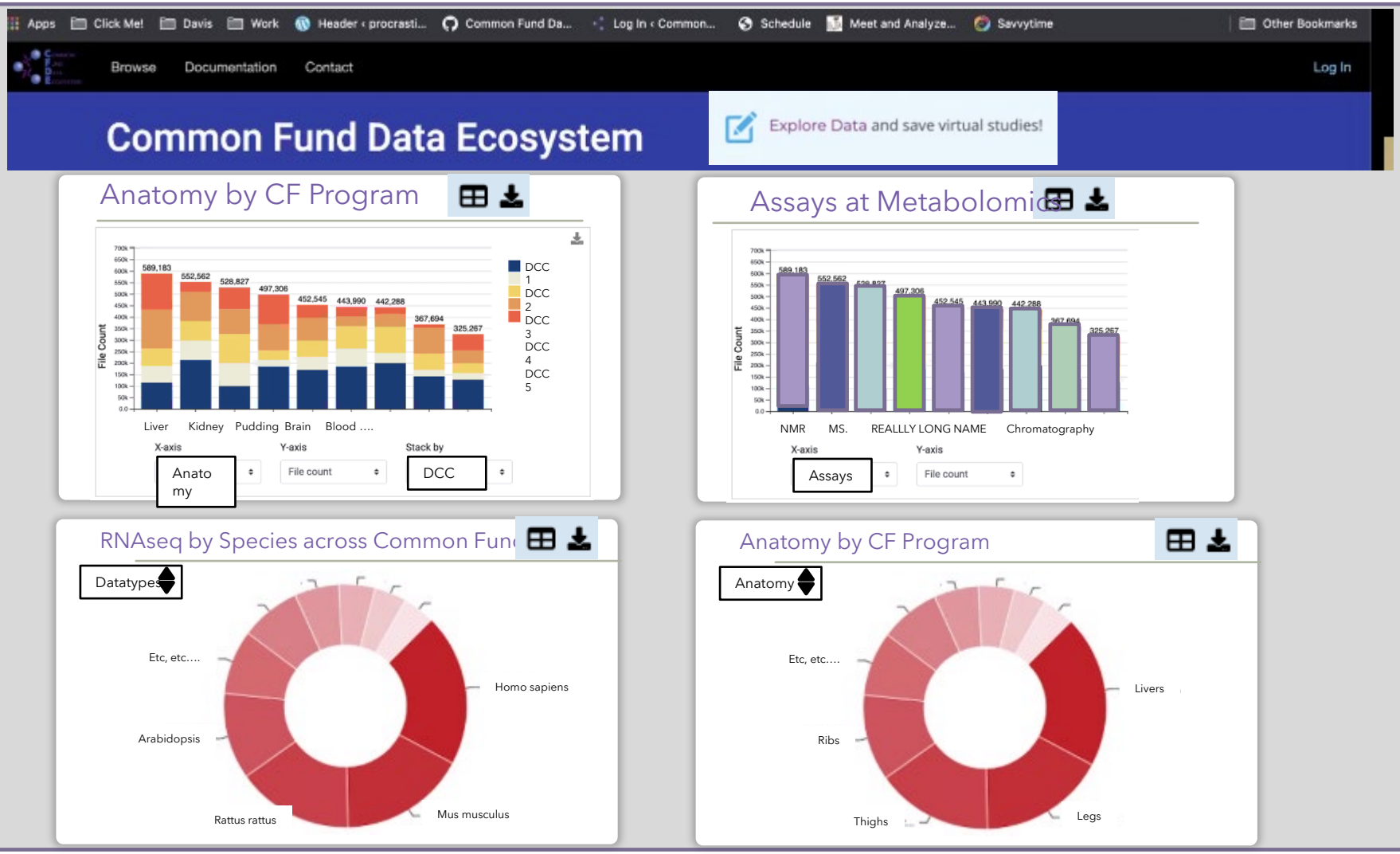
NIH_CFDE Retweeted

Lappalainen Lab
@tuuliel_lab

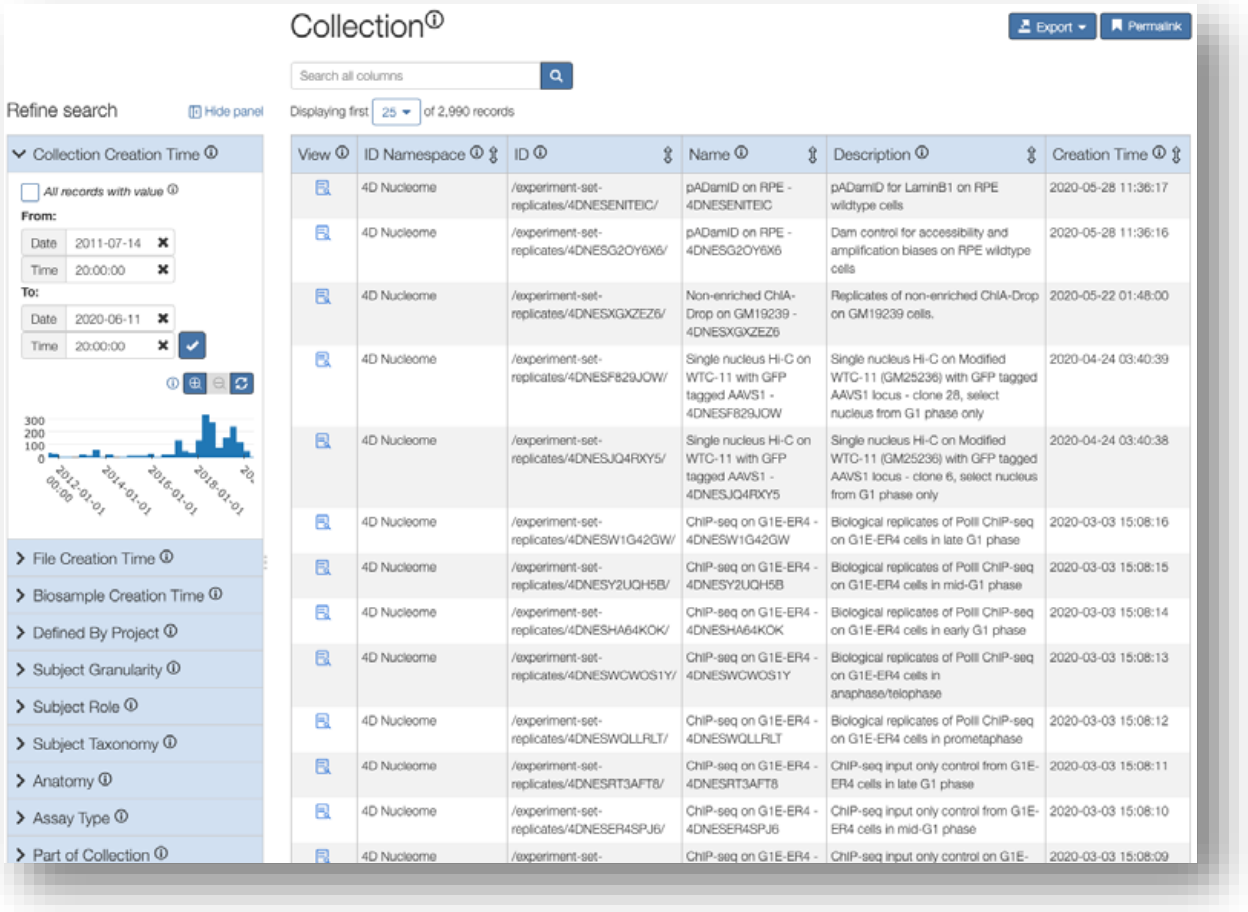
Very happy and proud to present: The GTEx papers. This set of 15 papers published today describes the final phase of this 10-year effort, providing the genomics community an atlas of genetic regulatory variants and a deep dive into the biology behind it. sciencemag.org/collections/ge…

Embed View on Twitter

http://nih-cfde.org


Apps
Click Me!
Davis
Work
Header ‹ procrasti...
Common Fund Da...
Log In ‹ Common...
Schedule
Meet and Analyze...
Savvytime
Other Bookmarks
Browse
Documentation
Contact
Log In
Common Fund Data Ecosystem
Explore Data and save virtual studies!
Anatomy by CF Program
589,183
552,562
528,827
497,306
452,545
443,990
442,288
367,694
325,267
File Count
Liver
Kidney
Pudding
Brain
Blood ....
DCC 1
DCC 2
DCC 3
DCC 4
DCC 5
X-axis
Anatomy
Y-axis
File count
Stack by
DCC
Assays at Metabolomics
NMR
MS.
REALLLY LONG NAME
Chromatography
X-axis
Assays
Y-axis
File count
RNAseq by Species across Common Fun
Datatypes
Etc, etc....
Homo sapiens
Arabidopsis
Rattus rattus
Mus musculus
Anatomy by CF Program
Anatomy
Etc, etc....
Livers
Ribs
Thighs
Legs

CFDE Portal for cross data set search – public beta at end of October!!



MISSION: INCREASING DATA REUSE

- Enhance
  - Reuse, and comparison to NIH biomedical resources
  - Access current and sunsetted programs
  - Application of your tools / systems
  - Ability to ask scientific questions across data sets

- Increase the on-ramp for more researchers with training

A CRITICAL TOOL: USER TRAINING

- Using DCC-specific resources
- Finding, accessing and combining resources between DCCs
- Analyzing data sets in the cloud
- Data-type specific training, e.g. WGS for clinicians
- Modalities include Web sites, videos, teleconference training, and (eventually) in-person.

A CRITICAL TOOL: USER TRAINING **FOR KIDS FIRST:**

We are planning tutorials and training on:

- Logging into and using the KF DRC portal to find data sets
- Running data analyses in the Cavatica cloud platform
  - WGS?
  - RNAseq?
- Data-type specific training, e.g. WGS for clinicians
- Modalities include Web sites, videos, teleconference training, and (eventually…) in-person.

A CRITICAL TOOL: USER TRAINING **FOR KIDS FIRST:**

We are planning tutorials and training on:

- Logging into and using the KF DRC portal to find data sets
- Running data analyses in the Cavatica cloud platform
  - WGS?
  - RNAseq?
- **Data-type specific training, e.g. WGS for clinicians**
- Modalities include Web sites, videos, teleconference training, and (eventually…) in-person.

- Also happy to talk with people who are developing tools about testing, documenting, training, etc! :)

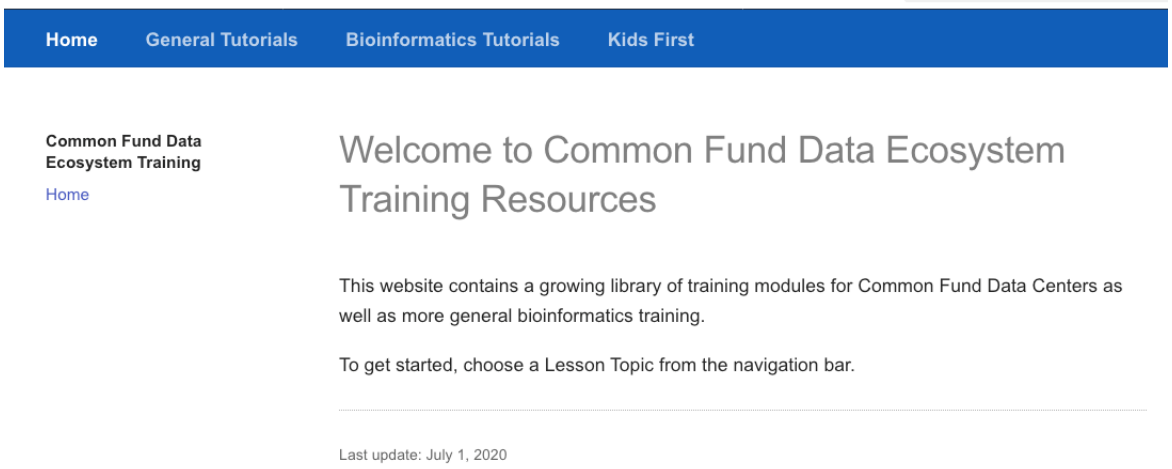
SUPPORT **FOR KIDS FIRST:**

We will also support Kids First DRC in joining forces with other NIH programs:

- Cross-program analysis (WGS, RNAseq) – e.g. GTEx
- scRNA (HuBMAP) and/or drug effects (LINCS) and/or …

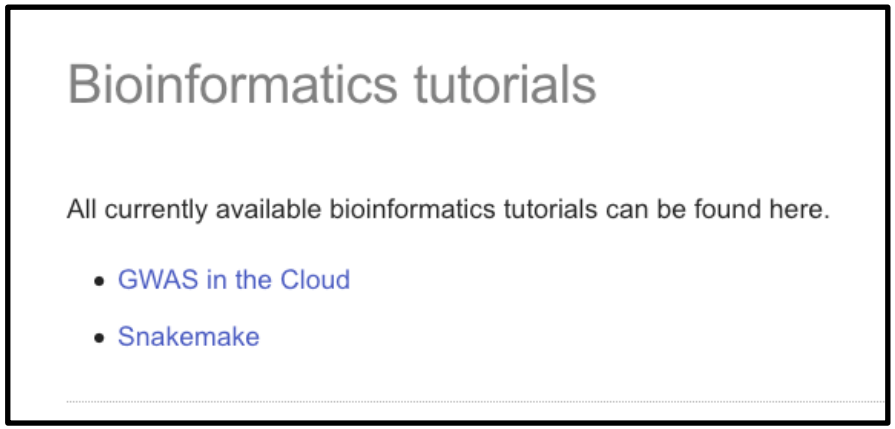
PUBLIC CFDE RESOURCES: TRAINING WEB SITE

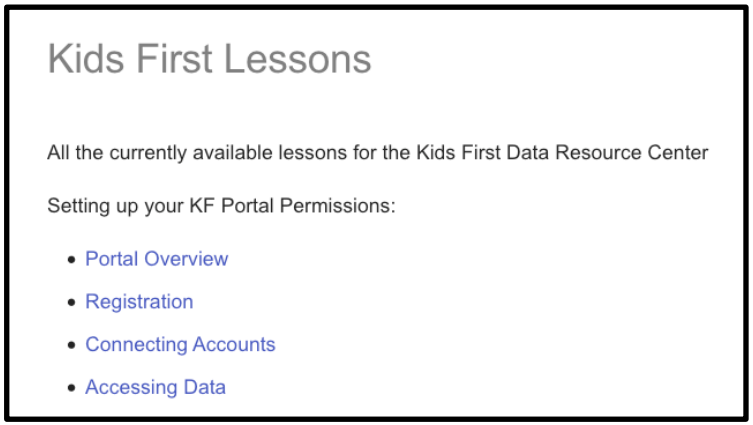
- Training Website https://cfde-training-and-engagement.readthedocs-hosted.com/
  - Modules for learning how to use CFDE resources, tutorials on general bioinformatics workflows, and DCC specific tutorials



PUBLIC CFDE RESOURCES: TRAINING WEB SITE

- Training Website https://cfde-training-and-engagement.readthedocs-hosted.com/
  - Modules for learning how to use CFDE resources, tutorials on general bioinformatics workflows, and DCC specific tutorials



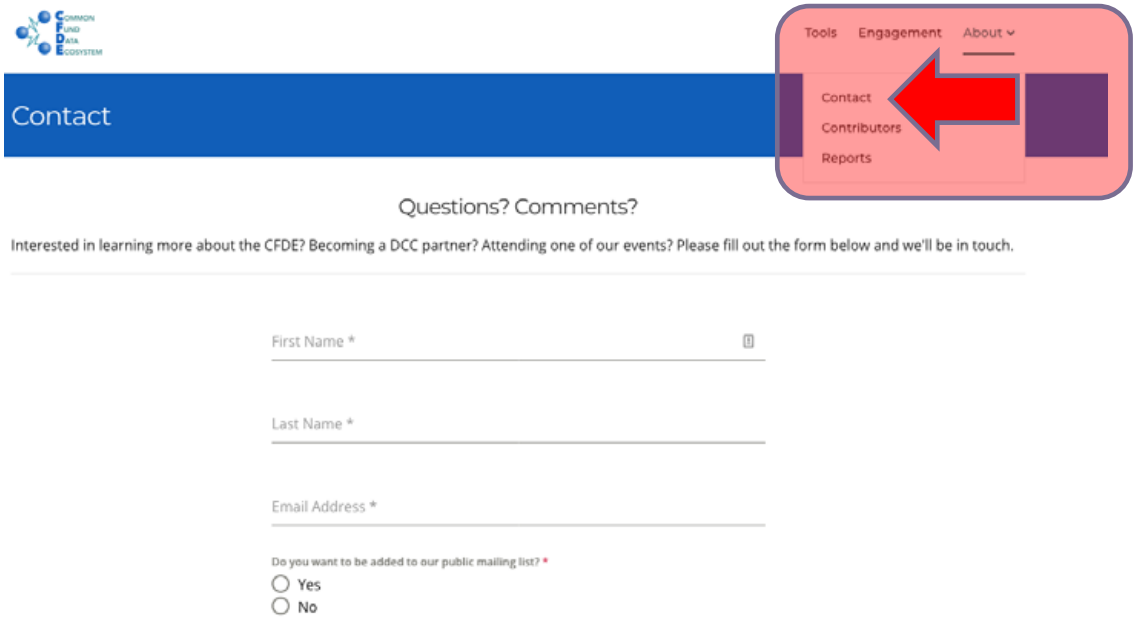


WE WELCOME FEEDBACK!

ctbrown@ucdavis.edu

Please let us know:

- What kind of training you are interested in!
- What kind of data set combinations you are interested in!



http://nih-cfde.org

Gabriella Miller Kids First Pediatric Research Program

Program Updates

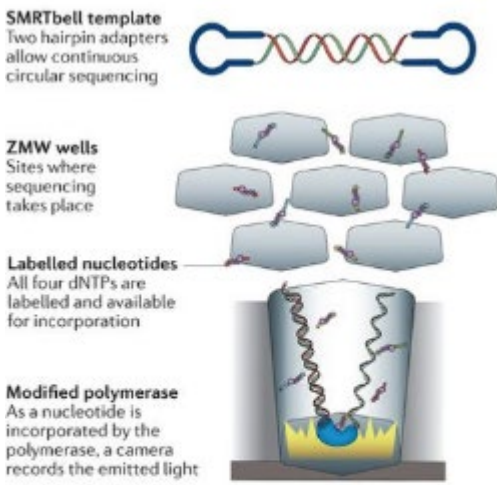
FY20 Long-Read Sequencing Pilot

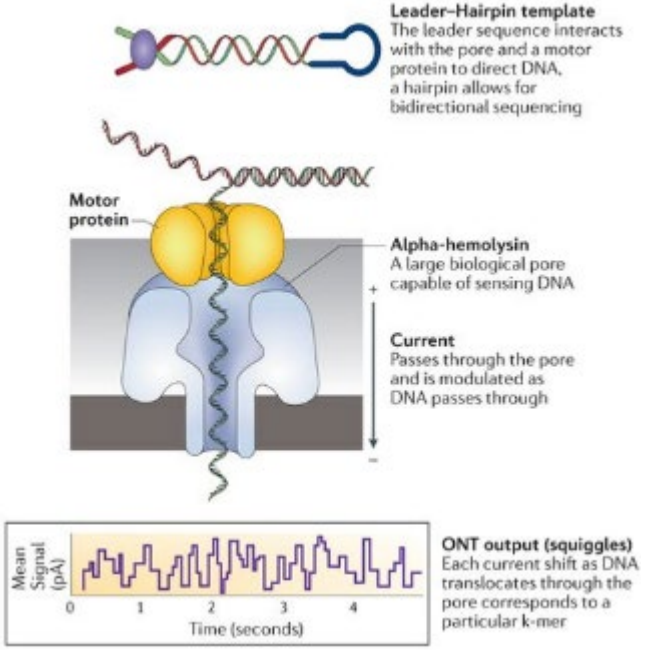
2020 Kids First X01
Long Read Sequencing Pilot

BROAD INSTITUTE

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY

St. Jude Children's Research Hospital





PACIFIC BIOSCIENCES

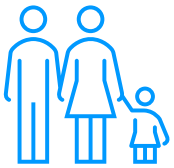
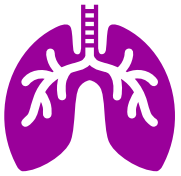
OXFORD NANOPORE

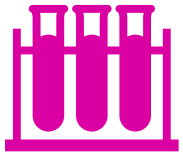
Adapted from Goodwin et. al 2016, https://doi.org/10.1038/nrg.2016.49

FY20 Long-Read Projects

https://commonfund.nih.gov/kidsfirst/longreadprojects

- **Nonsyndromic Craniosynostosis**
  - PI: Simeon Boyd
- **Congenital Diaphragmatic Hernia**
  - PI: Wendy Chung
- **Structural Heart & Other Defects**
  - PI: Bruce Gelb
- **Bladder Exstrophy Epispadias Complex (BEEC)**
  - PI: Angie Jelin
- **Acute Myeloid Leukemia**
  - PI: Soheil Meshinchi
- **Novel Cancer Susceptibility**
  - PI: Sharon Plon
- **Ollier Disease & Maffucci Syndrome**
  - PI: Nara Sobreira




Strategic Planning & Phase 2 Initiatives

What is next for Kids First?

2018 Strategic (Re-)Planning Exercise:

- **2018 Program Survey launched at ASHG:** Kids First & external investigators
- **Kids First Steering Committee**
- **External Program Consultants**
- **DRC Admin & Outreach Core** (feedback from the public, patients, families)
- **NIH Kids First Working Group & IC Director Co-Chairs**

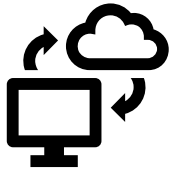
7 Consensus Recommendation Themes

1. **Innovation: Resource, infrastructure, or tool development.**
   Activities: Data Visualization tools; other tools for clinical/phenotypic data

2. **Clinical/phenotypic data extraction, harmonization, & curation.**
   Activities: Collect, extract, organize, curate, harmonize, and submit deep clinical and phenotypic data; annotate variants with pathogenicity, ClinGen scores.

3. **Collaborative validation and discovery.**
   Activities: Building synthetic cohorts; identify structural variants; test pipelines.
   *Engage trainees in data analysis projects**Bring users to the platform*

4. **Integration and interoperability of external pediatric datasets.**
   Activities: Using DRC workflow and best practices to harmonize external pediatric datasets; Building tools that can operate across multiple spaces



5. **Consent and data sharing.**
   Activities: Re-consenting cohorts in line with our data sharing expectations

6. **Validation with model organisms.**
   Activities: validating KF findings/variants, deep phenotyping of animal models

7. **Continue WGS & data generation**, invest in long-read, consider other –omics. Reissues of: https://grants.nih.gov/grants/guide/pa-files/PAR-19-104.html

NIH Council of Councils
September 11, 2020

Common Fund Concept Clearance: Gabriella Miller Kids First Pediatric Research Program: Plans for FY22-24 (Phase 2)

James Coulombe, Ph.D.
Chief, Developmental Biology and Structural Variation Branch
Eunice Kennedy Shriver National Institute of Child Health and Human Development

Goal of Phase 2 Initiatives: Enhance the value and impact of the Kids First Data Resource to accelerate pediatric research to improve preventative measures, diagnostics, and therapeutic interventions.

The archived videocast of the Council of Councils meeting is publicly available and can be viewed here (Kids First discussion begins at 4:48:00).

The presentation materials are available here.

Phase 2 Initiatives Approved

$12.6M/year (FY22-24)

1) **Additional generation of childhood cancer and structural birth defects-related -omics data**
   - Add epigenomic and proteomic assays
2) **Continue development & improvement of the Data Resource**
   - Plan for sustaining the Data Resource beyond FY24
3) **Expert-driven activities to increase the value of Kids First data**
   - Engage Kids First & community experts in activities such as integration, curation, and/or harmonization of rich clinical and phenotypic data

Collaborations

In parallel, continue trans-NIH collaborations to address common goals

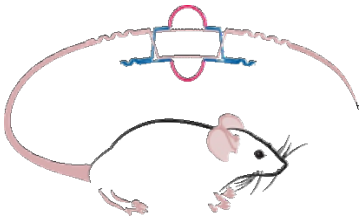
Collaborations amplify the impact of Kids First

COMMON FUND DATA ECOSYSTEM

db GaP GENOTYPES and PHENOTYPES

INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE)

BioData CATALYST



KOMP2 IMPC International Mouse Phenotyping Consortium

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

monarch INITIATIVE

AnVILproject

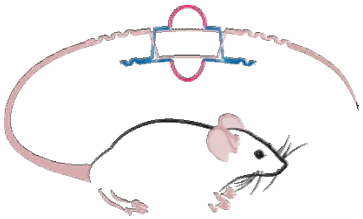
Collaborations amplify the impact of Kids First

COMMON FUND DATA ECOSYSTEM

db GaP GENOTYPES and PHENOTYPES

INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE)

BioData CATALYST



KOMP2 IMPC International Mouse Phenotyping Consortium

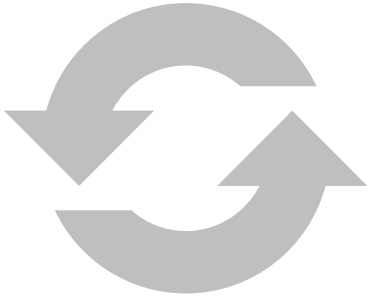
NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

monarch INITIATIVE

AnVILproject

NIH Cloud Based Platforms Interoperability (NCPI)

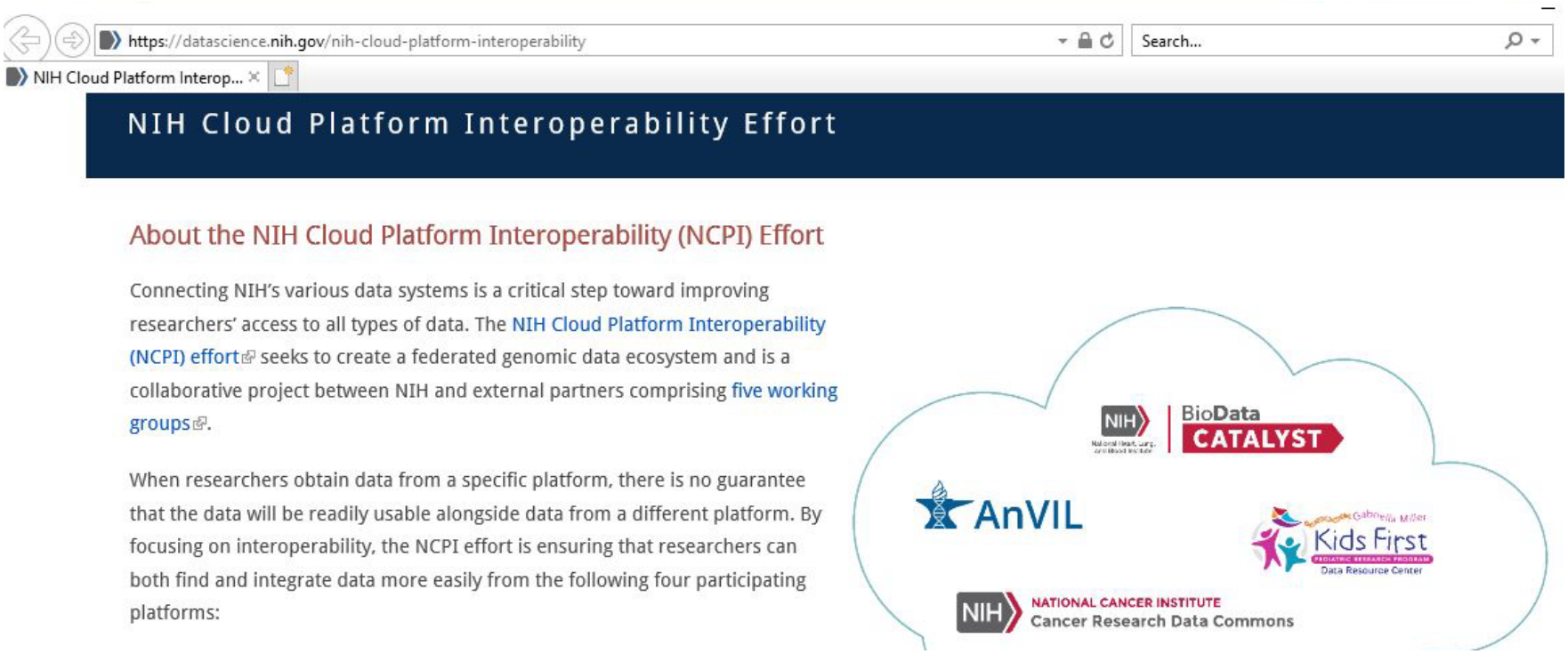
Goal: Empower end-user analyses across platforms through federation & interoperability



New NCPI Webpage!

https://datascience.nih.gov/nih-cloud-platform-interoperability

National Institutes of Health
Office of Data Science Strategy



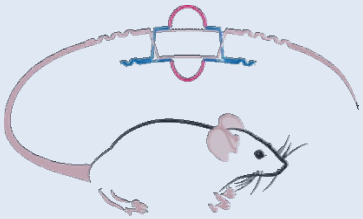
Collaborations amplify the impact of Kids First

COMMON FUND DATA ECOSYSTEM

db GaP GENOTYPES and PHENOTYPES

INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndrome (INCLUDE)

BioData CATALYST



KOMP2 IMPC International Mouse Phenotyping Consortium

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

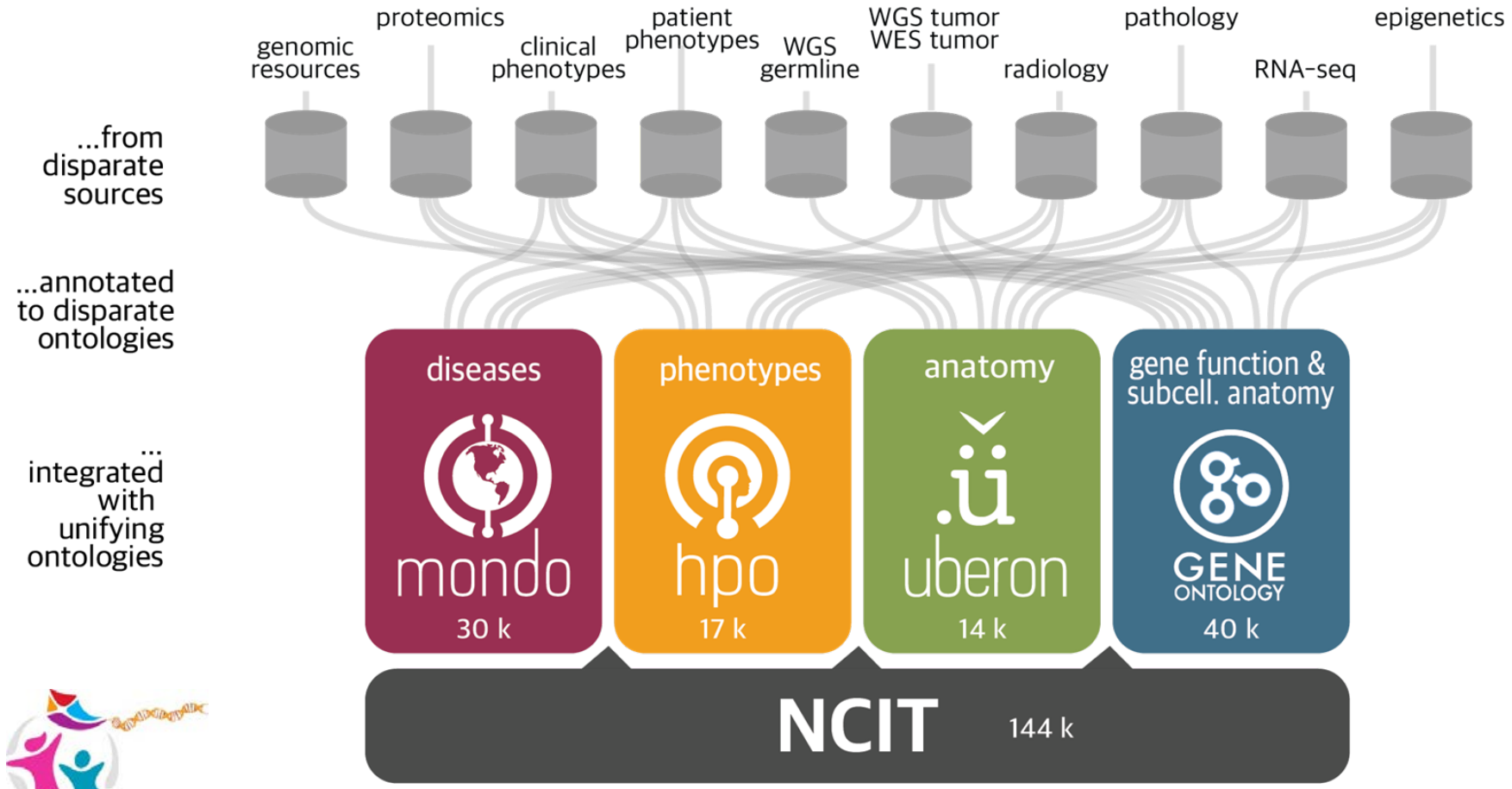
monarch INITIATIVE

AnVILproject

Innovation across the Phenotypic Translational Divide Webinar

Information: https://monarch-initiative.github.io/phenomics/pages/clin-phen-webinar.html

Curation with ontologies that support heterogenous data types in Kids First



Innovation across the Phenotypic Translational Divide Webinar Series

Part 1

Pediatric Cardiac Genomics Consortium (PCGC) - Betsy Goldmuntz	**Down Syndrome -Congenital Heart Disease** - Joaquin Espinosa	**Neuropsycho-logical Data Harmonization** - Stephanie Sherman
Down Syndrome-ALL & Rhabdomyo-sarcoma - Phillip Lupo		**Enchondroma-toses and Related Malignant Tumors** - Nara Sobreira
Orofacial Clefts - Mary Marazita	**Adolescent Idiopathic Scoliosis** - Carole Wise	**Cornelia de Lange Syndrome** - Sarah Raible

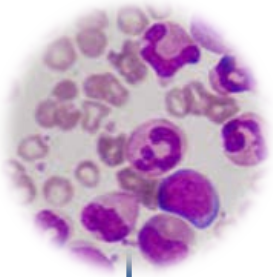
Part 2

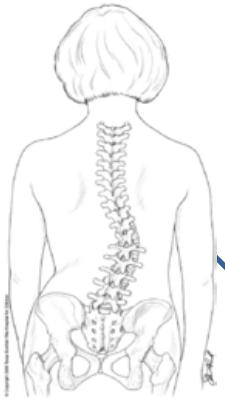
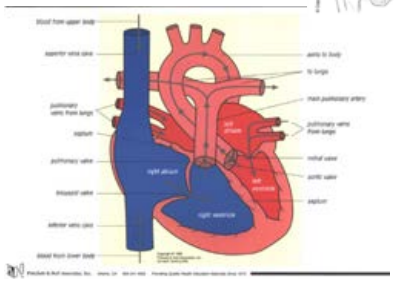
- **Kidney and Urinary Tract Defects**
  Ali Gharavi, Columbia University
- **Structural birth defects and childhood cancer**
  Xiang Wang, Children's Hospital of Philadelphia
- **Neuroblastoma**
  Sharon Diskin, Children's Hospital of Philadelphia
- **CHARGE syndrome**
  Donna Martin, Michigan Medicine
- **Laterality Birth Defects**
  Stephanie Ware, Indiana University
- **Structural Brain Defects, Neural Tube Defects**
  Joe Gleeson, University of California, San Diego

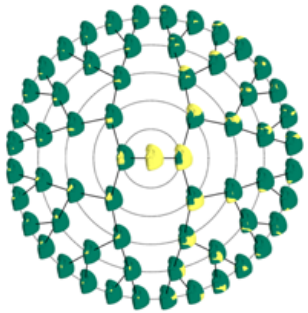
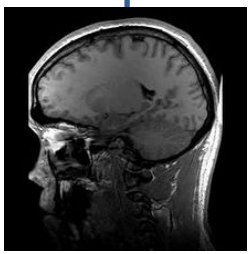
Webinar Information:
https://monarch-initiative.github.io/phenomics/pages/clin-phen-webinar.html

Part 3: ***Cross-Species Genotype-Phenotype Analysis***

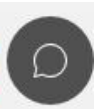
Save-the-Date: *Dec 9th 3-6pm ET*
Information will be shared here



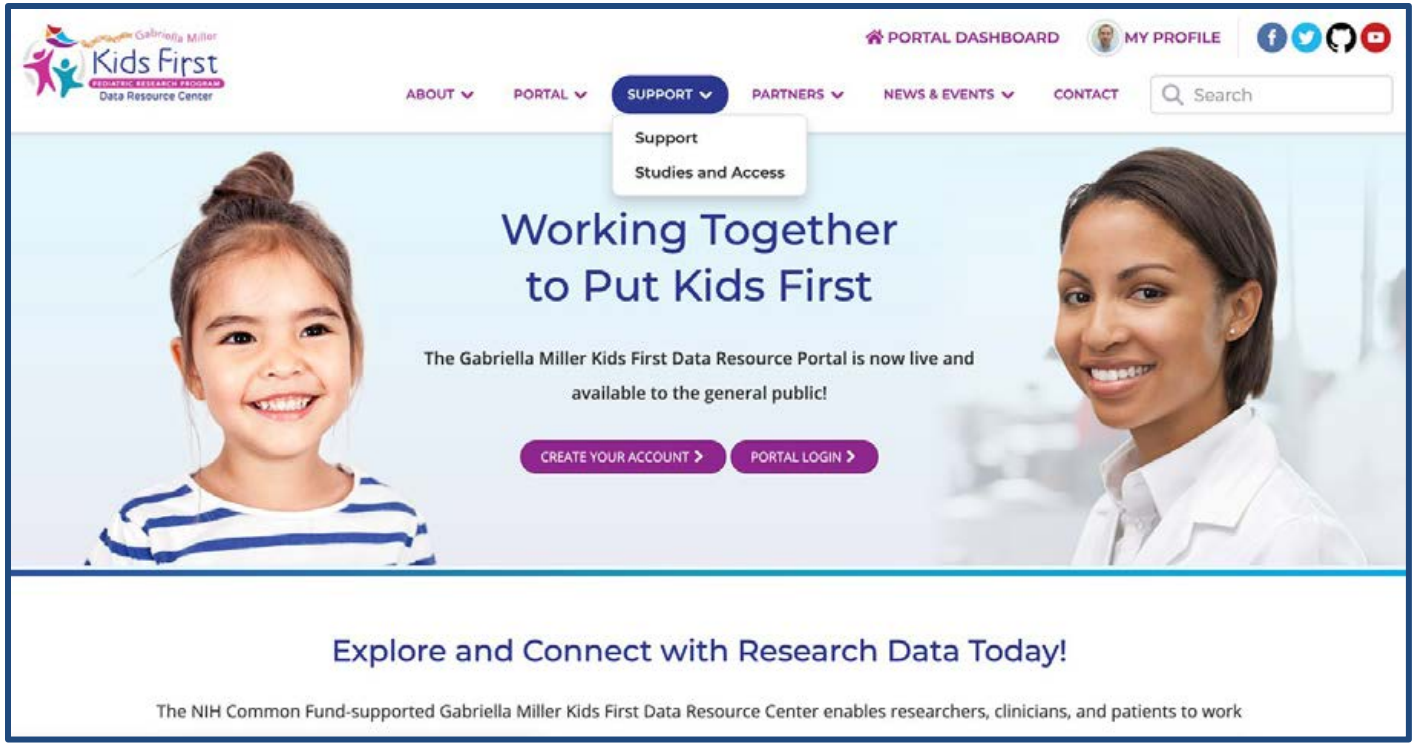

Q & A

- Use the Q&A bar (lower right of your screen) to send your questions to "**All Panelists"**. We will read your questions out loud and answer them.
- You can ask also use the "chat"  service to send private messages to the host or presenters.

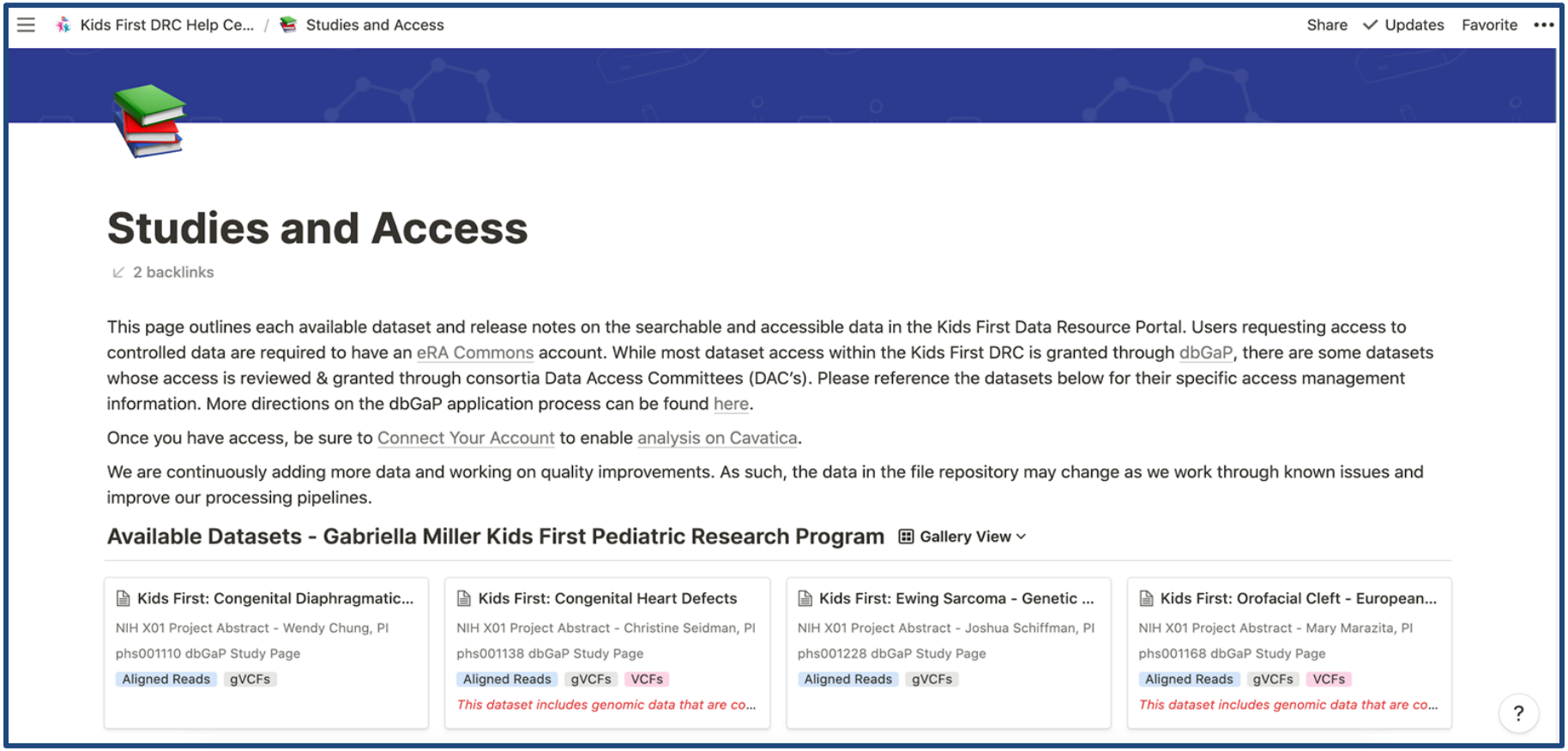
What support is available for learning how to use the Kids First platforms?

What support is available for learning how to use the Kids First platforms?

Support Pages on kidsfirstdrc.org



What support is available for learning how to use the Kids First platforms?



Studies and Access page in the Kids First DRC Help Center

What support is available for learning how to use the Kids First platforms?

Kids First Support Inbox

Send any questions to…

Kids First Support Inbox
Support@KidsFirstDRC.org

What support is available for learning how to use the Kids First platforms?

Kids First Office Hours

What: Monthly user support session hosted by the Kids First DRC

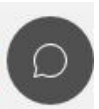
When: 3pm on the 2nd Tuesday of each month

Who: Users of all levels are welcome!

Our next session is **October 13th**.

Contact David Higgins (HigginsD@email.chop.edu) for login information.

Q & A

- Use the Q&A bar (lower right of your screen) to send your questions to “**All Panelists”**. We will read your questions out loud and answer them.
- You can ask also use the “chat”  service to send private messages to the host or presenters.

What funding opportunities are available?

How do I access data?

How can I get involved?

What funding opportunities are available?

Discovery of the Genetic Basis of Childhood Cancers and of Structural Birth Defects: Gabriella Miller Kids First Pediatric Research Program (X01 Clinical Trial Not Allowed)

Kids First cohort sequencing opportunity:

- Propose samples from cohorts for whole genome sequencing at a Kids First-supported sequencing center
  - DNA from germline/normal from affected child and parents/family members (if available)
  - DNA/RNA from tumors and/or affected tissue
- 1 more reissue of PAR-19-390 **for 2021**
- View FY20 materials here
- Considerations include broad data sharing & value of incorporating the data into the Data Resource

What funding opportunities are available?

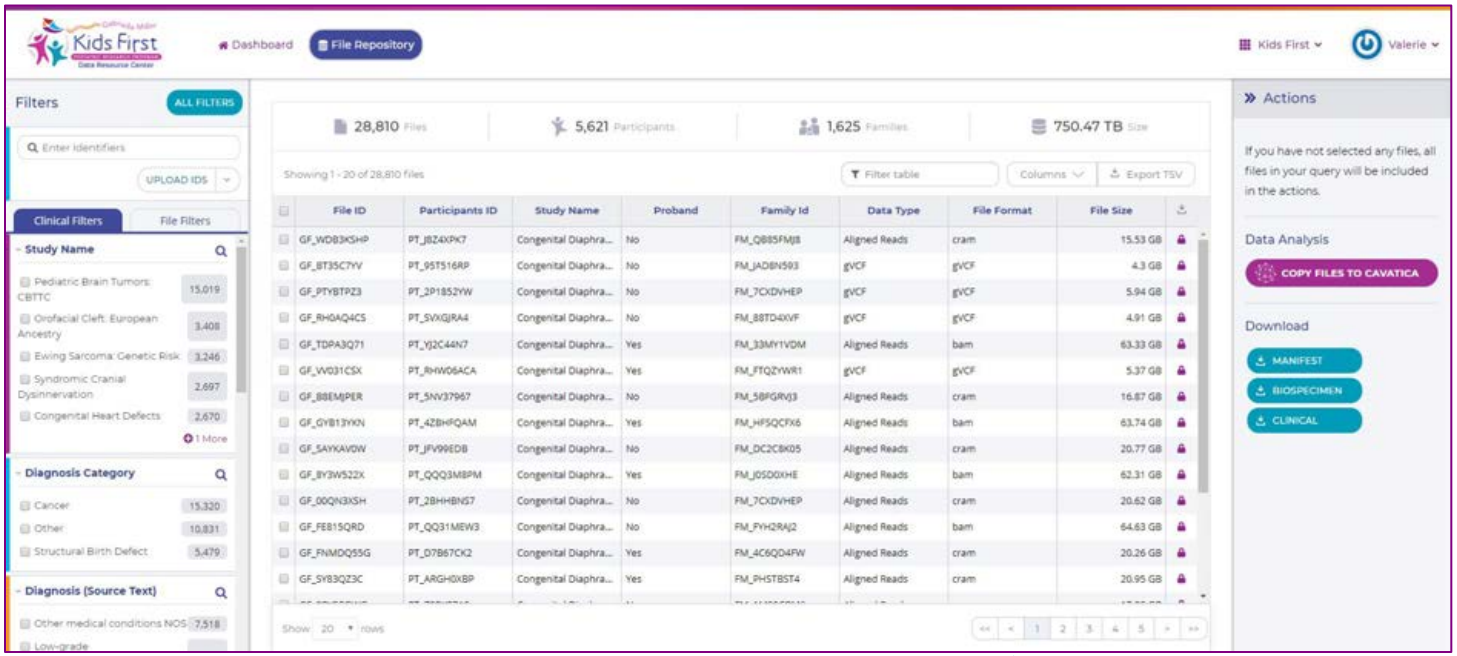
See: FAQs for Funding Opportunities Announcements (FOAs) to Support Data Analyses of Kids First Datasets (https://commonfund.nih.gov/kidsfirst/FAQ)

- **Analyze Kids First data with support from:**
  - **"Kids First R03 PAR": PAR-19-375**
  - **NIH "Parent" R01**: PA-19-056
  - **NIH Parent R03**: PA-19-052
- **Validate variants with support from**:
  - ORIP's Development of Animal Models and Related Biological Materials for Research (R21): https://grants.nih.gov/grants/guide/pa-files/PA-16-141.html
  - Mechanistic Studies of Gene-Environment Interplay in Dental, Oral, Craniofacial, and Other Diseases and Conditions (R01) (PAR-19-292).
  - Development of Novel and Robust Systems for Mechanistic Studies of Gene-Environment Interplay in Dental, Oral, Craniofacial, and Other Diseases and Conditions (R21) (PAR-19-293).
  - To pursue collaborations with the Knockout Mouse Phenotyping Program (KOMP2), contact: KidsFirstKOMP@nih.gov

- **To receive updates about future Kids First opportunities, sign up for the listserv:**
  - https://commonfund.nih.gov/kidsfirst/register

How do I access data?

Anyone can register & login to the portal to filter, search, visualize datasets



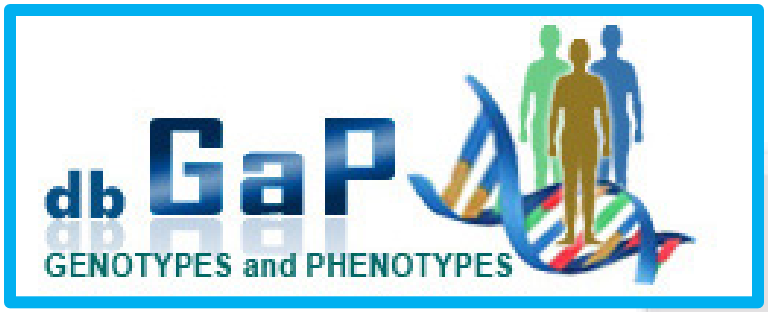
NIH Kids First Data Access Committee

1. Find available data on the portal or Kids First X01 page
2. Submit dbGaP Data Access Requests (DARs) for **individual-level sequence data**
3. **Push approved sequence data to Cavatica from the portal:** https://kidsfirstdrc.org/support/analyze-data/

dbGaP GENOTYPES and PHENOTYPES

Individual-level sequence data

- To learn more about submitting dbGaP Data Access Requests (DARs) watch: https://www.youtube.com/watch?v=39cba0gF2tw&index=3&t=503s&list=PLoXwgZflAe4aMwWpVQU_WVeWHzyhI3BCu



Also see:
https://dbgap.ncbi.nlm.nih.gov/aa/dbgap_request_process.pdf

How can I get involved?

- **Connect with and provide feedback to the DRC:** support@kidsfirstdrc.org
- **Contact the program for questions or feedback:** kidsfirst@od.nih.gov
- **Learn more about the program & DRC:** https://commonfund.nih.gov/kidsfirst & https://kidsfirstdrc.org/
- **Search data available through the Kids First Data Resource Portal:** https://portal.kidsfirstdrc.org/

Q & A

- Use the Q&A bar (lower right of your screen) to send your questions to “**All Panelists”**. We will read your questions out loud and answer them.
- You can ask also use the “chat” service to send private messages to the host or presenters.

Thank You!

Email Additional Questions and Comments to the Kids First Mailbox: **kidsfirst@od.nih.gov**

Kids First Investigators: Past Presentations

- **Congenital Diaphragmatic Hernia**,
  PI: Wendy Chung (April 2019):
  https://www.youtube.com/watch?v=3CS6AphmCp0&t=978s
- **Neuroblastoma**,
  PI: Sharon Diskin (September 2019):
  https://www.youtube.com/watch?v=Gq8kK2UGI4s
- **Orofacial Clefts**
  PI: Mary Marazita (May 2020)
  https://www.youtube.com/watch?v=TddkIx3lZpI