Webinar Instructions

Welcome to the Gabriella Miller Kids First Pediatric Research Program's Public Webinar!

- Every participant is muted upon entry.
- To ask public questions, use the **Q&A** bar (right side of your screen). We encourage you to save these for the question period.
- You can ask also use the "chat" 💬 service to send private messages to the host or presenters throughout the webinar.
- After the webinar, additional program-related questions can be emailed to: KidsFirst@od.nih.gov.

This webinar will be recorded.
We will start at 3pm (EDT)

Gabriella Miller Kids First Pediatric Research Program

Public Webinar

May 18, 2020
3:00 pm EDT

NIH
The Common Fund

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM

May 18th Webinar Agenda

- **3:00pm** - Introduction
- **3:05pm** - Kids First Orofacial Cleft Project Findings
- **3:40pm** - Kids First Data Resource Center
  - New Portal Features
  - Cavatica: Cloud User Workspace Introduction
  - User Workspace Demonstration
  - Kids First DRC Roadmap
- **4:30pm** - Kids First Program & Collaboration Update
- **4:50pm** - Questions and Answers

NIH The Common Fund

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

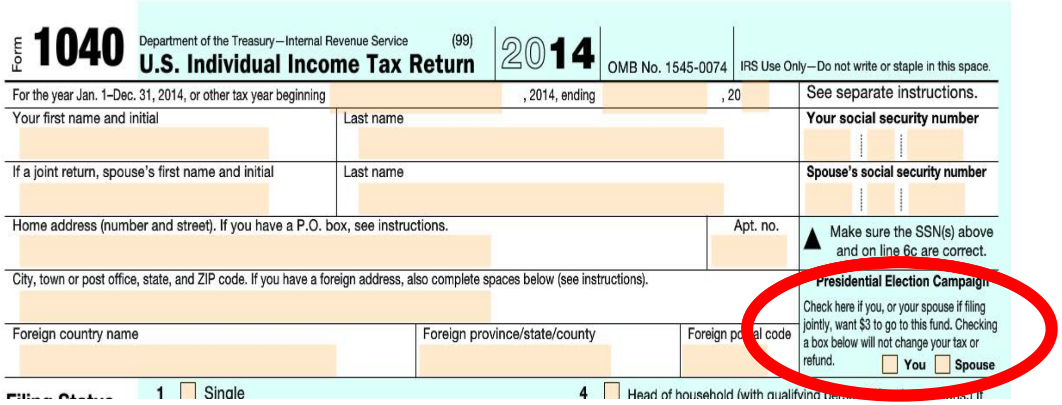
Valerie Cotton

Kids First Program Manager

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

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

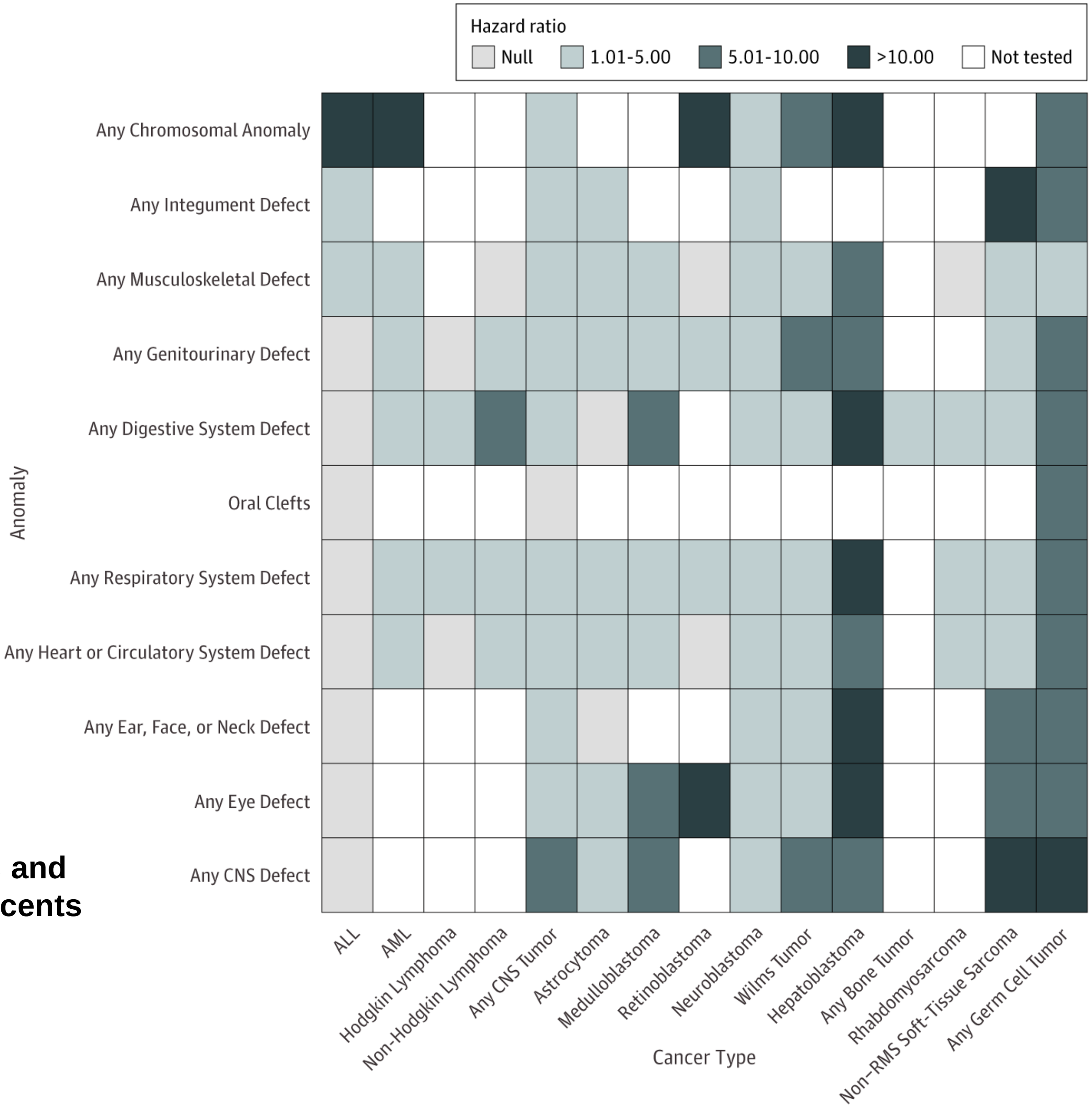
Vision

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM

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?

Birth defects associated with increased risk of cancer among children



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

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**)

Other Working Group Representation:

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

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

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

?

The Kids First Dataset is Growing!

39 projects | 37,000 genomes | 15,000 cases | 10 released datasets

BROAD INSTITUTE

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY

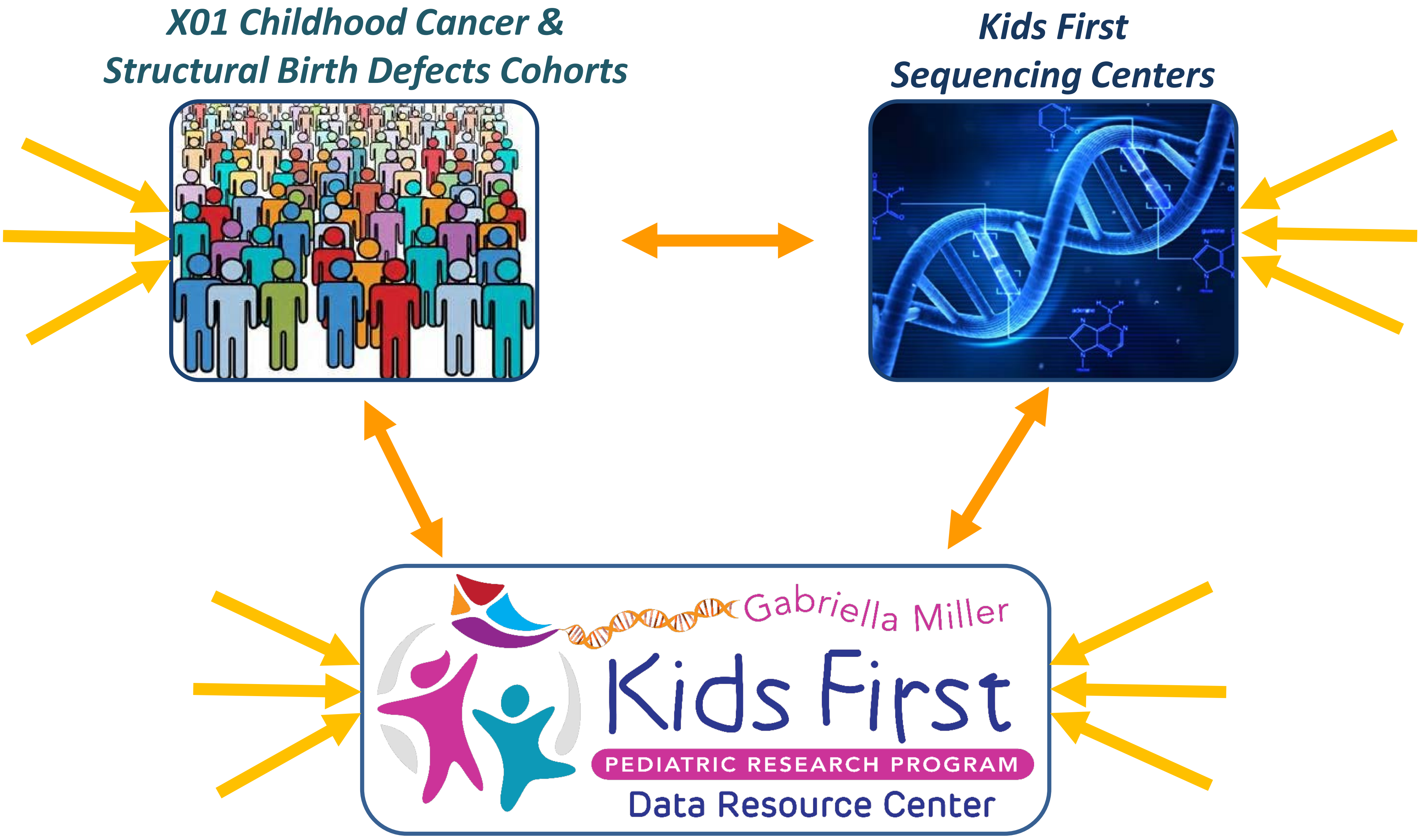
St. Jude Children's Research Hospital

- 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
- Craniofacial Microsomia
- Hemangiomas, Vascular Anomalies & Overgrowth
- Nonsyndromic Craniosynostosis
- Patients with both childhood cancer and birth defects

Kidney and Urinary Tract Defects

- Microtia
- Hearing Loss
- Bladder Exstrophy
- Cornelia de Lange Syndrome
- Intracranial & Extracranial Germ Cell Tumors
- Esophageal Atresia and Tracheoesophageal Fistulas
- Fetal Alcohol Spectrum Disorders
- Myeloid Malignancies + overlap with Down syndrome
- Congenital Heart Defects and 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

The Kids First Community is Growing!



X01 Childhood Cancer & Structural Birth Defects Cohorts

Kids First Sequencing Centers

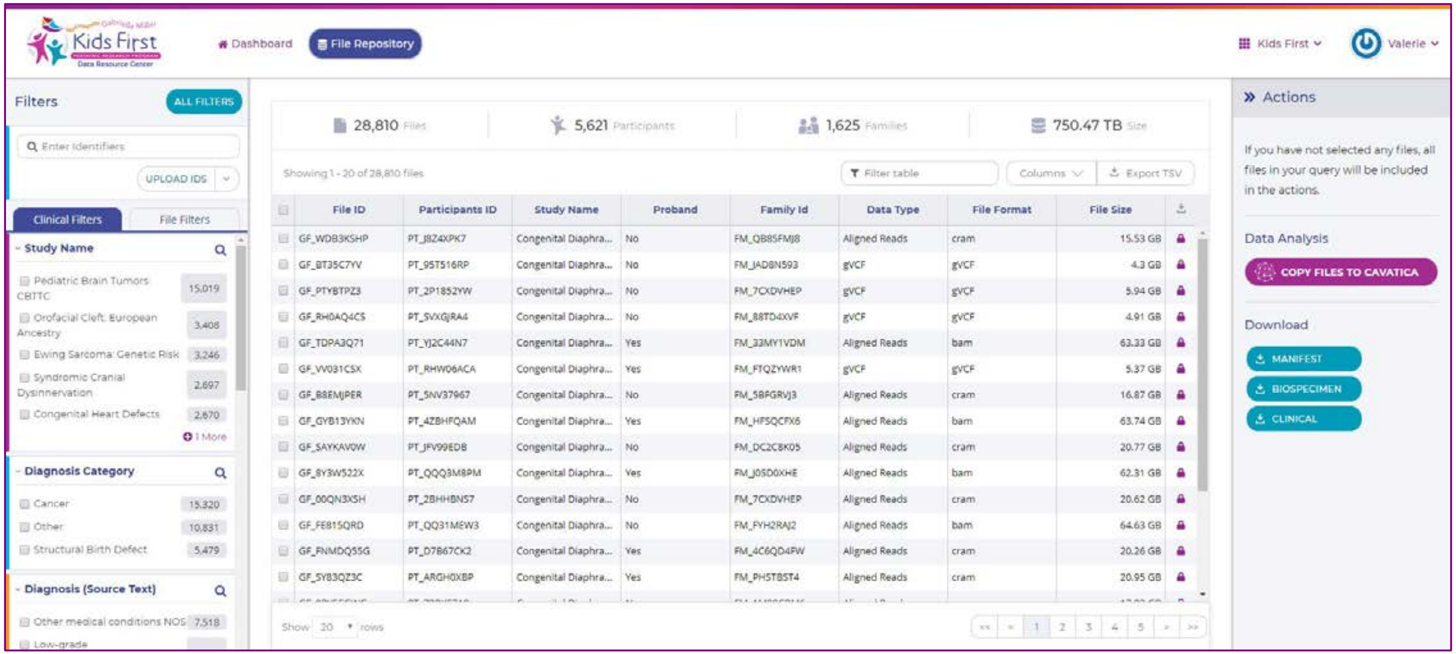
Gabriella Miller Kids First

PEDIATRIC RESEARCH PROGRAM

Data Resource Center

More researchers are accessing Kids First data!

~1500 registered users since 2018 launch



Kids First Pediatric Research Program Data Resource Center

Dashboard | File Repository | Kids First | Valerie

Filters — ALL FILTERS

Enter identifiers

UPLOAD IDS

Clinical Filters | File Filters

Study Name

□ Pediatric Brain Tumors CBTTC	15,019
□ Orofacial Cleft: European Ancestry	3,408
□ Ewing Sarcoma: Genetic Risk	3,246
□ Syndromic Cranial Dysinnervation	2,697
□ Congenital Heart Defects	2,670

1 More

Diagnosis Category

□ Cancer	15,320
□ Other	10,831
□ Structural Birth Defect	5,479

Diagnosis (Source Text)

□ Other medical conditions NOS	7,518
□ Low-grade	

28,810 Files | 5,621 Participants | 1,625 Families | 750.47 TB Size

Showing 1 - 20 of 28,810 files

Filter table | Columns | Export TSV

□	File ID	Participants ID	Study Name	Proband	Family Id	Data Type	File Format	File Size	
□	GF_WDB3KSHP	PT_J8Z4XPK7	Congenital Diaphra...	No	FM_QB85FMJ8	Aligned Reads	cram	15.53 GB	🔒
□	GF_BT35C7YV	PT_95T516RP	Congenital Diaphra...	No	FM_JAD8N593	gVCF	gVCF	4.3 GB	🔒
□	GF_PTYBTPZ3	PT_2P1852YW	Congenital Diaphra...	No	FM_7CXDVHEP	gVCF	gVCF	5.94 GB	🔒
□	GF_RH0AQ4CS	PT_SVXGJRA4	Congenital Diaphra...	No	FM_88TD4XVF	gVCF	gVCF	4.91 GB	🔒
□	GF_TDPA3Q71	PT_YJ2C44N7	Congenital Diaphra...	Yes	FM_33MY1VDM	Aligned Reads	bam	63.33 GB	🔒
□	GF_VW031CSX	PT_RHW06ACA	Congenital Diaphra...	Yes	FM_FTQZYWR1	gVCF	gVCF	5.37 GB	🔒
□	GF_B8EMJPER	PT_5NV37967	Congenital Diaphra...	No	FM_5BFGRVJ3	Aligned Reads	cram	16.87 GB	🔒
□	GF_GYB13YKN	PT_4ZBHFQAM	Congenital Diaphra...	Yes	FM_HFSQCFX6	Aligned Reads	bam	63.74 GB	🔒
□	GF_SAYKAV0W	PT_JFV99EDB	Congenital Diaphra...	No	FM_DC2C8K05	Aligned Reads	cram	20.77 GB	🔒
□	GF_8Y3W522X	PT_QQQ3M8PM	Congenital Diaphra...	Yes	FM_J0SD0XHE	Aligned Reads	bam	62.31 GB	🔒
□	GF_00QN3XSH	PT_2BHHBNS7	Congenital Diaphra...	No	FM_7CXDVHEP	Aligned Reads	cram	20.62 GB	🔒
□	GF_FE815QRD	PT_QQ31MEW3	Congenital Diaphra...	No	FM_FYH2RAJ2	Aligned Reads	bam	64.63 GB	🔒
□	GF_FNMDQ55G	PT_D7B67CK2	Congenital Diaphra...	Yes	FM_4C6QD4FW	Aligned Reads	cram	20.26 GB	🔒
□	GF_SY83QZ3C	PT_ARGH0XBP	Congenital Diaphra...	Yes	FM_PHSTBST4	Aligned Reads	cram	20.95 GB	🔒

Show 20 rows

<< < 1 2 3 4 5 > >>

Actions

If you have not selected any files, all files in your query will be included in the actions.

Data Analysis

COPY FILES TO CAVATICA

Download

MANIFEST

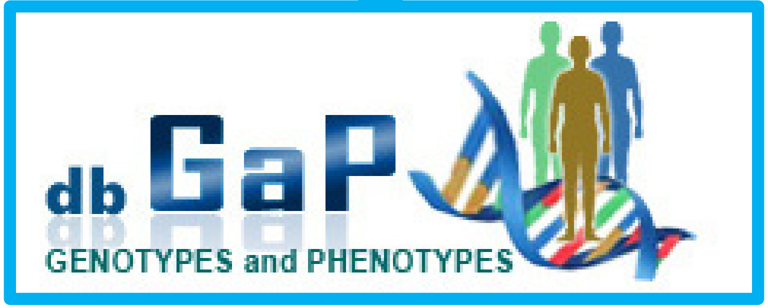
BIOSPECIMEN

CLINICAL

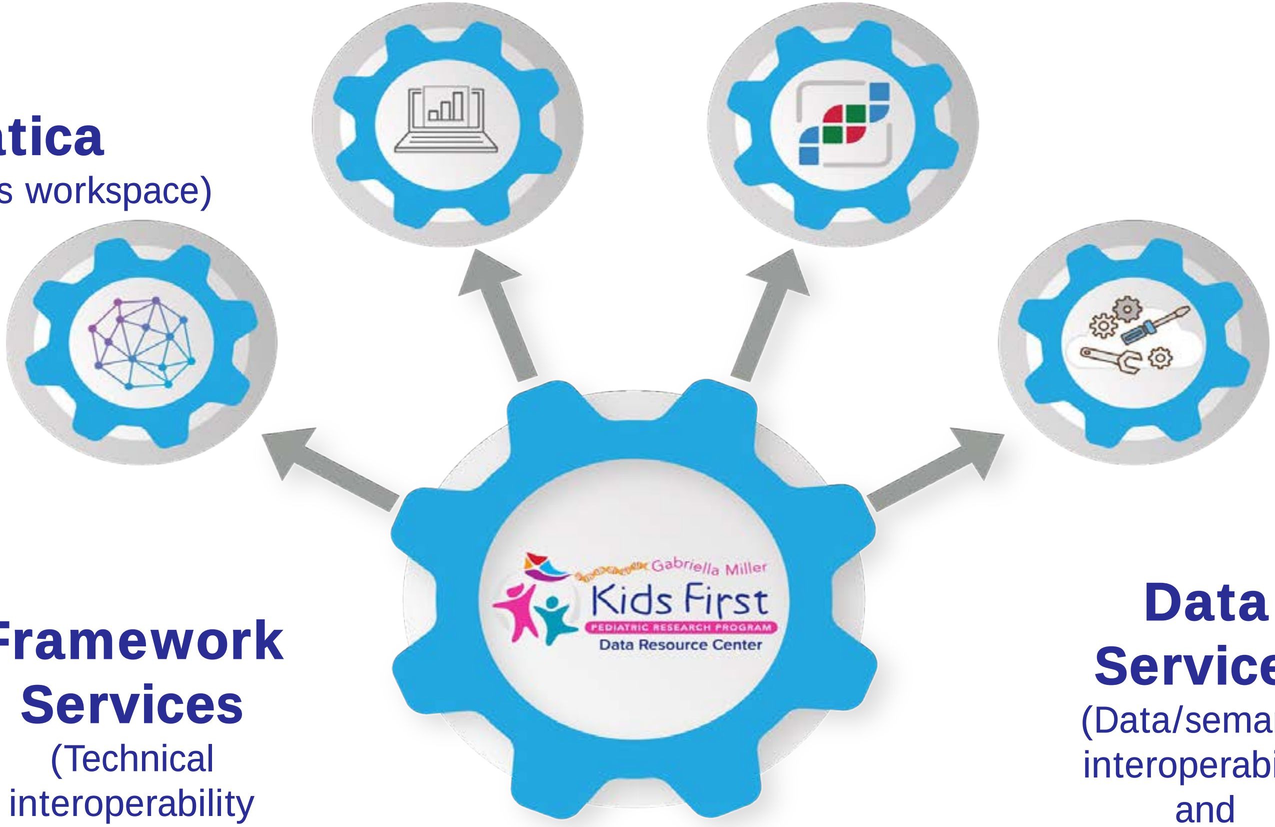
Individual-level sequence data

<u>>100</u> Data Access Requests approved by the Kids First Data Access Committee across **10** Kids First genomic datasets available

NIH Kids First Data Access Committee



The Kids First Data Resource for Collaborative Discovery



DRC Portal
(Entry point, search & build synthetic cohorts)

PedcBioPortal
(Cancer data visualization)

Cavatica
(Cloud analysis workspace)

New Applications & Integrations
(e.g. germline data visualization and analysis)

Framework Services
(Technical interoperability and IAM)

Data Services
(Data/semantic interoperability and coordination)

Kids First X01 Investigators: Orofacial Clefts

Mary Marazita, PhD
University of Pittsburgh

Eleanor Feingold, PhD
University of Pittsburgh

Elizabeth Leslie, PhD
Emory University

Harrison Brand, PhD
Broad Institute

Gabriel Miller Kids First: Orofacial Cleft (OFC) Studies

Mary L. Marazita, Ph.D.; Eleanor Feingold, Ph.D.

and GMKF OFC team

Center for Craniofacial and Dental Genetics

University of Pittsburgh

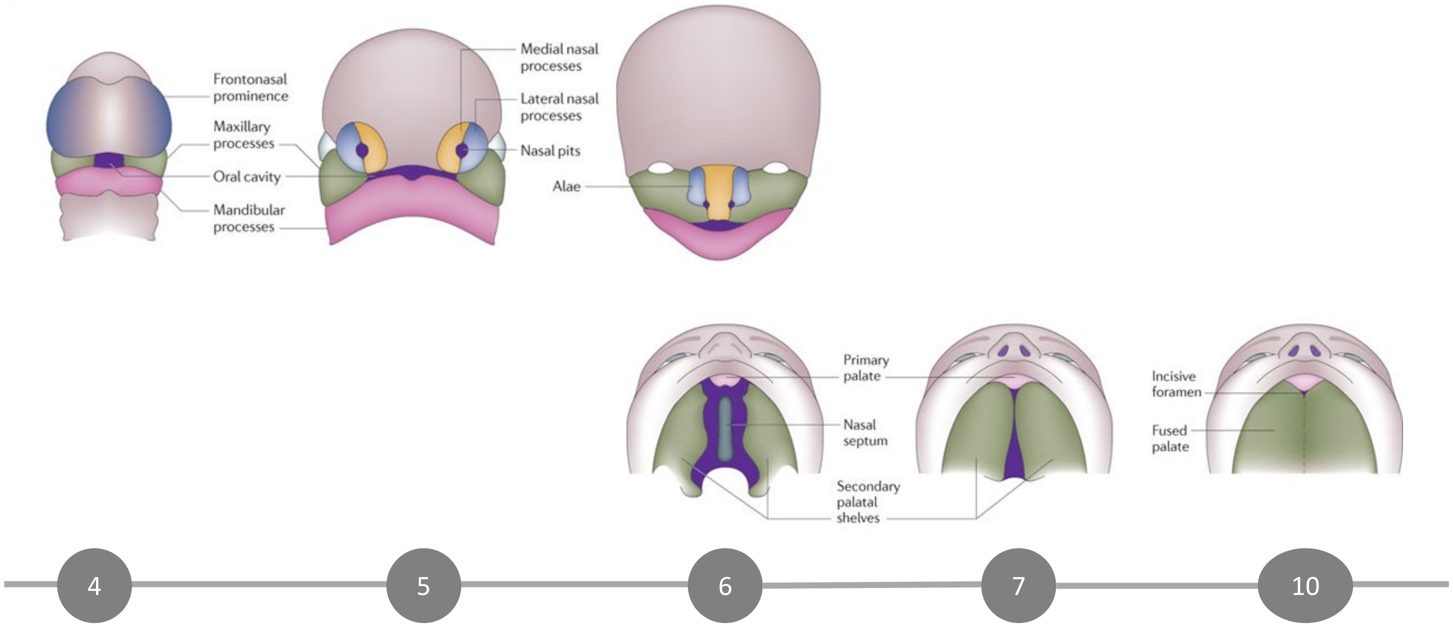
Kids First 2020 Spring Public Webinar
May 18, 2020

Goal is to elucidate the genetic basis of OFC, one of the most common structural birth defects in humans worldwide

operationsmile.org

Lip and Palate Development

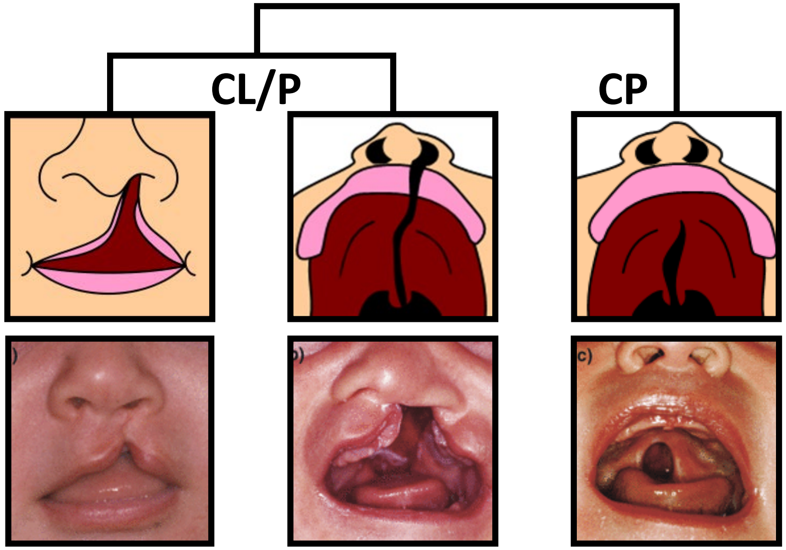
Lip Development **Palate Development**



Human Gestational Week

Dixon, Marazita, Beaty (2010), Nature Reviews Genetics

Orofacial Clefts (OFCs), Sub-Phenotypes



CL/P

CP

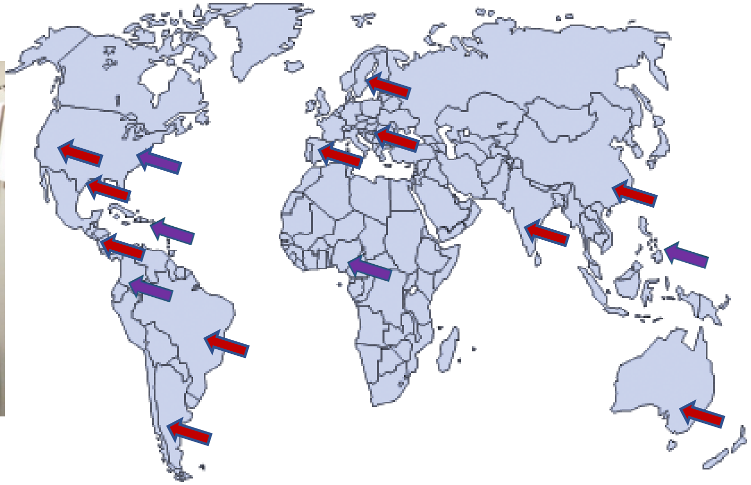
Cleft Lip (CL)

Cleft Lip and Palate (CLP)

Cleft Palate (CP)

Jugessur et al. (2009),
Oral Diseases

MULTI-ETHNIC, DEEP PHENOTYPING AND GENETIC RESOURCES



Houston, TX
Denver, CO
Iowa City, IA
St. Louis, MO
Puerto Rico
Guatemala
Colombia
Brazil
Argentina
Denmark
Spain
Hungary
India
China
Australia
Philippines
Nigeria

And more

Deep Phenotyping Protocol

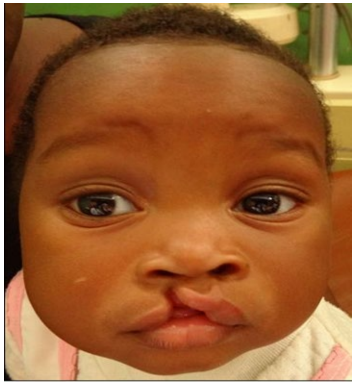
- **Questionnaires**
  - Demographics, personal and family medical history (eg birth defects, cancer, systems review), pregnancy history, developmental milestones
- **Physical examination**
  - Facial birth defects, Height/weight/head circumference, oral exam (plus imaging), limbs, hands/digits, speech sample for VPI perceptual screening
- **Imaging (to derive traits such as facial measurements)**
  - 3d facial image, intraoral photograph, palate video during speech (smCP), hand scans, ultrasound of upper lip region (OOM), dental casts plus 3d, upper and lower lip photos (lip pits)

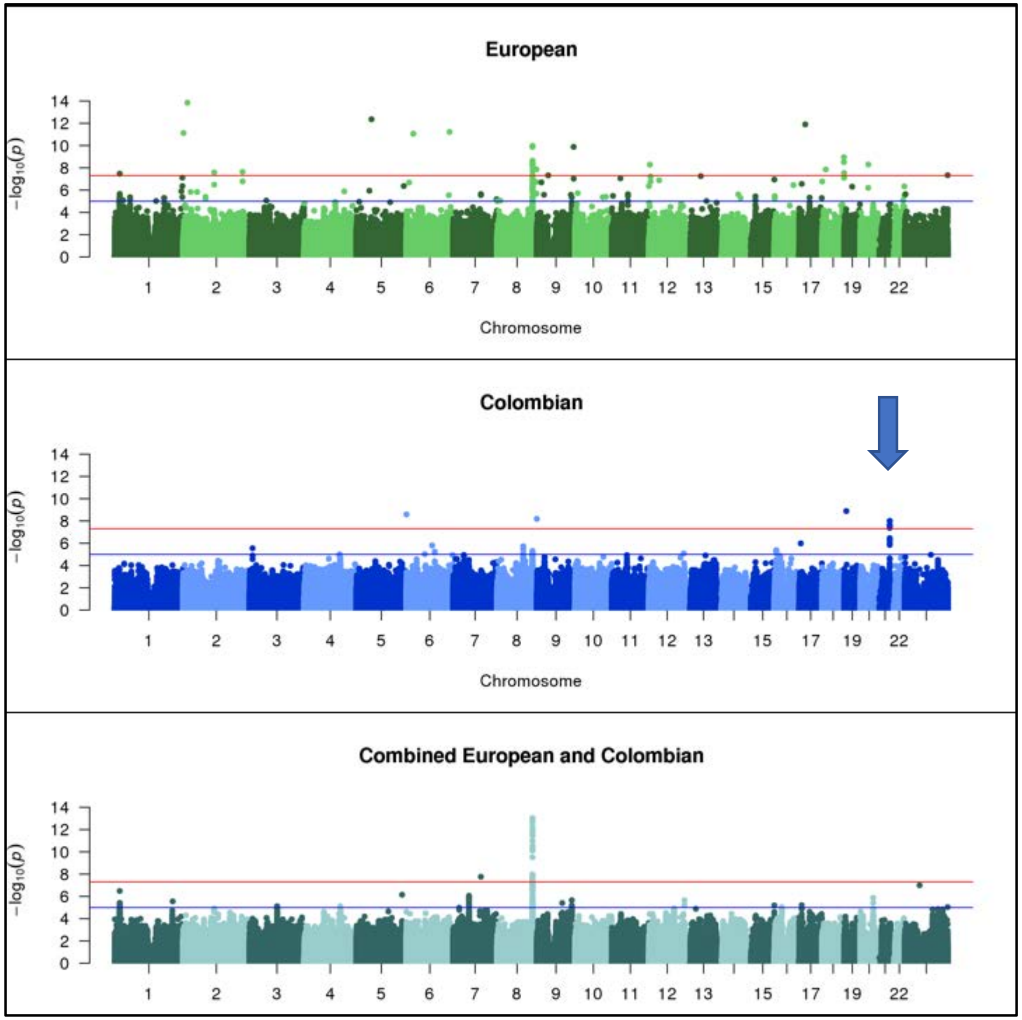
GMKF: OFC Whole Genome Sequencing

APPROVED (total = 1,413 OFC trios to date):

- European descent: 447
- Latin American (Colombia) : 265
- African (Nigeria and Ghana) : 137 trios
- Asian (Taiwan) : 124 trios
- Asian (Philippines) : 373 trios (COVID-19 delay)

- In review: additional Latin American trios





Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21.

Mukhopadhyay N, *et al*. *Human Genetics.* 139(2), 215-226. Epub ahead of print: 2019 Dec 17. doi: 10.1007/s00439-019-02099-1. URL: http://link.springer.com/article/10.1007/s00439-019-02099-1

Nandita Mukhopadhyay

AFRICAN AND ASIAN ANALYSES

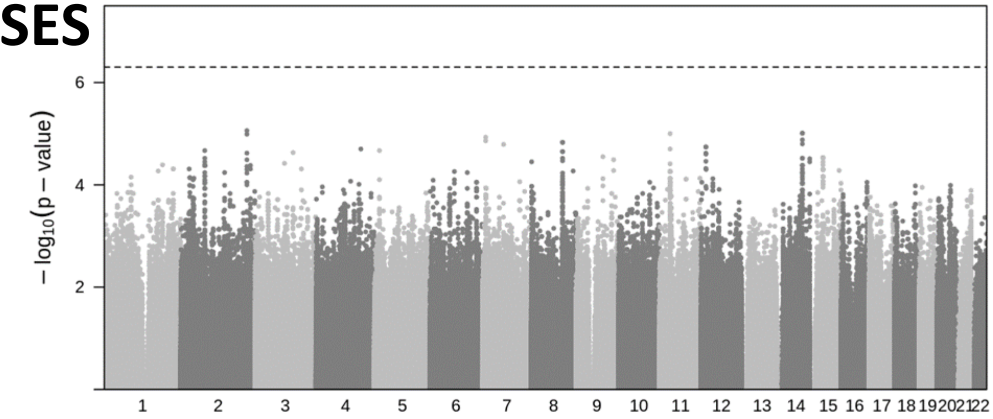
(Butaliand Baneatalyy)s

& 124 Taiwanese trios

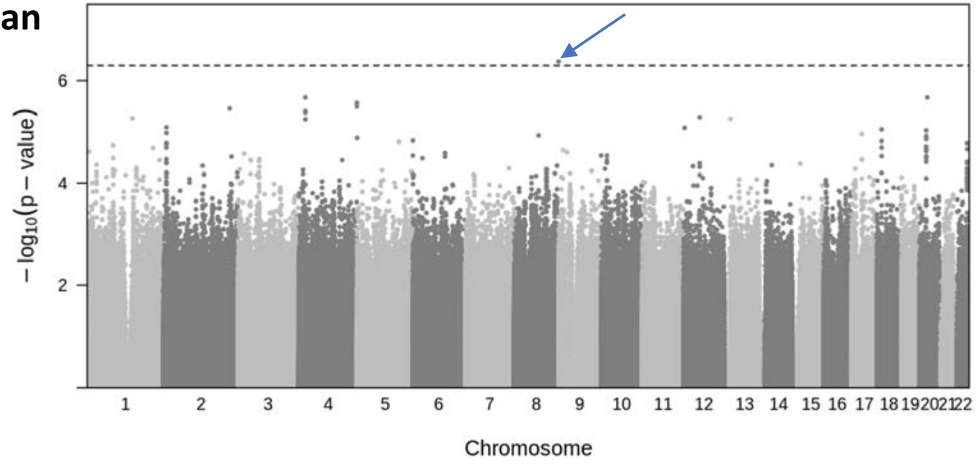
gTDT analysis of 130 African trios

Chr8q.24 signal replicated in African TDT and at the same E-06 as was reported in the African only GWAS for CL/P

Asian gTDT Analysis for CLP



African gTDT Analysis for CL/P



MANY THANKS TO OUR PARTICIPANTS WORLDWIDE

CCDG
CENTER FOR CRANIOFACIAL & DENTAL GENETICS

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

NIH
National Institute of Dental and Craniofacial Research

U of Pittsburgh:
Mary L. Marazita
Seth M. Weinberg
Eleanor Feingold
Nandita Mukhopadhyay
Ross Long (Lancaster)

Emory U:
Elizabeth J. Leslie
Madison Bishop
Pankaj Chopra
Michael Mortillo
Dave Cutler
Michael Epstein

U of Iowa:
Jeffrey C. Murray
Azeez Butali
Lina M. Moreno
Luz Consuelo Valencia-Ramirez
George L. Wehby
Andrew Lidral

Johns Hopkins University:
Terri Beaty
Ingo Ruczinski
Margaret Taub
Allen Scott
Jackie Hetmanski
Debashree Ray

Taiwan:
YH Wu-Chou
PK Chen

Africa:
NHGRI: Adebowale Adeyemo (NHGRI)
Kwame Nkurumah (Ghana)
Lord J.J Gowans (Ghana)
Lanre W Adeyemo (Nigeria)
Peter Mossey (U of Dundee)

Other:
Harrison Brand (Harvard)
Jacqueline T. Hecht (U of TX)
Frederic Deleyiannis (U of CO)
Carmencita Padilla (U of Manila)
Mauricio Arcos-Burgos
Andrew Czeizel
Eduardo Castilla
Ieda Orioli
Fernando Poletta

FUNDING: GMKF European sequencing: X01-HL132363, McDonnel 3U54HG003079-12S1; **GMKF other sequencing:** X01-HL136465 **(Colombia),** X01-HL140516 **(Africa and Taiwan),** X01- HD100701 **(Philippines);** Broad Inst. U24-HD090743. **Cohorts and analysis (Eur, Col, Phil):** R01-DE016148, R03-DE026469, R01-DE012472, U01-DD000295, R01-DE014581, R01-DE011931, R37-DE008559, R21-DE016930, R01-DE015667, R03-DE027193, R00-DE025060. **(Taiwan):**; R03-DE-027121; R01-DE-01458, U01-HG-018993 & U01-DE-020073, **(Africa)**: R00 DE022378, R01 DE028300; R01 DE016148, R37 DE-08559.

Uncovering the Genome-Wide Contribution of *De Novo* Mutations in Orofacial Clefts

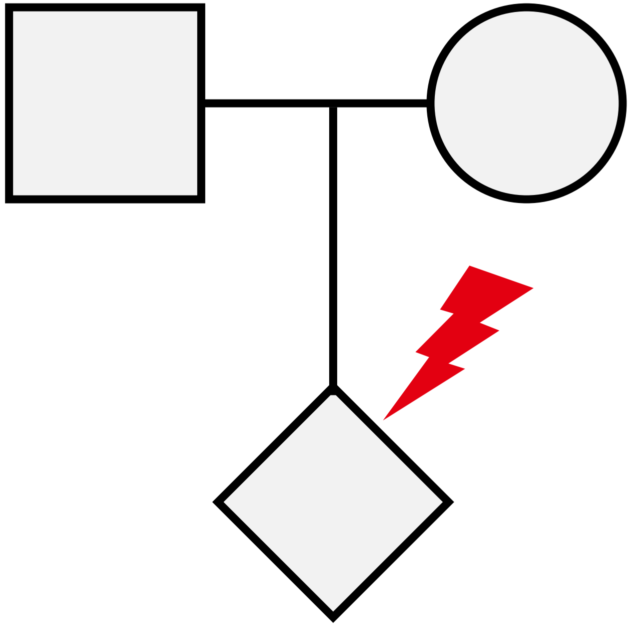
Elizabeth J. Leslie, PhD

Department of Human Genetics

Emory University

May 2020 Kids First Public Webinar

De novo mutations



***de novo* mutations per genome:**

~70-90 single nucleotide variants

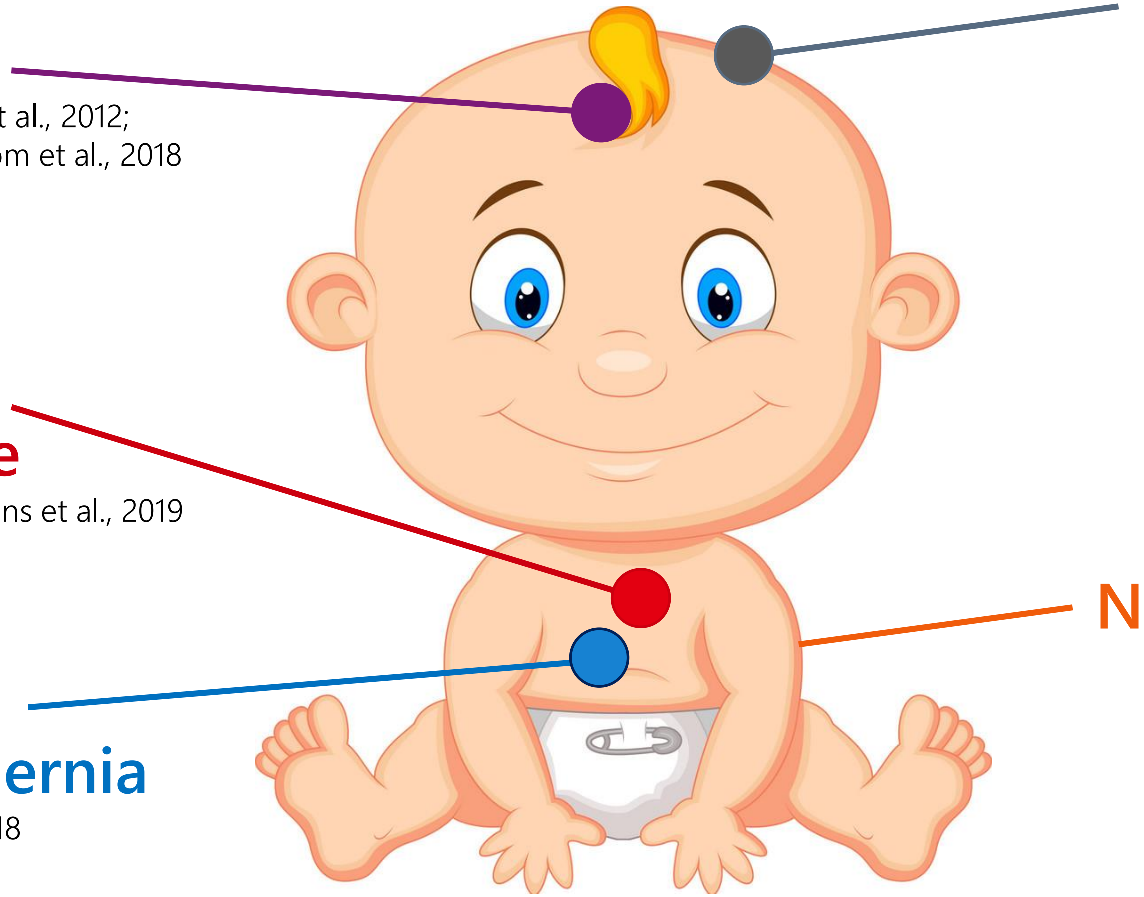
~6 insertion/deletions

0.02 copy number variants

~1 *de novo* mutation per exome

1 out of 20,563 protein-coding genes are hit per generation

De novo mutations are common causes of congenital and developmental anomalies



Autism

Neale et al., 2012; Sanders et al., 2012;
De Rubeis et al., 2014; Satterstrom et al., 2018

Craniosynostosis

Timberlake et al., 2017, 2018

Congenital Heart Disease

Homsy et al., 2015; Jin et al., 2017; Watkins et al., 2019

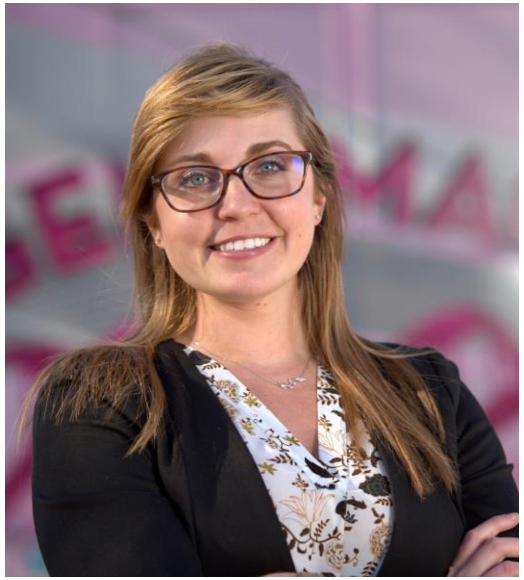
Neural Tube Defects

Lemay et al., 2015

Congenital Diaphragmatic Hernia

Yu et al., 2014; Qi et al., 2018

What is the role of de novo mutations in OFCs?



Madison Bishop, PhD

1. Are coding DNMs enriched in OFC cases?
2. What is the biological relevance of DNMs in OFC cases?
3. What is the clinical significance of DNMs in OFC cases?

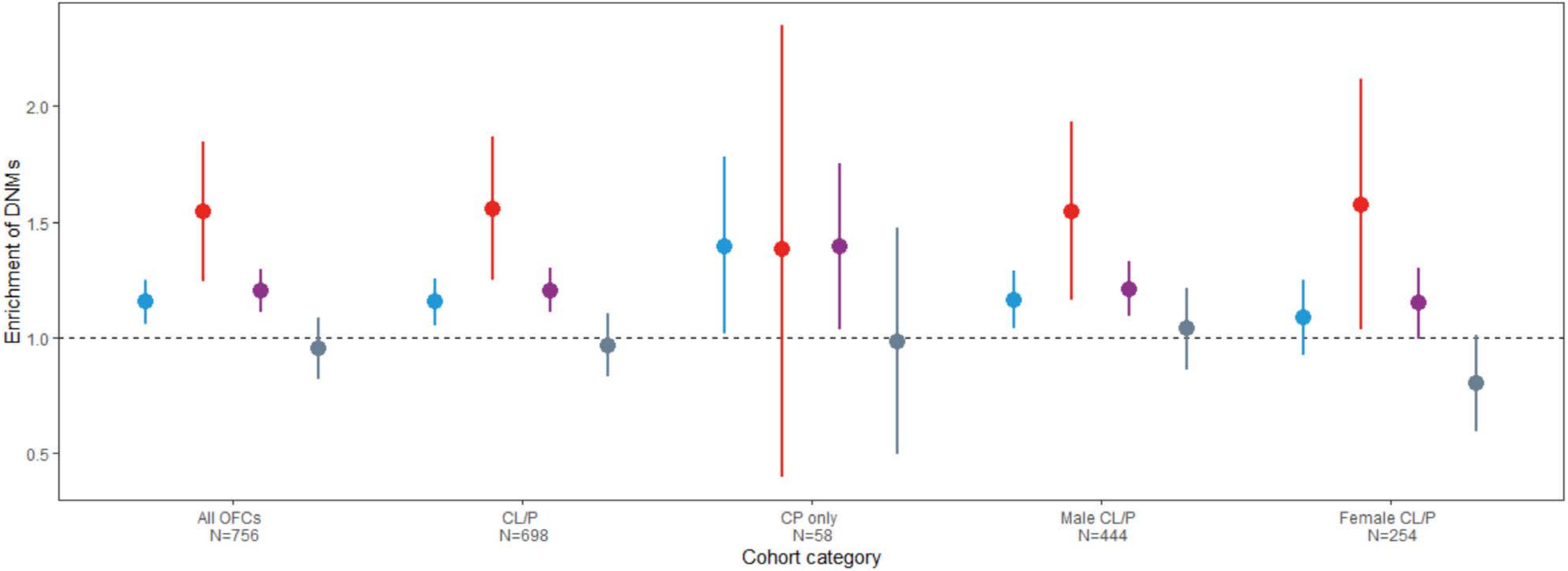
OFC de novos: by the numbers

- **756 trios (US/European, Colombian, and Taiwanese)**
  - 80 cleft lip only
  - 618 cleft lip and palate
  - 58 cleft palate only

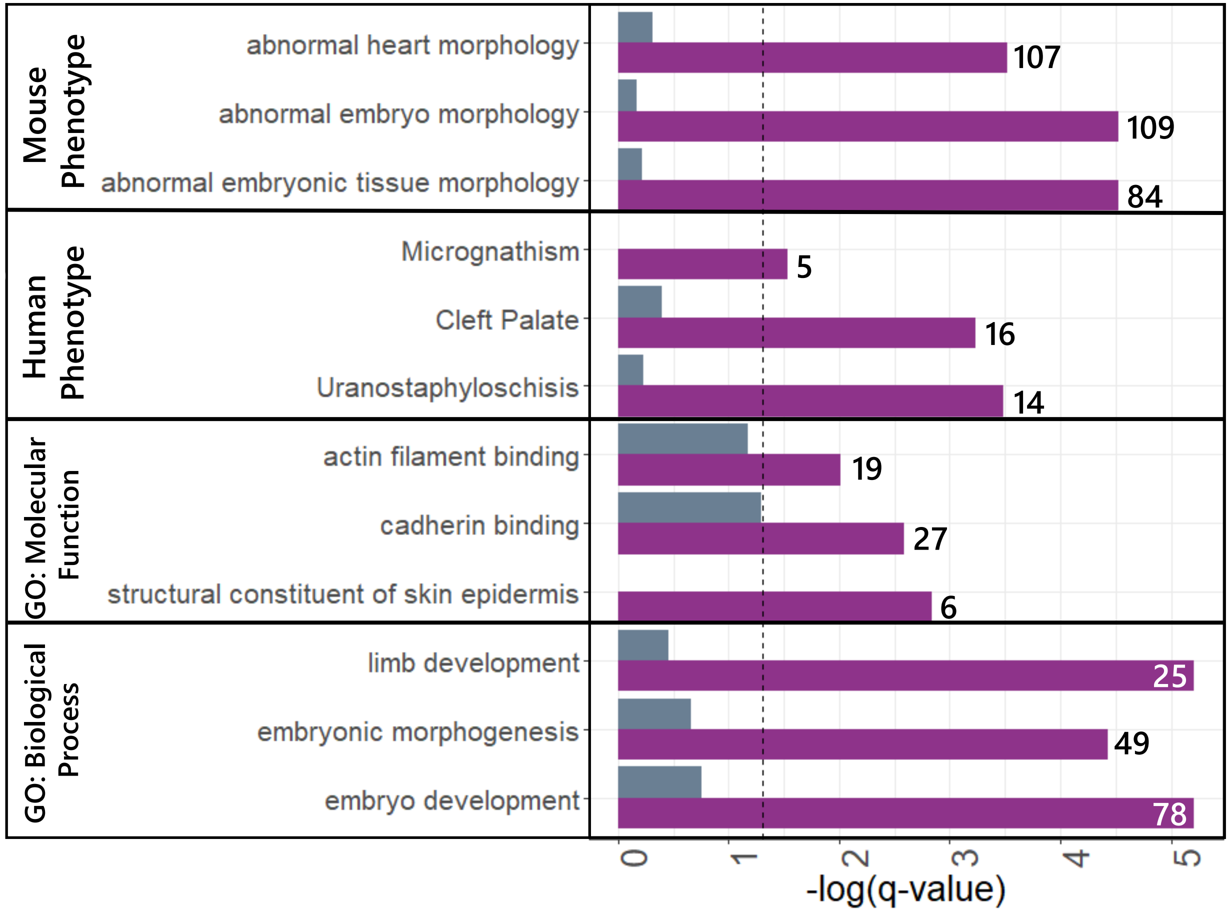
- **73,027 DNMs genome-wide**

- **862 coding DNMs**

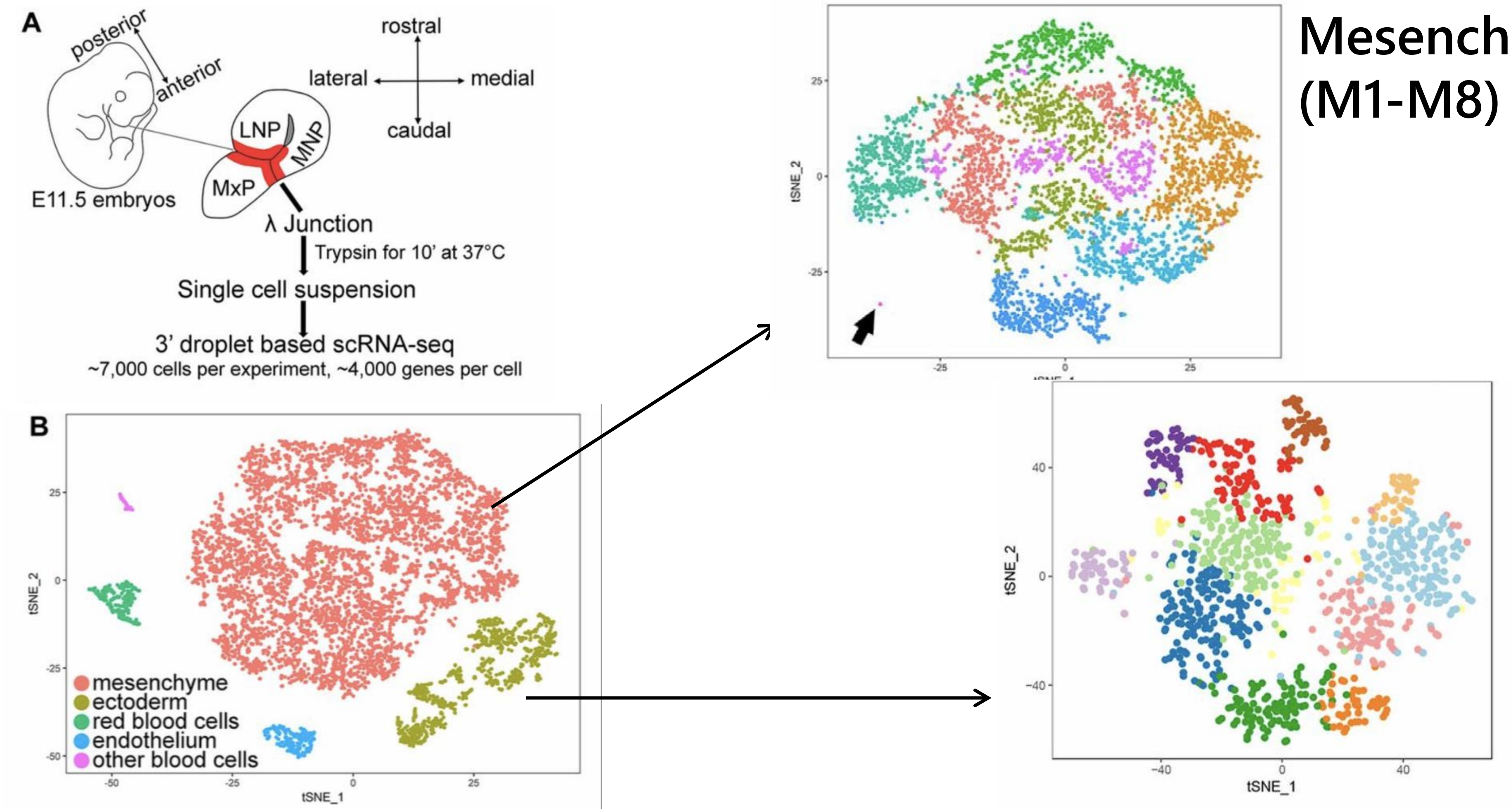
Protein-Altering DNMs are enriched in OFCs



DNMs are enriched in developmental genes



Craniofacial-Specific Annotations?



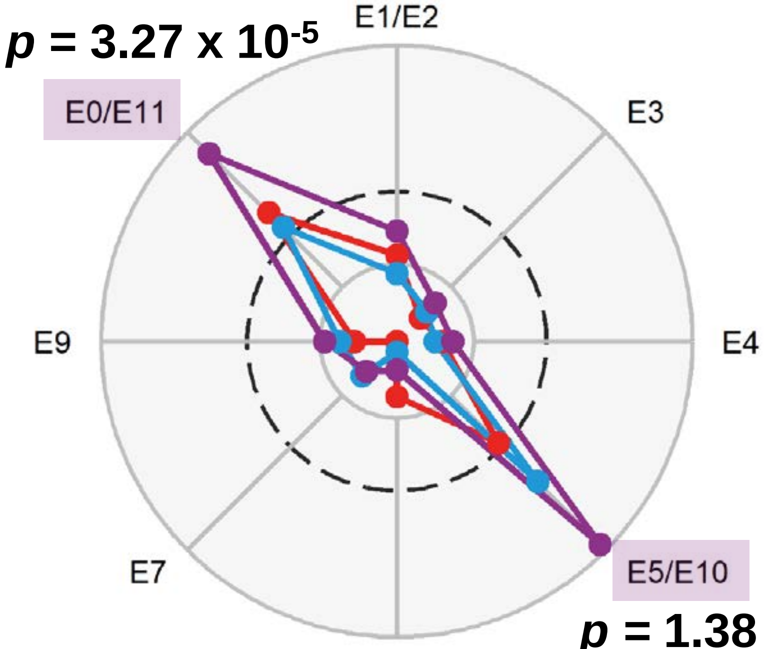
Mesenchyme Clusters (M1-M8)

Ectoderm Clusters (E1-E11)

Li et al., *Development*, 2019

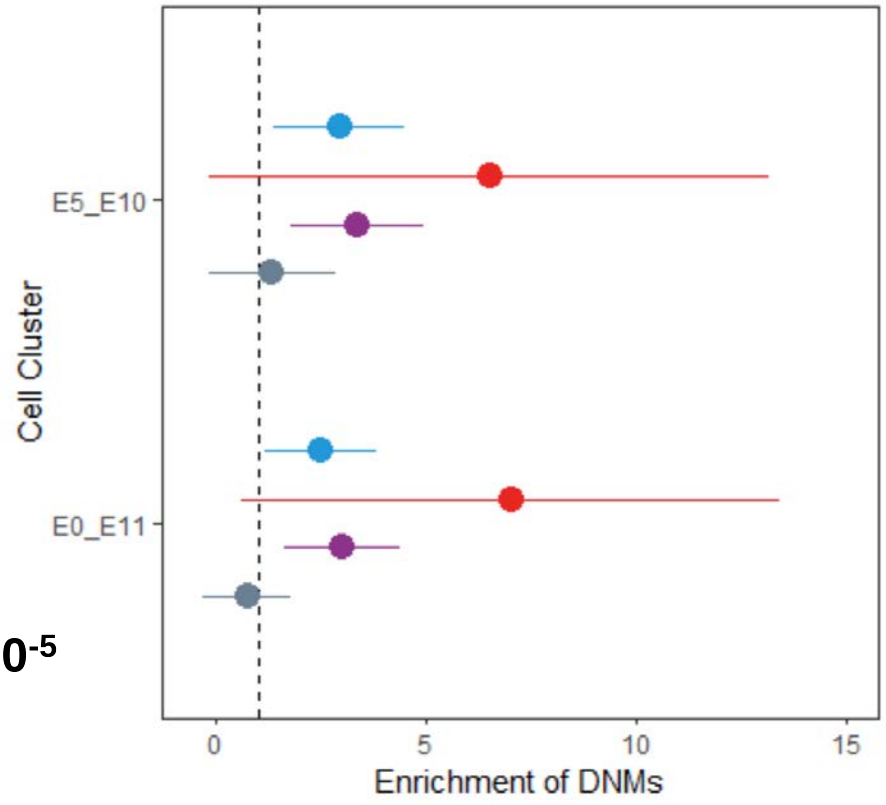
Excess of DNMs in genes expressed at point of fusion

p = 3.27 x 10^{-5}

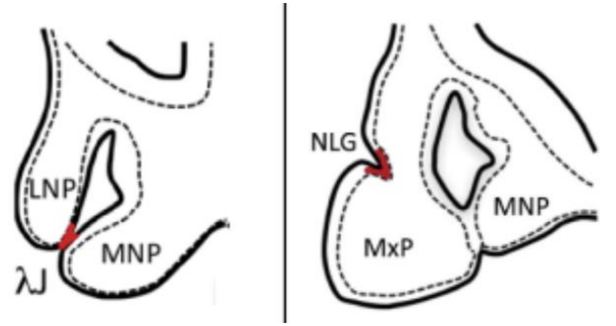


p = 1.38 x 10^{-5}

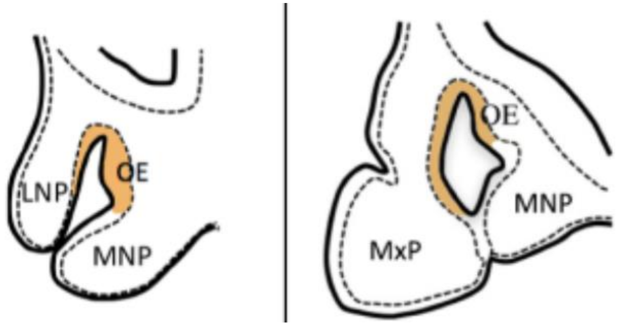
E11.5 Ectoderm Clusters



Nasal process fusion zone

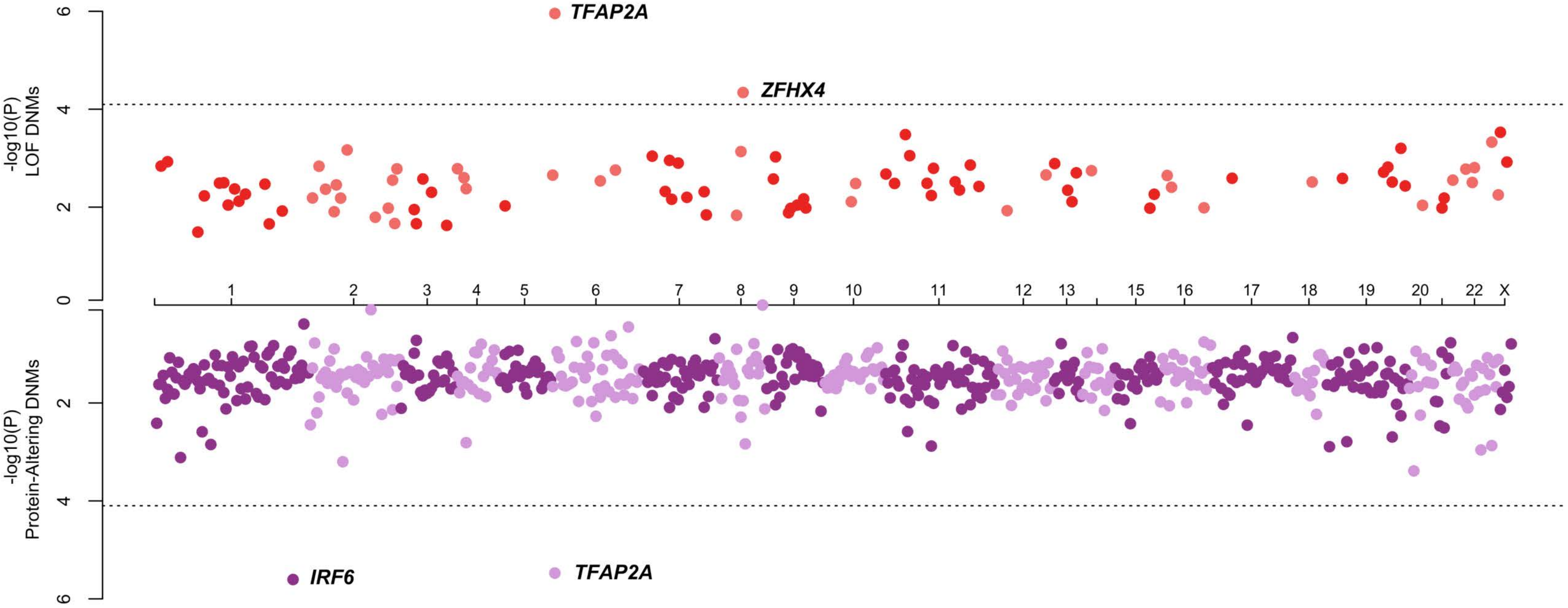


Olfactory epithelium

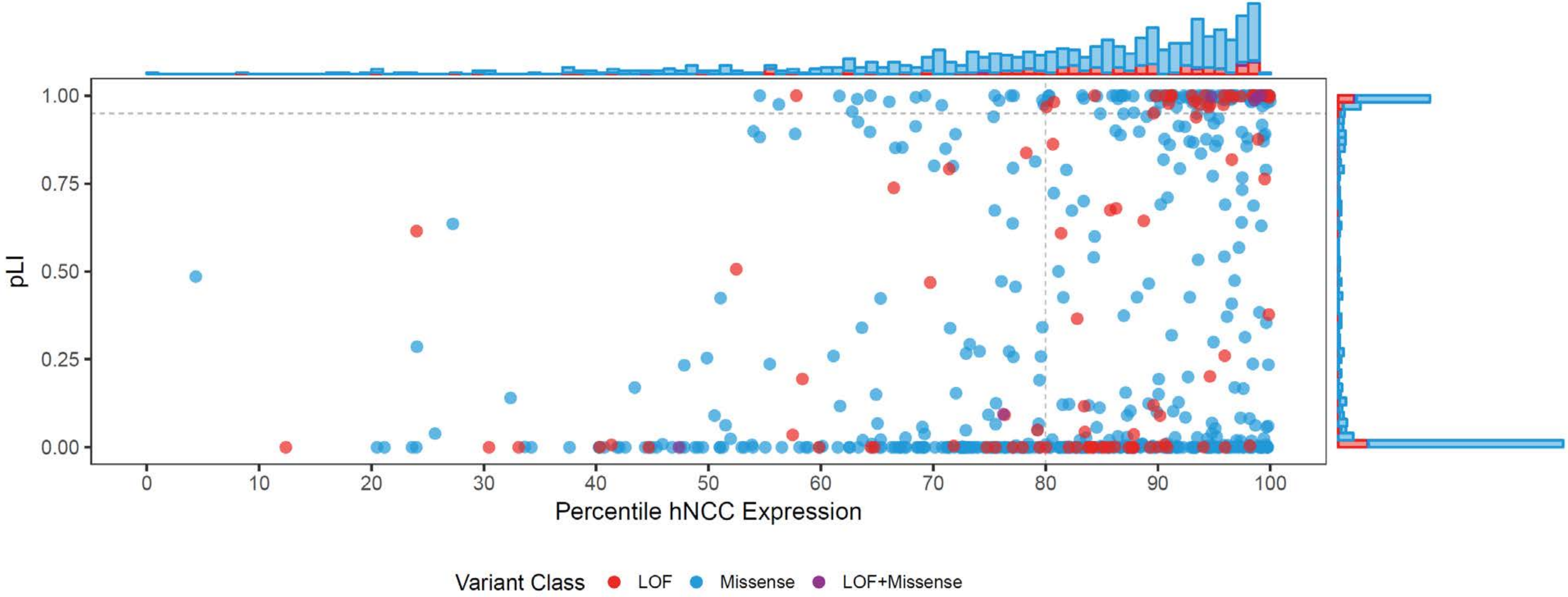


Li et al., *Development*, 2019

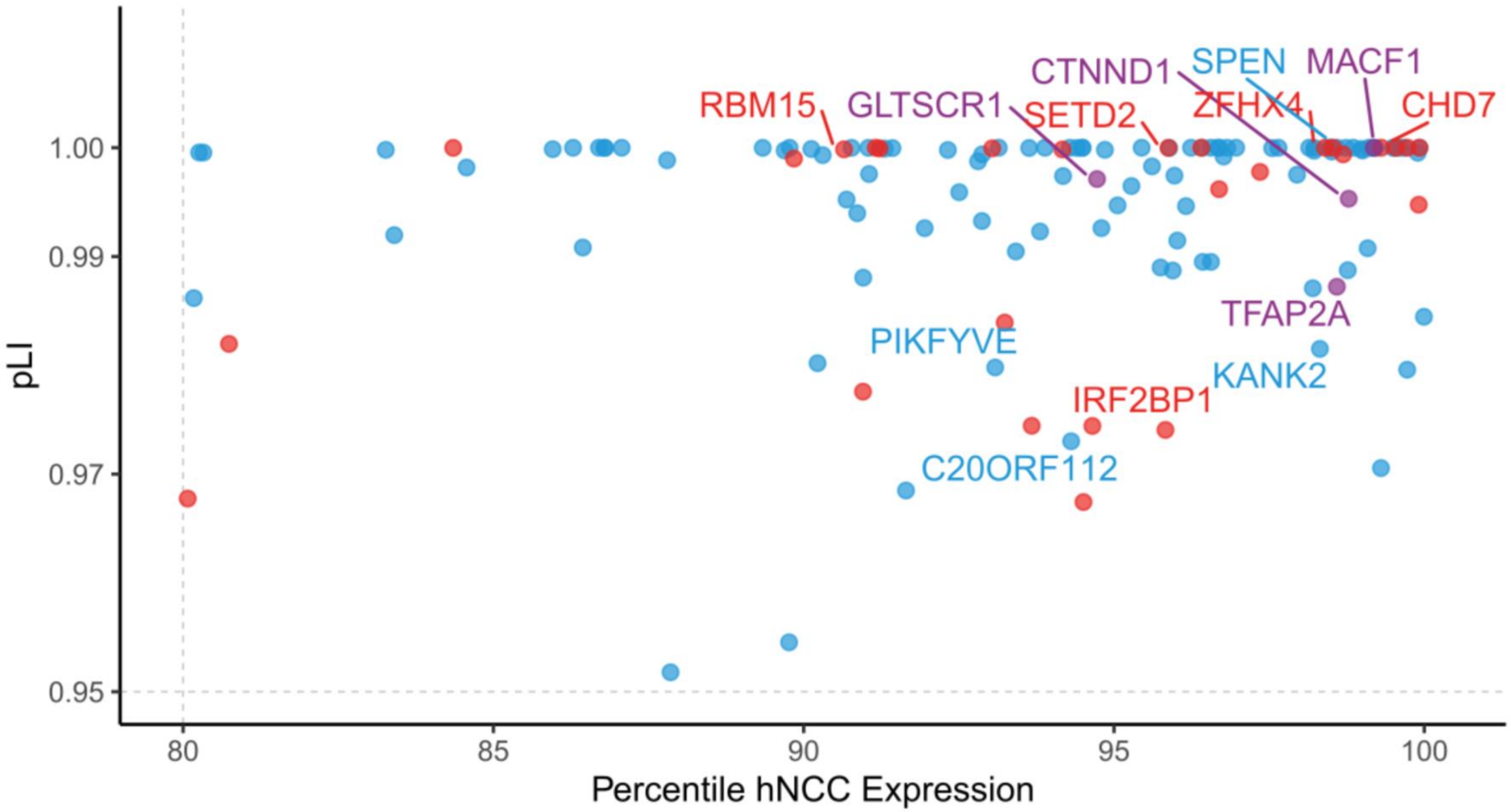
De novo mutations in *IRF6*, *TFAP2A*, and *ZFHX4* are associated with OFCs

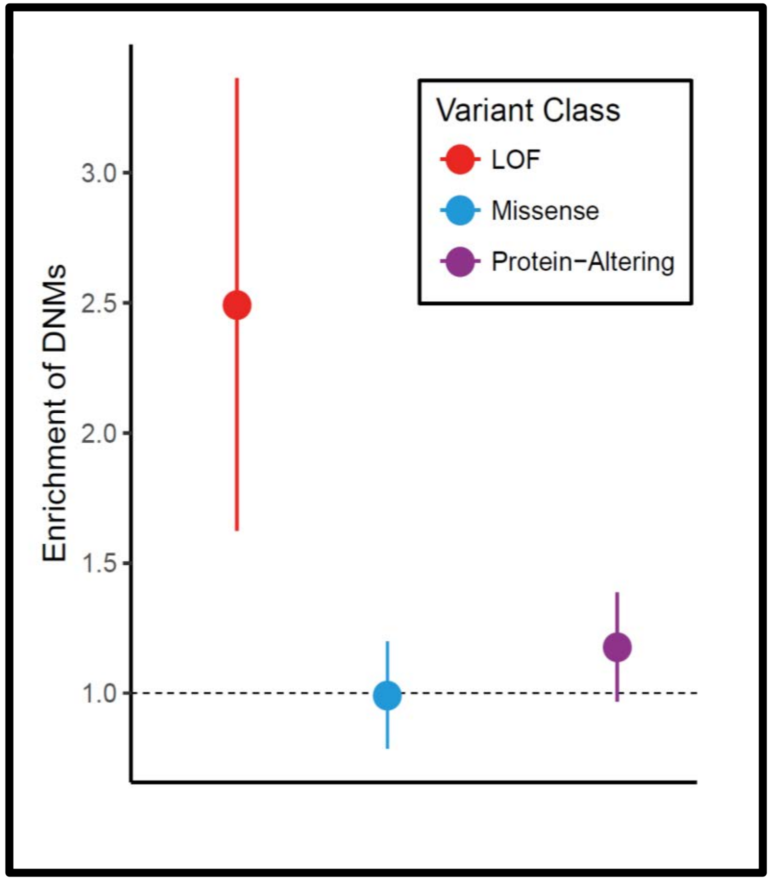


Excess of DNMs in cranial neural crest genes



Excess of DNMs in cranial neural crest genes





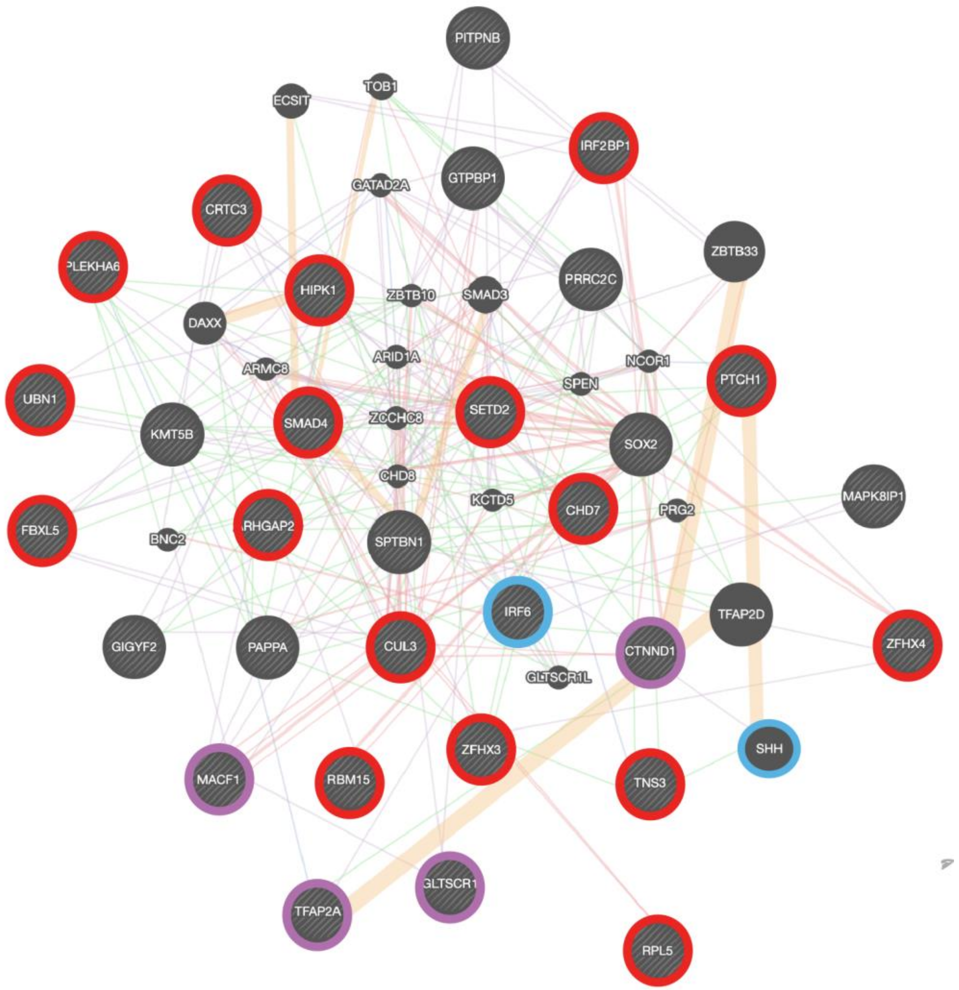
DNMs in SOX2-interactome

29 genes with Loss of Function DNMs + pLI > 0.95 + top 20% hNCC expression

8 genes interact with SOX2 (FDR p = 9.5×10^{-4})
MACF1, RBM15, SETD2, CHD7, CTNND1, ZFHX4, IRF2BP1,TFAP2A

126 genes with Protein-Altering DNMs + pLI > 0.95 + top 20% hNCC expression

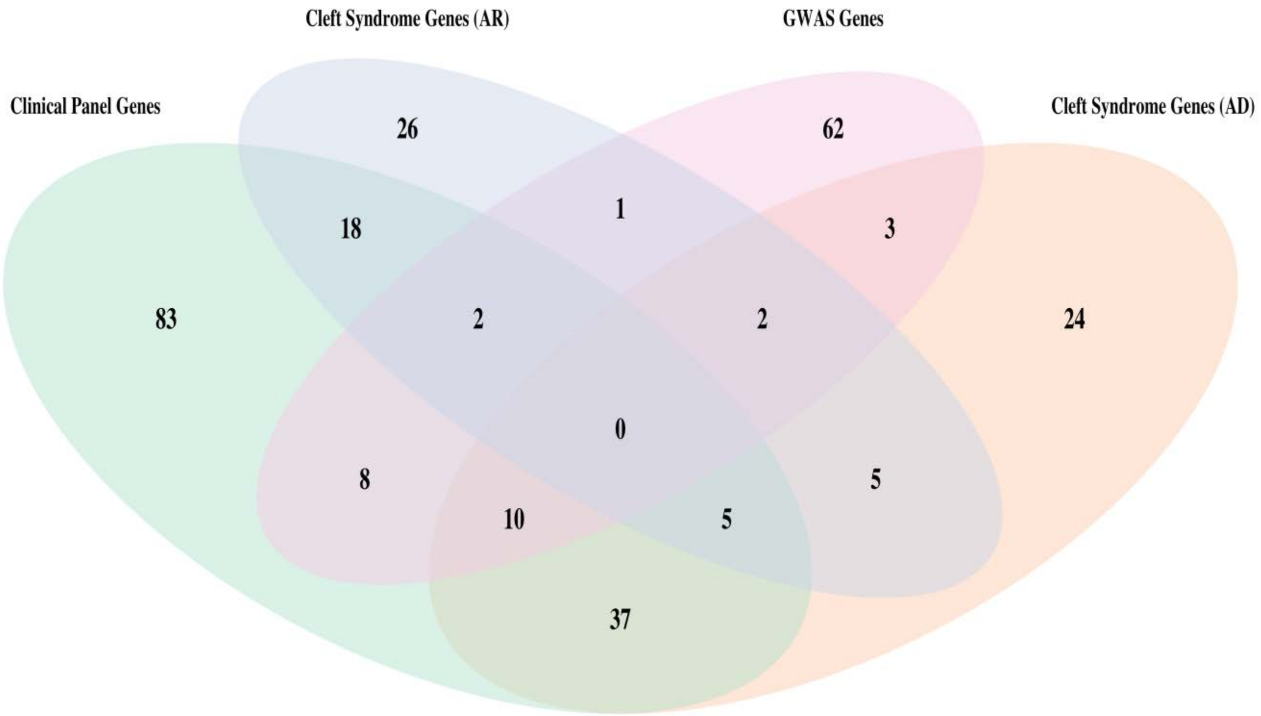
16 genes interact with SOX2 (FDR p = 5.1×10^{-5})
TCF20, RFX1, PPP2R5D, MDN1, NFIA, SPEN, NIPBL, ZNF292



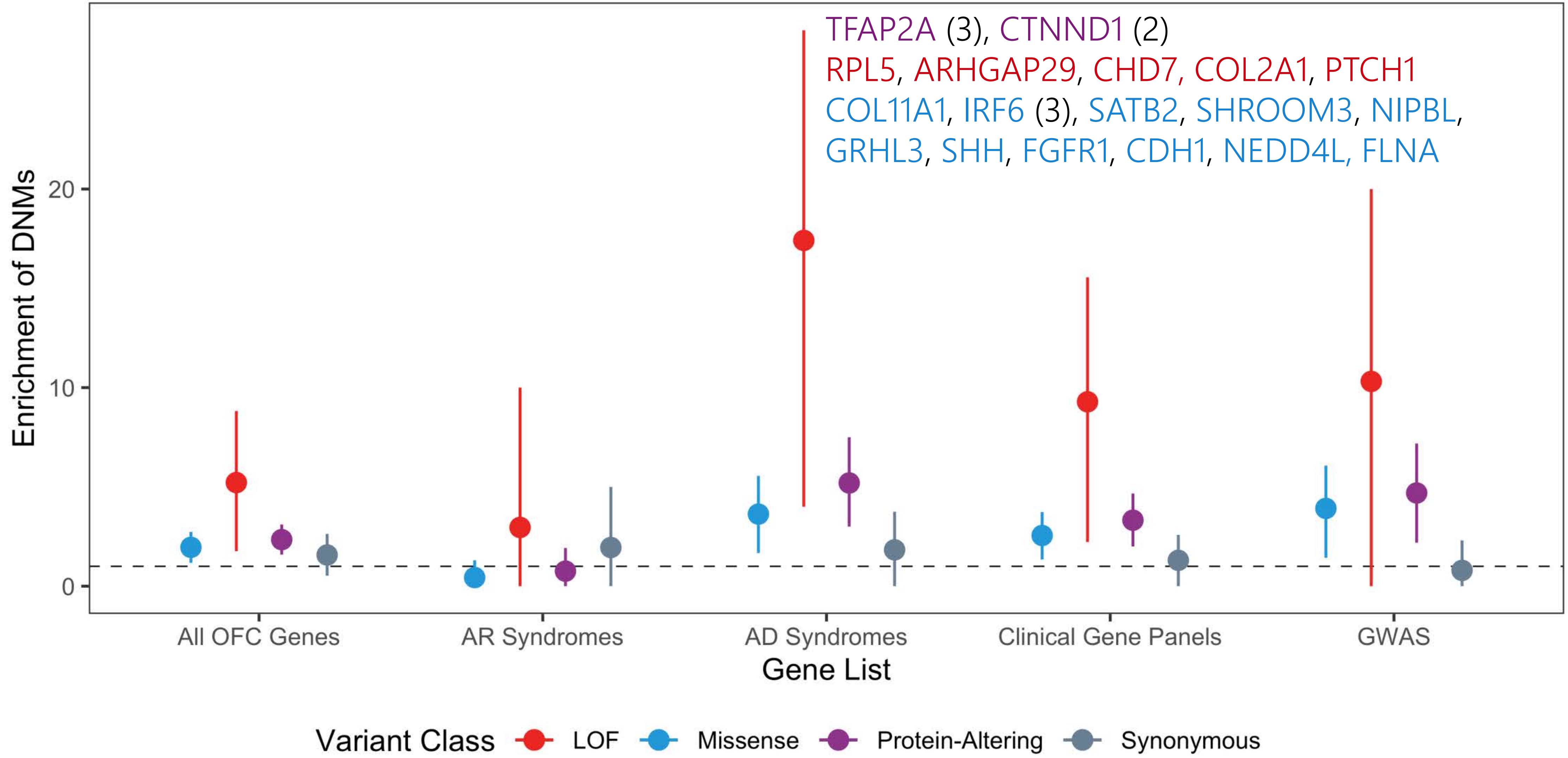
Towards a clinical gene panel for OFCs?

Curated an OFC gene list containing 289 genes

- PubMed
- OMIM
- ClinVar
- Existing gene panels
  - NCBI Genetic Testing Registry: 6 genes
  - Fulgent: 24 genes
  - Prevention Genetics: 172 genes
  - Blueprint Genetics: 22 genes



DNMs are enriched in clinically-relevant genes



TFAP2A (3), CTNND1 (2)
RPL5, ARHGAP29, CHD7, COL2A1, PTCH1
COL11A1, IRF6 (3), SATB2, SHROOM3, NIPBL,
GRHL3, SHH, FGFR1, CDH1, NEDD4L, FLNA

Summary

- We identified an excess of protein-altering DNMs in OFC trios (~1.2x more than expected)

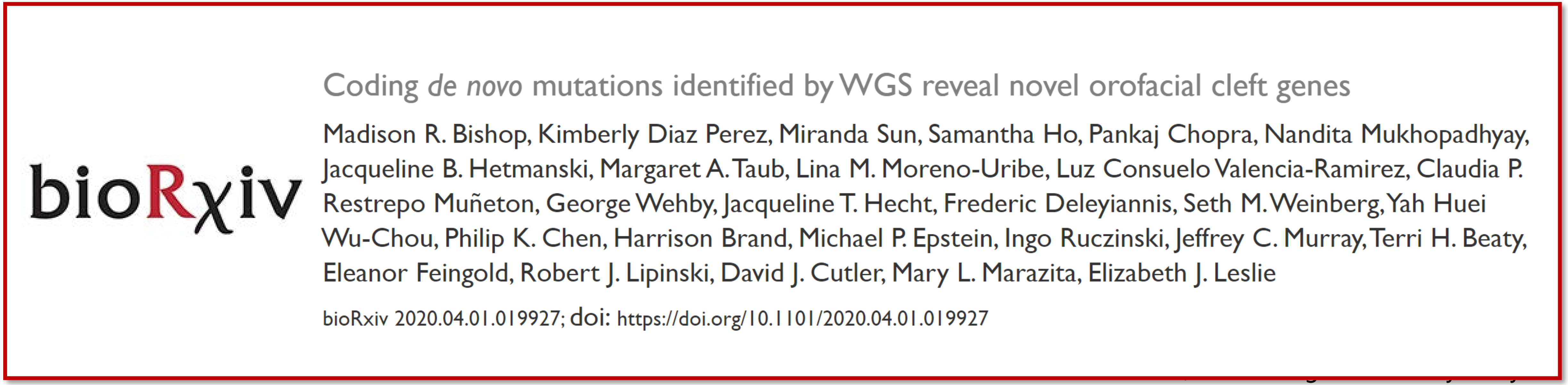
- DNMs are in biologically relevant genes:
  - marker genes expressed in cells at the point of fusion of developing lip
  - genes in the top 20% in human cranial neural crest cells that are constrained to mutation

- DNMs are found in clinically relevant genes (~18x more than expected in AD OFC genes)

- 3 genes (*IRF6*, *TFAP2A*, *ZFHX4*) had individual excesses of DNMs

- ZFHX4 is a novel gene for OFCs

Acknowledgments



bioRxiv

Coding *de novo* mutations identified by WGS reveal novel orofacial cleft genes

Madison R. Bishop, Kimberly Diaz Perez, Miranda Sun, Samantha Ho, Pankaj Chopra, Nandita Mukhopadhyay, Jacqueline B. Hetmanski, Margaret A. Taub, Lina M. Moreno-Uribe, Luz Consuelo Valencia-Ramirez, Claudia P. Restrepo Muñeton, George Wehby, Jacqueline T. Hecht, Frederic Deleyiannis, Seth M. Weinberg, Yah Huei Wu-Chou, Philip K. Chen, Harrison Brand, Michael P. Epstein, Ingo Ruczinski, Jeffrey C. Murray, Terri H. Beaty, Eleanor Feingold, Robert J. Lipinski, David J. Cutler, Mary L. Marazita, Elizabeth J. Leslie

bioRxiv 2020.04.01.019927; doi: https://doi.org/10.1101/2020.04.01.019927

The Leslie Lab

- Grace Carlock
- **Kim Diaz-Perez**
- Courtney Willett
- Dan Chang
- **Madison Bishop, PhD**

- Sarah Curtis, PhD
- Kelly Manning
- **Samantha Ho**
- Sydney Chung
- Shade Awoniyi

others

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

Funding: R03 DE027193
R00 DE025060
X01 HL132363
X01 HL136465

SV CALLING IN OROFACIAL CLEFTS

Harrison Brand
Assistant Professor
MGH, Harvard Medical School, & Broad Institute

May 18th, 2020

Introduction

- Impact of structural variation (SV) in non-syndromic forms of orofacial clefts (OFC) is largely uncharacterized

- We applied GATK-SV, our computational SV discovery pipeline, to 2,746 WGS samples that passed quality control
  - ➢ Includes 837 complete trios for *de novo* SV analysis

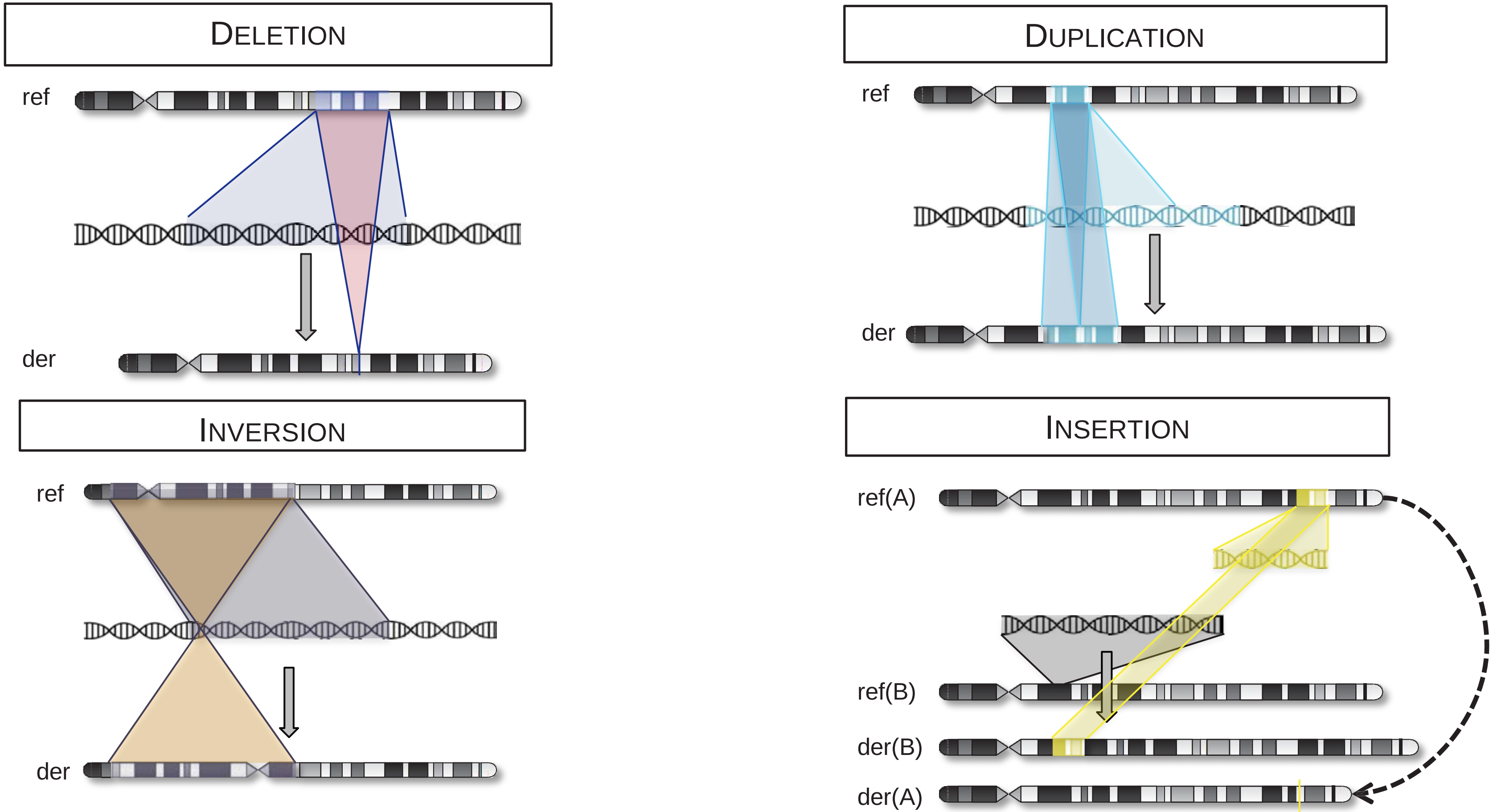
- OFC samples from 3 GMKF studies (140543, 136465, 132377) and 4 distinct populations (African, Asian, Latino, Caucasian)

- GATK-SV recently applied to 14,891 individuals in the gnomAD reference database (Collins*, Brand*, *et al. Nature,* in press)

Structural Variation Background

Structural Variation

Four basic classes of SV in the human genome



DELETION

DUPLICATION

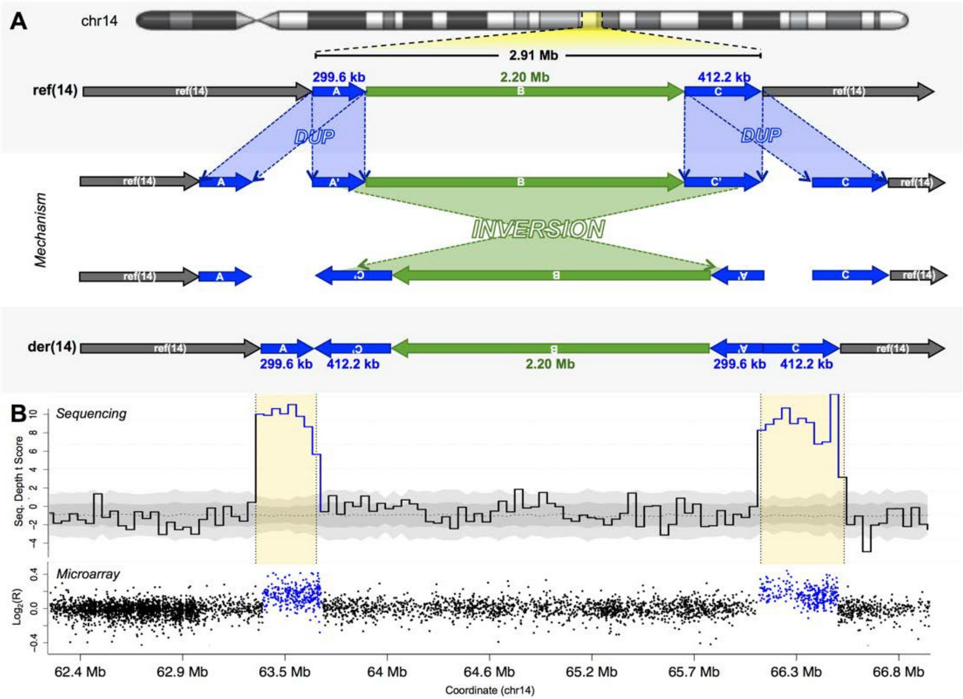
INVERSION

INSERTION

Complex SVs

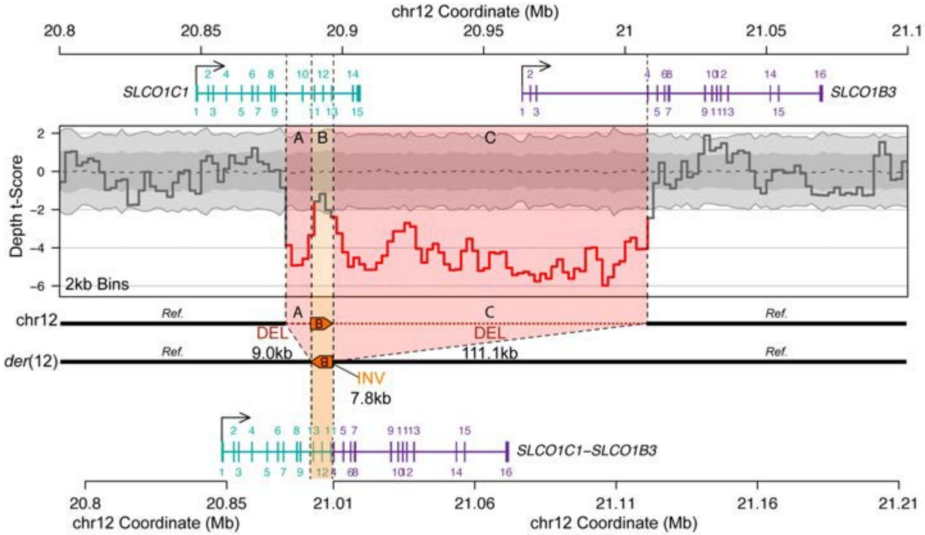
Complex SVs are comprised of combinations of the four basic SV classes

Paired-duplication inversion (dupINVdup)



Brand *et al.*, *Am. J. Hum. Genet.* (2014 & 2015)

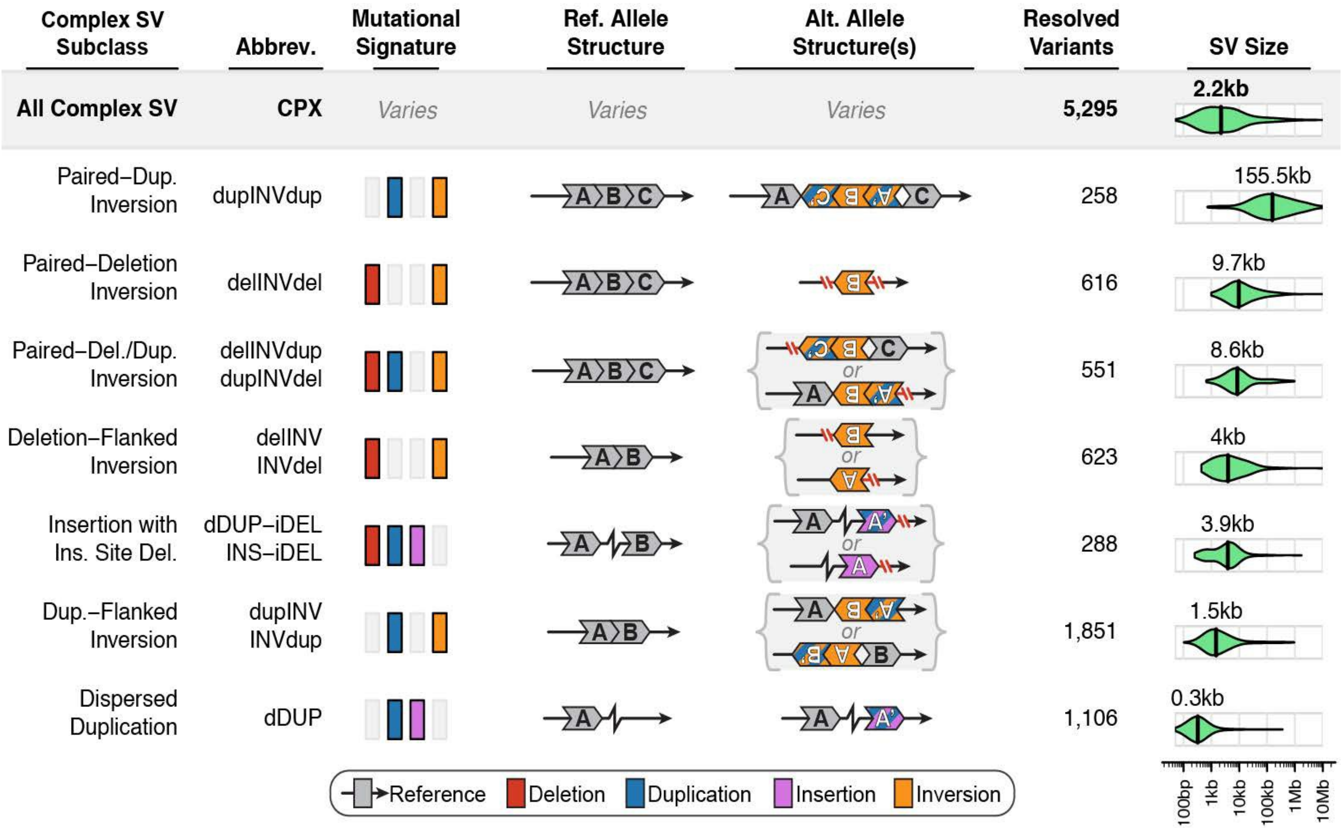
Paired-deletion inversion (delINVdel)



Collins, Brand *et al.*, *Genome Biology* (2017)

Abundance of Complex SVs in the Genome

Complex SVs are surprisingly abundant in the genome



Complex SV Subclass	Abbrev.	Mutational Signature	Ref. Allele Structure	Alt. Allele Structure(s)	Resolved Variants	SV Size
All Complex SV	CPX	Varies	Varies	Varies	5,295	2.2kb
Paired–Dup. Inversion	dupINVdup	🟦🟧	A B C	A C B A C	258	155.5kb
Paired–Deletion Inversion	delINVdel	🟥🟧	A B C	B	616	9.7kb
Paired–Del./Dup. Inversion	delINVdup dupINVdel	🟥🟦🟧	A B C	C B C or A B A	551	8.6kb
Deletion–Flanked Inversion	delINV INVdel	🟥🟧	A B	B or A	623	4kb
Insertion with Ins. Site Del.	dDUP–iDEL INS–iDEL	🟥🟦🟪	A B	A A' or A	288	3.9kb
Dup.–Flanked Inversion	dupINV INVdup	🟦🟧	A B	A B A or B A B	1,851	1.5kb
Dispersed Duplication	dDUP	🟦🟪	A	A A'	1,106	0.3kb

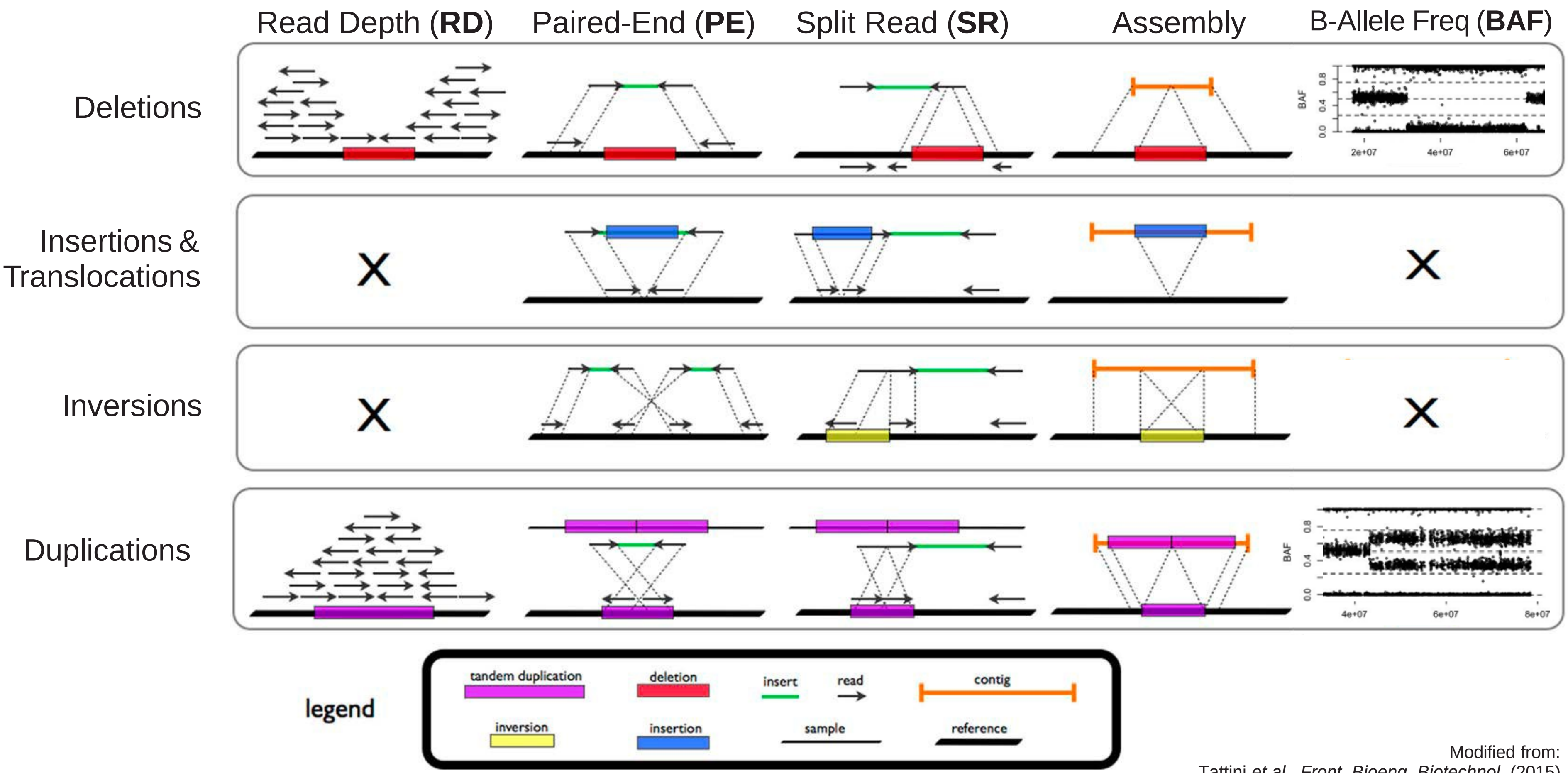
Complex SV counts in the gnomAD SV cohort

Collins, Brand et al., Nature 2020, in press

METHODS

SV Discovery in Whole Genome Sequencing (WGS)

Different classes of SVs leave distinct signatures in Illumina WGS data



Read Depth (**RD**)

Paired-End (**PE**)

Split Read (**SR**)

Assembly

B-Allele Freq (**BAF**)

Deletions

Insertions & Translocations

Inversions

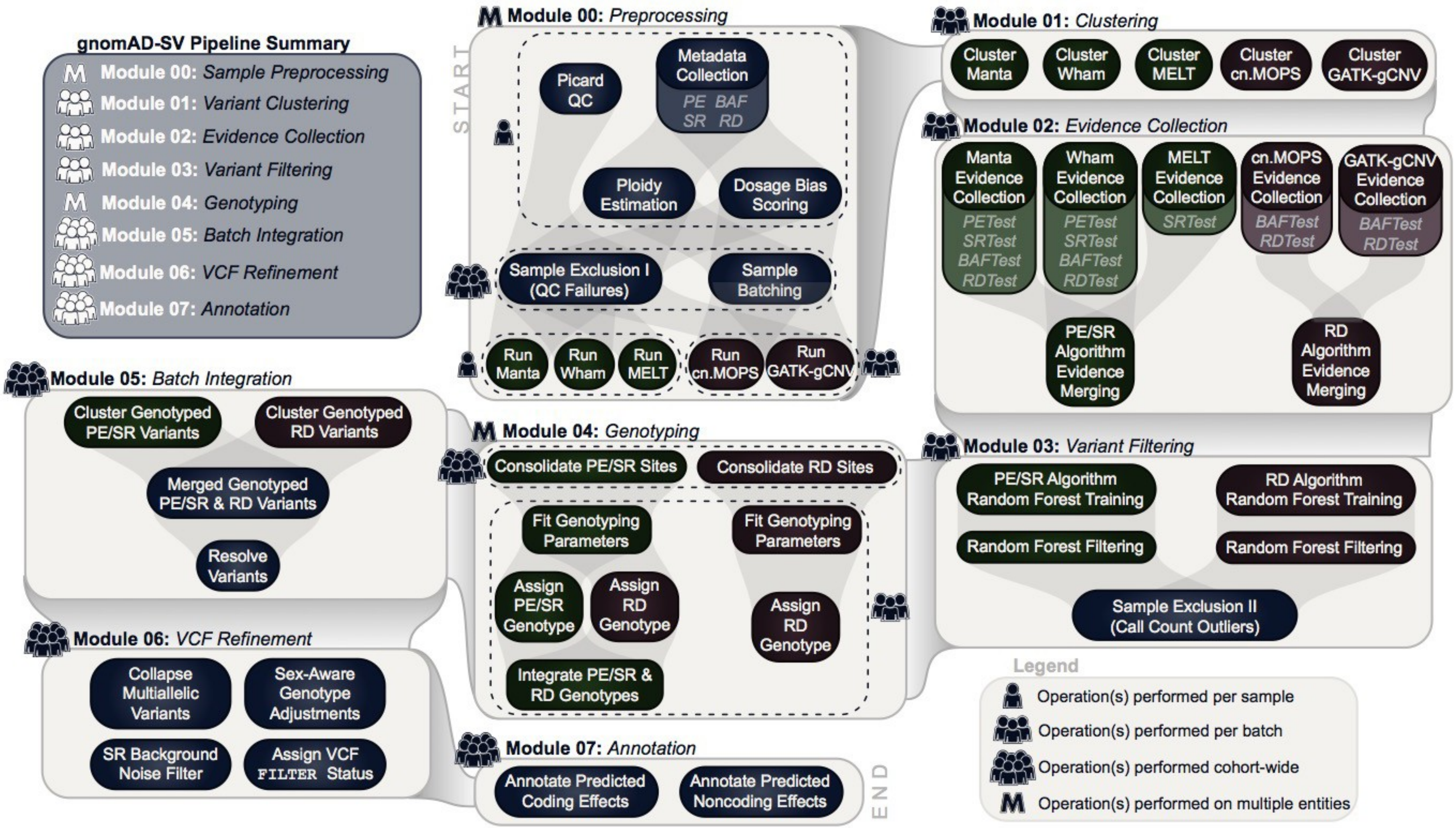
Duplications

Modified from:
Tattini *et al.*, *Front. Bioeng. Biotechnol.* (2015)

GATK-SV: Cloud Enabled SV Pipeline

gnomAD-SV Pipeline Summary

- M **Module 00:** *Sample Preprocessing*
- **Module 01:** *Variant Clustering*
- **Module 02:** *Evidence Collection*
- **Module 03:** *Variant Filtering*
- M **Module 04:** *Genotyping*
- **Module 05:** *Batch Integration*
- **Module 06:** *VCF Refinement*
- **Module 07:** *Annotation*



START

M **Module 00:** *Preprocessing*

Picard QC

Metadata Collection: PE BAF SR RD

Ploidy Estimation

Dosage Bias Scoring

Sample Exclusion I (QC Failures)

Sample Batching

Run Manta, Run Wham, Run MELT, Run cn.MOPS, Run GATK-gCNV

Module 01: *Clustering*

Cluster Manta, Cluster Wham, Cluster MELT, Cluster cn.MOPS, Cluster GATK-gCNV

Module 02: *Evidence Collection*

Manta Evidence Collection: PETest SRTest BAFTest RDTest

Wham Evidence Collection: PETest SRTest BAFTest RDTest

MELT Evidence Collection: SRTest

cn.MOPS Evidence Collection: BAFTest RDTest

GATK-gCNV Evidence Collection: BAFTest RDTest

PE/SR Algorithm Evidence Merging

RD Algorithm Evidence Merging

Module 03: *Variant Filtering*

PE/SR Algorithm Random Forest Training

RD Algorithm Random Forest Training

Random Forest Filtering

Random Forest Filtering

Sample Exclusion II (Call Count Outliers)

M **Module 04:** *Genotyping*

Consolidate PE/SR Sites

Consolidate RD Sites

Fit Genotyping Parameters

Fit Genotyping Parameters

Assign PE/SR Genotype

Assign RD Genotype

Assign RD Genotype

Integrate PE/SR & RD Genotypes

Module 05: *Batch Integration*

Cluster Genotyped PE/SR Variants

Cluster Genotyped RD Variants

Merged Genotyped PE/SR & RD Variants

Resolve Variants

Module 06: *VCF Refinement*

Collapse Multiallelic Variants

Sex-Aware Genotype Adjustments

SR Background Noise Filter

Assign VCF `FILTER` Status

Module 07: *Annotation*

Annotate Predicted Coding Effects

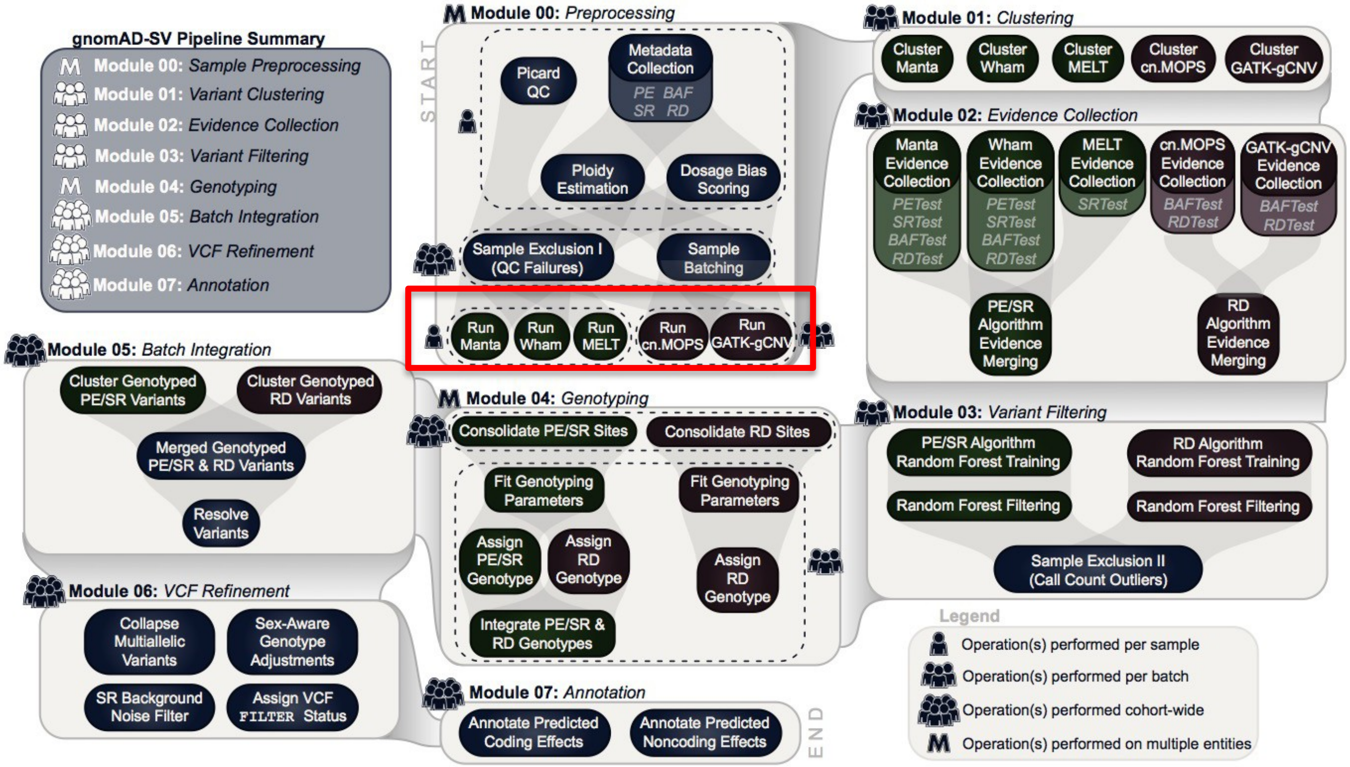
Annotate Predicted Noncoding Effects

END

Legend

- Operation(s) performed per sample
- Operation(s) performed per batch
- Operation(s) performed cohort-wide
- M Operation(s) performed on multiple entities

GATK-SV: Cloud Enabled SV Pipeline



gnomAD-SV Pipeline Summary

- Module 00: *Sample Preprocessing*
- Module 01: *Variant Clustering*
- Module 02: *Evidence Collection*
- Module 03: *Variant Filtering*
- Module 04: *Genotyping*
- Module 05: *Batch Integration*
- Module 06: *VCF Refinement*
- Module 07: *Annotation*

START

Module 00: *Preprocessing*

Picard QC

Metadata Collection
PE BAF
SR RD

Ploidy Estimation

Dosage Bias Scoring

Sample Exclusion I (QC Failures)

Sample Batching

Run Manta · Run Wham · Run MELT · Run cn.MOPS · Run GATK-gCNV

Module 01: *Clustering*

Cluster Manta · Cluster Wham · Cluster MELT · Cluster cn.MOPS · Cluster GATK-gCNV

Module 02: *Evidence Collection*

Manta Evidence Collection: *PETest SRTest BAFTest RDTest*

Wham Evidence Collection: *PETest SRTest BAFTest RDTest*

MELT Evidence Collection: *SRTest*

cn.MOPS Evidence Collection: *BAFTest RDTest*

GATK-gCNV Evidence Collection: *BAFTest RDTest*

PE/SR Algorithm Evidence Merging

RD Algorithm Evidence Merging

Module 03: *Variant Filtering*

PE/SR Algorithm Random Forest Training

RD Algorithm Random Forest Training

Random Forest Filtering

Random Forest Filtering

Sample Exclusion II (Call Count Outliers)

Module 04: *Genotyping*

Consolidate PE/SR Sites

Consolidate RD Sites

Fit Genotyping Parameters

Fit Genotyping Parameters

Assign PE/SR Genotype

Assign RD Genotype

Assign RD Genotype

Integrate PE/SR & RD Genotypes

Module 05: *Batch Integration*

Cluster Genotyped PE/SR Variants

Cluster Genotyped RD Variants

Merged Genotyped PE/SR & RD Variants

Resolve Variants

Module 06: *VCF Refinement*

Collapse Multiallelic Variants

Sex-Aware Genotype Adjustments

SR Background Noise Filter

Assign VCF `FILTER` Status

Module 07: *Annotation*

Annotate Predicted Coding Effects

Annotate Predicted Noncoding Effects

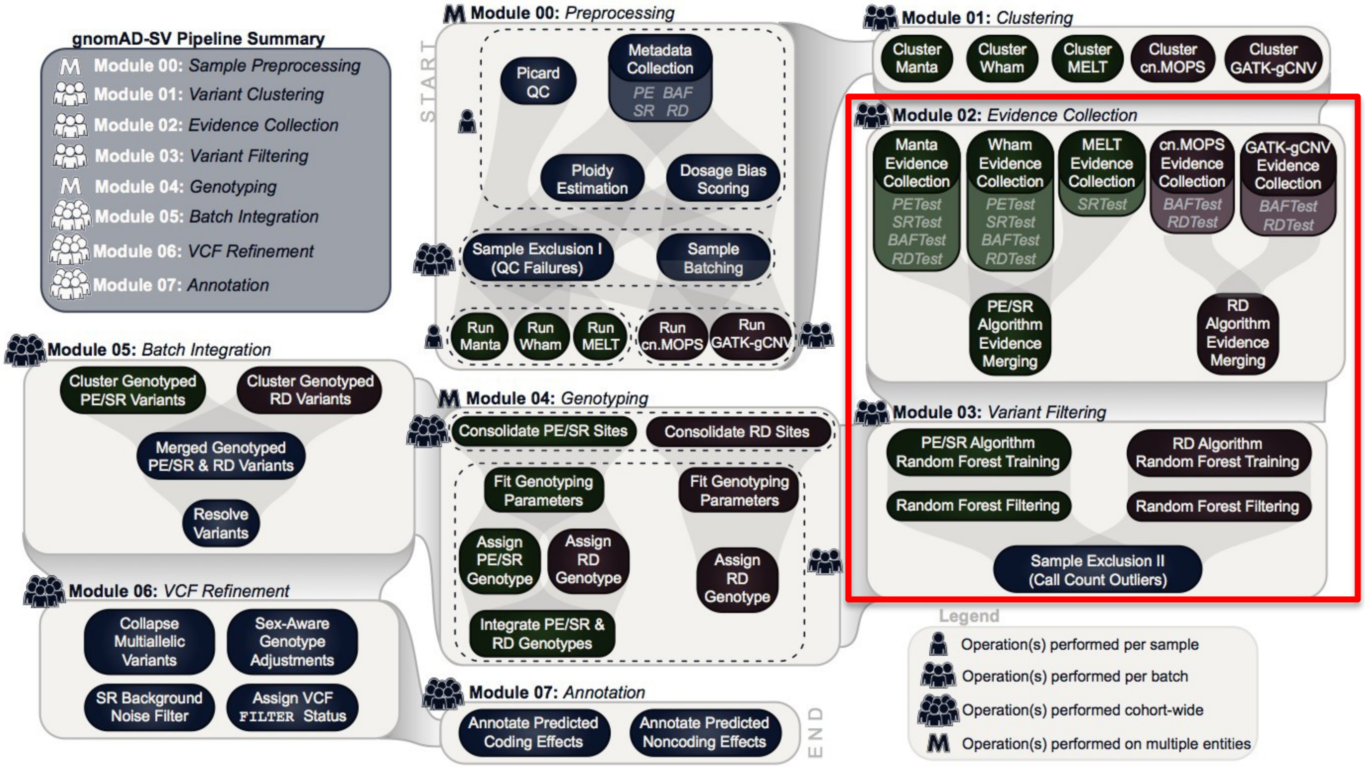
END

Legend

- Operation(s) performed per sample
- Operation(s) performed per batch
- Operation(s) performed cohort-wide
- M Operation(s) performed on multiple entities

- Run several unfiltered algorithms to **maximize sensitivity**
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Has been adapted to work on Google Cloud via Broad Institute's Terra Platform

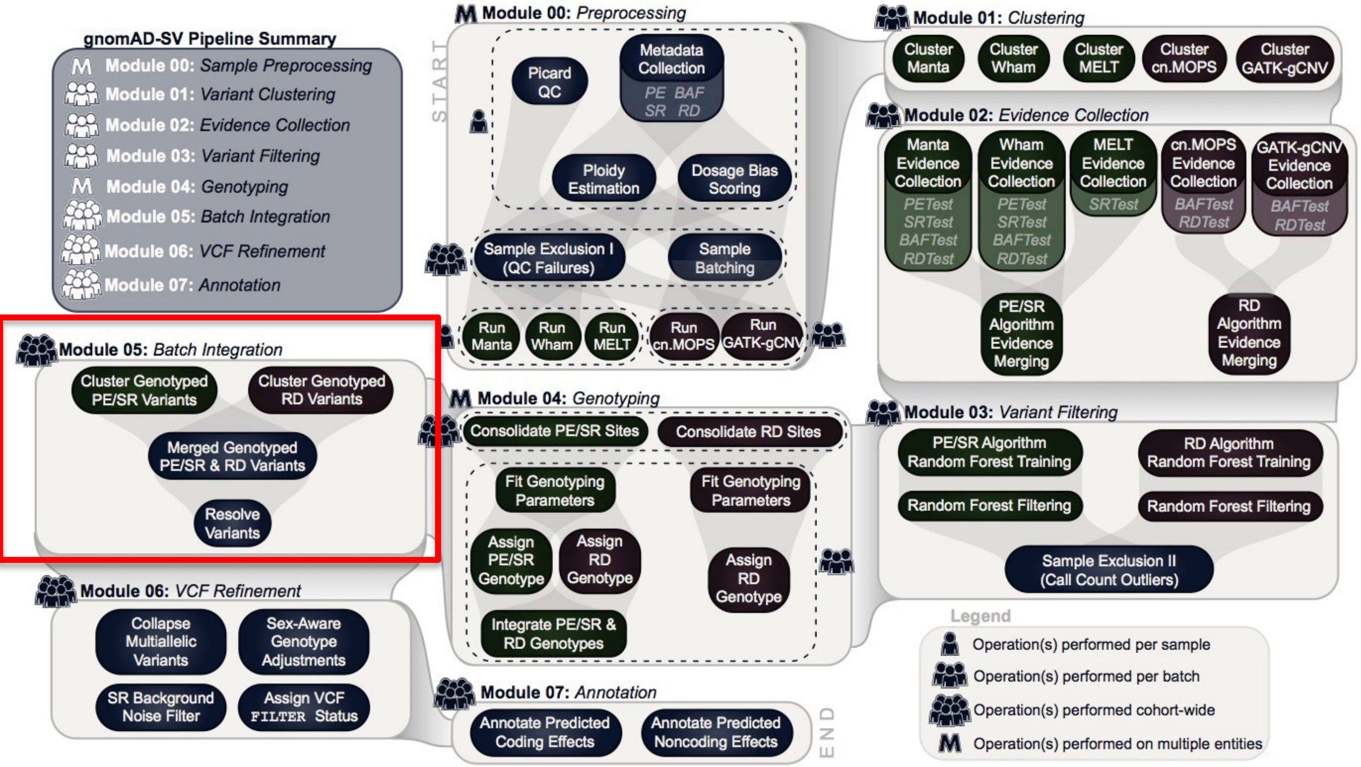
GATK-SV: Cloud Enabled SV Pipeline


gnomAD-SV Pipeline Summary
Module 00: Sample Preprocessing
Module 01: Variant Clustering
Module 02: Evidence Collection
Module 03: Variant Filtering
Module 04: Genotyping
Module 05: Batch Integration
Module 06: VCF Refinement
Module 07: Annotation
START
Module 00: Preprocessing
Picard QC
Metadata Collection
PE BAF
SR RD
Ploidy Estimation
Dosage Bias Scoring
Sample Exclusion I (QC Failures)
Sample Batching
Run Manta
Run Wham
Run MELT
Run cn.MOPS
Run GATK-gCNV
Module 01: Clustering
Cluster Manta
Cluster Wham
Cluster MELT
Cluster cn.MOPS
Cluster GATK-gCNV
Module 02: Evidence Collection
Manta Evidence Collection
PETest SRTest BAFTest RDTest
Wham Evidence Collection
PETest SRTest BAFTest RDTest
MELT Evidence Collection
SRTest
cn.MOPS Evidence Collection
BAFTest RDTest
GATK-gCNV Evidence Collection
BAFTest RDTest
PE/SR Algorithm Evidence Merging
RD Algorithm Evidence Merging
Module 03: Variant Filtering
PE/SR Algorithm Random Forest Training
RD Algorithm Random Forest Training
Random Forest Filtering
Random Forest Filtering
Sample Exclusion II (Call Count Outliers)
Module 04: Genotyping
Consolidate PE/SR Sites
Consolidate RD Sites
Fit Genotyping Parameters
Fit Genotyping Parameters
Assign PE/SR Genotype
Assign RD Genotype
Assign RD Genotype
Integrate PE/SR & RD Genotypes
Module 05: Batch Integration
Cluster Genotyped PE/SR Variants
Cluster Genotyped RD Variants
Merged Genotyped PE/SR & RD Variants
Resolve Variants
Module 06: VCF Refinement
Collapse Multiallelic Variants
Sex-Aware Genotype Adjustments
SR Background Noise Filter
Assign VCF FILTER Status
Module 07: Annotation
Annotate Predicted Coding Effects
Annotate Predicted Noncoding Effects
END
Legend
Operation(s) performed per sample
Operation(s) performed per batch
Operation(s) performed cohort-wide
Operation(s) performed on multiple entities

- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to **improve specificity**
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Has been adapted to work on Google Cloud via Broad Institute's Terra Platform

CENTER FOR GENOMIC MEDICINE

GATK-SV: Cloud Enabled SV Pipeline



gnomAD-SV Pipeline Summary

- Module 00: *Sample Preprocessing*
- Module 01: *Variant Clustering*
- Module 02: *Evidence Collection*
- Module 03: *Variant Filtering*
- Module 04: *Genotyping*
- Module 05: *Batch Integration*
- Module 06: *VCF Refinement*
- Module 07: *Annotation*

START

Module 00: *Preprocessing*

Picard QC | Metadata Collection (PE BAF SR RD) | Ploidy Estimation | Dosage Bias Scoring | Sample Exclusion I (QC Failures) | Sample Batching | Run Manta | Run Wham | Run MELT | Run cn.MOPS | Run GATK-gCNV

Module 01: *Clustering*

Cluster Manta | Cluster Wham | Cluster MELT | Cluster cn.MOPS | Cluster GATK-gCNV

Module 02: *Evidence Collection*

Manta Evidence Collection (PETest SRTest BAFTest RDTest) | Wham Evidence Collection (PETest SRTest BAFTest RDTest) | MELT Evidence Collection (SRTest) | cn.MOPS Evidence Collection (BAFTest RDTest) | GATK-gCNV Evidence Collection (BAFTest RDTest)

PE/SR Algorithm Evidence Merging | RD Algorithm Evidence Merging

Module 03: *Variant Filtering*

PE/SR Algorithm Random Forest Training | RD Algorithm Random Forest Training | Random Forest Filtering | Random Forest Filtering | Sample Exclusion II (Call Count Outliers)

Module 04: *Genotyping*

Consolidate PE/SR Sites | Consolidate RD Sites | Fit Genotyping Parameters | Fit Genotyping Parameters | Assign PE/SR Genotype | Assign RD Genotype | Assign RD Genotype | Integrate PE/SR & RD Genotypes

Module 05: *Batch Integration*

Cluster Genotyped PE/SR Variants | Cluster Genotyped RD Variants | Merged Genotyped PE/SR & RD Variants | Resolve Variants

Module 06: *VCF Refinement*

Collapse Multiallelic Variants | Sex-Aware Genotype Adjustments | SR Background Noise Filter | Assign VCF `FILTER` Status

Module 07: *Annotation*

Annotate Predicted Coding Effects | Annotate Predicted Noncoding Effects

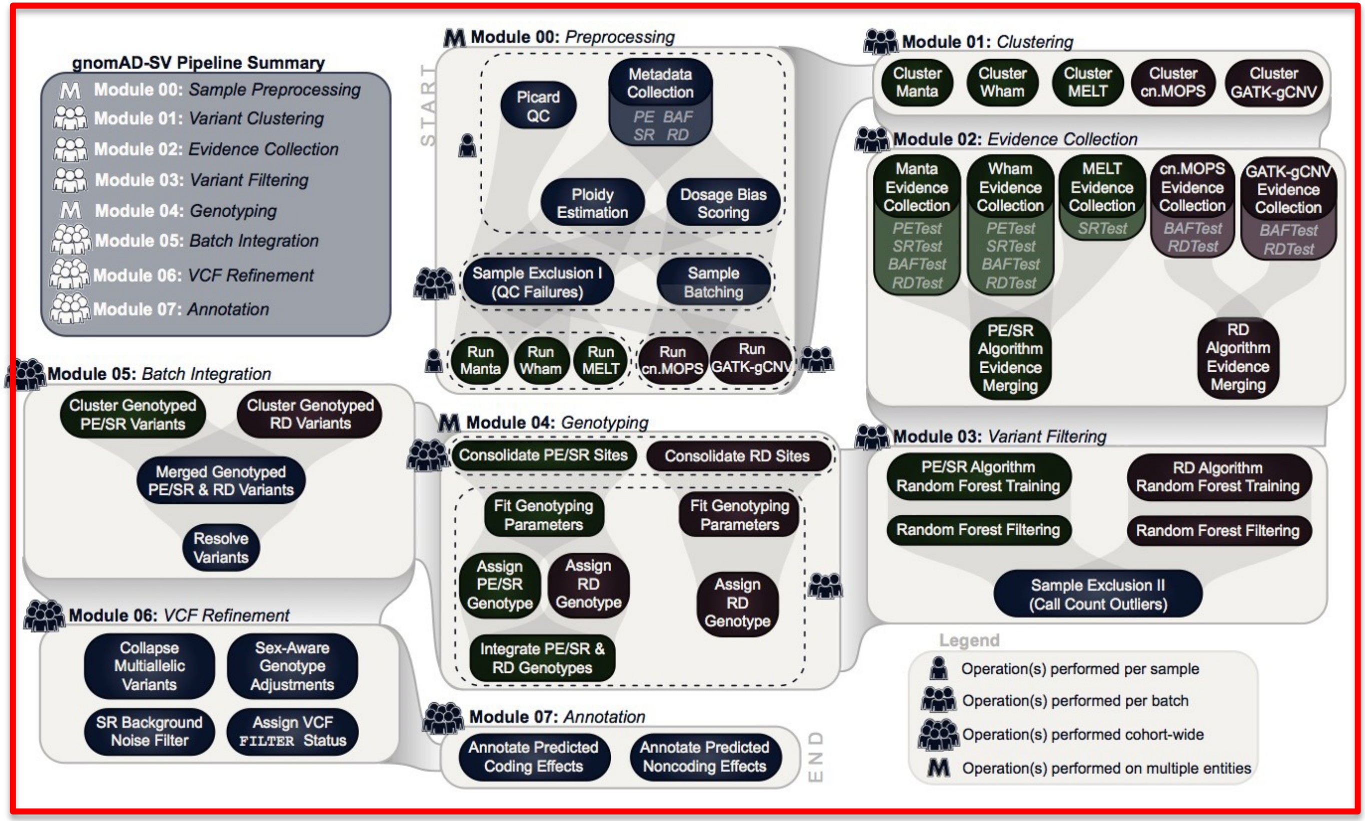
END

Legend

- Operation(s) performed per sample
- Operation(s) performed per batch
- Operation(s) performed cohort-wide
- Operation(s) performed on multiple entities

- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Has been adapted to work on Google Cloud via Broad Institute’s Terra Platform

GATK-SV; Cloud Enabled SV Pipeline



gnomAD-SV Pipeline Summary

- Module 00: *Sample Preprocessing*
- Module 01: *Variant Clustering*
- Module 02: *Evidence Collection*
- Module 03: *Variant Filtering*
- Module 04: *Genotyping*
- Module 05: *Batch Integration*
- Module 06: *VCF Refinement*
- Module 07: *Annotation*

START

Module 00: *Preprocessing*

Picard QC · Metadata Collection (PE BAF SR RD) · Ploidy Estimation · Dosage Bias Scoring · Sample Exclusion I (QC Failures) · Sample Batching · Run Manta · Run Wham · Run MELT · Run cn.MOPS · Run GATK-gCNV

Module 01: *Clustering*

Cluster Manta · Cluster Wham · Cluster MELT · Cluster cn.MOPS · Cluster GATK-gCNV

Module 02: *Evidence Collection*

Manta Evidence Collection (PETest, SRTest, BAFTest, RDTest) · Wham Evidence Collection (PETest, SRTest, BAFTest, RDTest) · MELT Evidence Collection (SRTest) · cn.MOPS Evidence Collection (BAFTest, RDTest) · GATK-gCNV Evidence Collection (BAFTest, RDTest) · PE/SR Algorithm Evidence Merging · RD Algorithm Evidence Merging

Module 03: *Variant Filtering*

PE/SR Algorithm Random Forest Training · RD Algorithm Random Forest Training · Random Forest Filtering · Random Forest Filtering · Sample Exclusion II (Call Count Outliers)

Module 04: *Genotyping*

Consolidate PE/SR Sites · Consolidate RD Sites · Fit Genotyping Parameters · Fit Genotyping Parameters · Assign PE/SR Genotype · Assign RD Genotype · Assign RD Genotype · Integrate PE/SR & RD Genotypes

Module 05: *Batch Integration*

Cluster Genotyped PE/SR Variants · Cluster Genotyped RD Variants · Merged Genotyped PE/SR & RD Variants · Resolve Variants

Module 06: *VCF Refinement*

Collapse Multiallelic Variants · Sex-Aware Genotype Adjustments · SR Background Noise Filter · Assign VCF `FILTER` Status

Module 07: *Annotation*

Annotate Predicted Coding Effects · Annotate Predicted Noncoding Effects

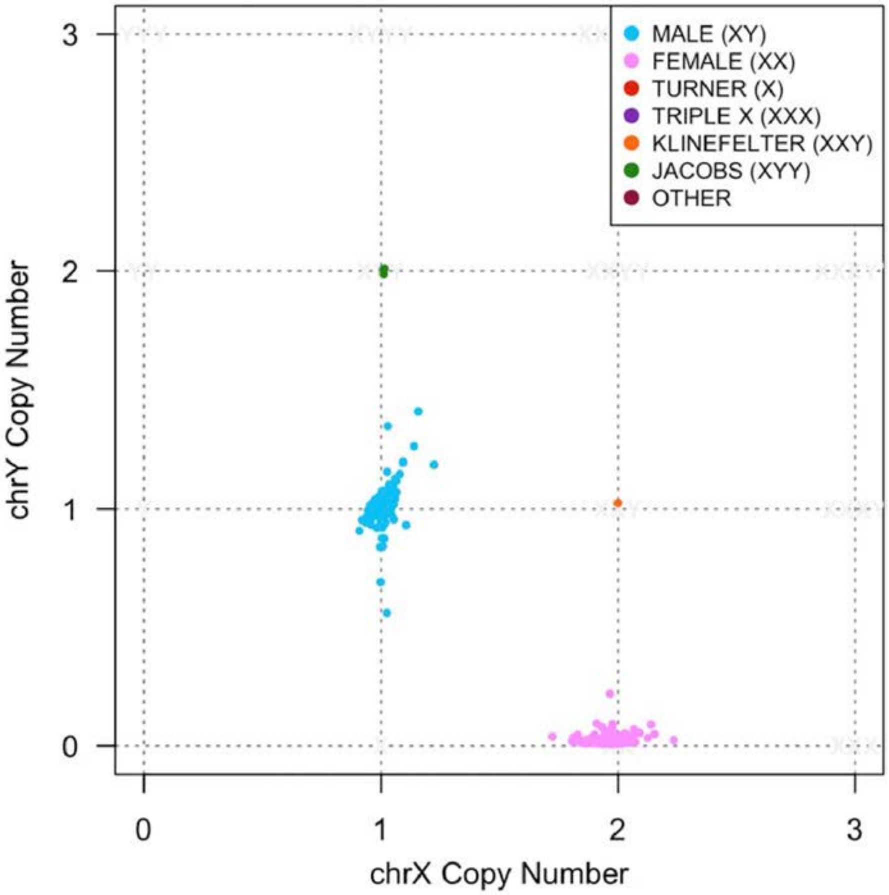
END

Legend

- Operation(s) performed per sample
- Operation(s) performed per batch
- Operation(s) performed cohort-wide
- M Operation(s) performed on multiple entities

- Run several unfiltered algorithms to maximize sensitivity
- Re-evaluate evidence directly from BAMs to improve specificity
- Captures both unbalanced (CNV) and balanced (inversion, translocation) SV
- Integrates SV signatures to resolve complex events
- Has been adapted to work on Google Cloud via Broad Institute’s Terra Platform

Sex Chromosome Aneuploidies

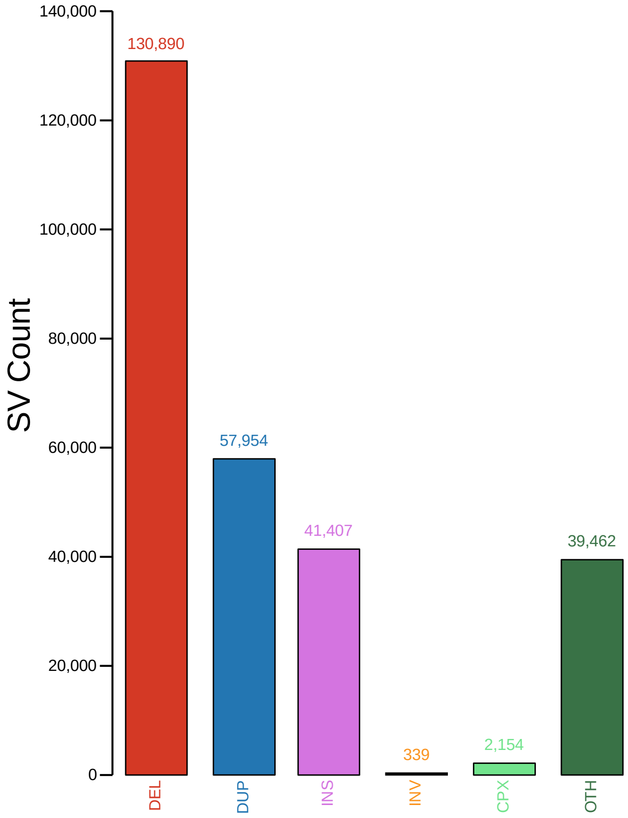


- 1 sample of Klinefelter Syndrome (XXY)
  - ➢ 1 Proband

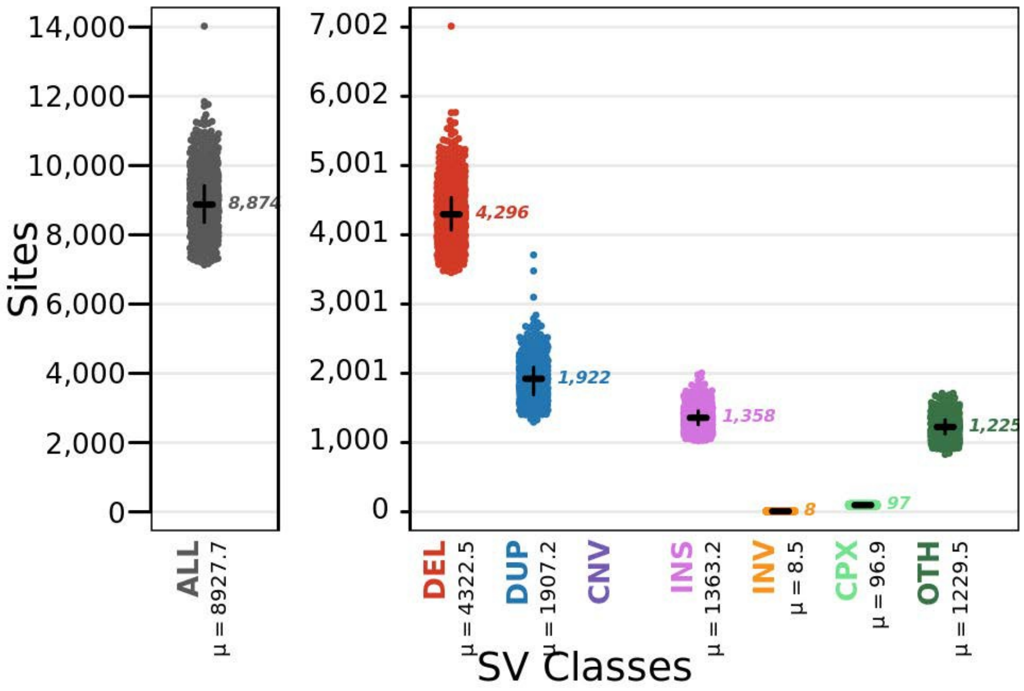
- 4 samples with Jacob Syndrome (XYY)
  - ➢ 3 Fathers
  - ➢ 1 Proband

SV Counts from OFC Cohort

Total Variant Count



SV Per Sample



De Novo SV Detection

Filtering GATK-SV for *De Novo* Events

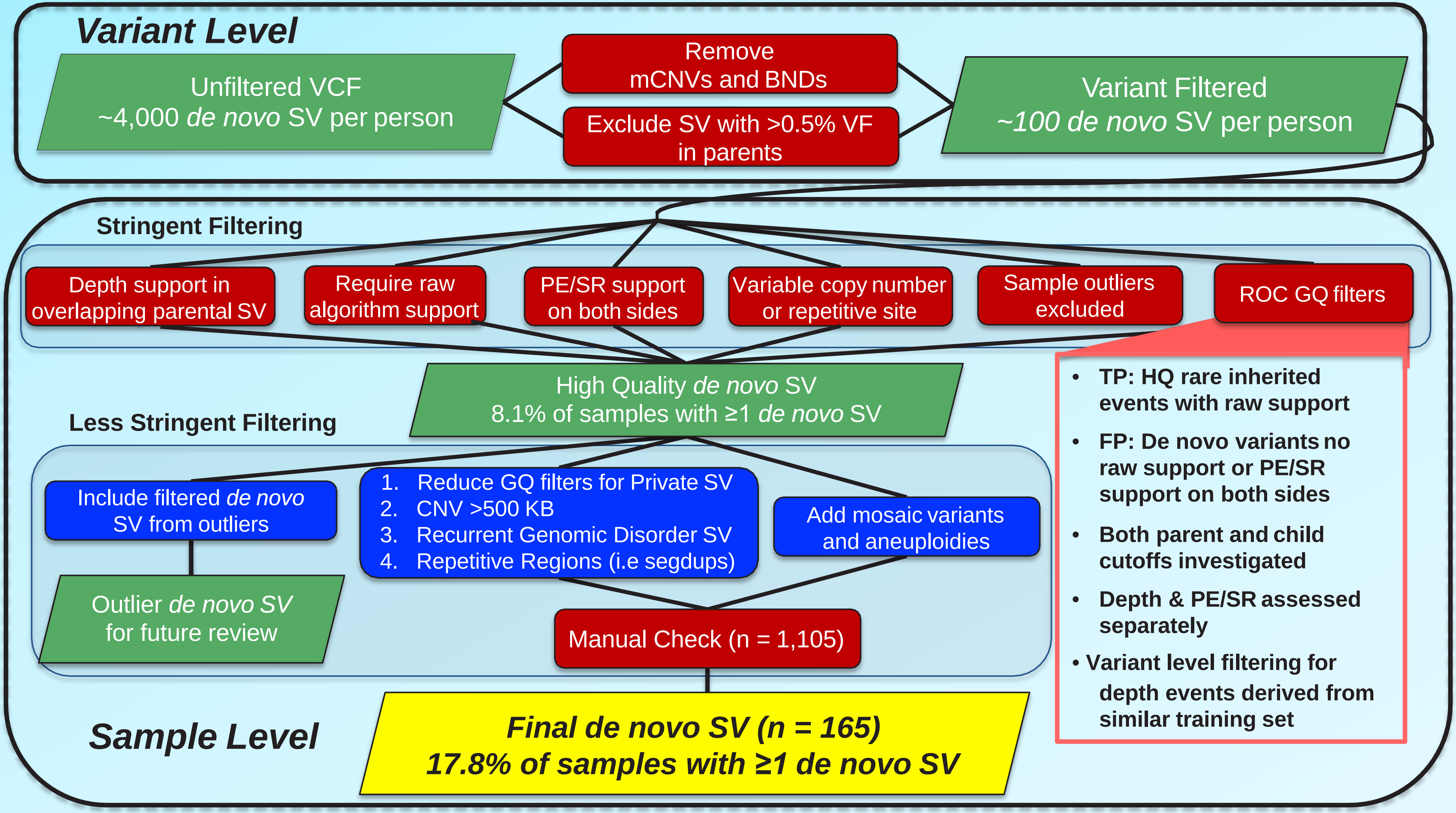
- Similar to SNVs and indels, **careful filtering is required to identify** *de novo* SVs at high precision

- SVs can be **misclassified** due to **parental mosaicism**, **lack of phasing**, and/or **inconsistent evidence**

- We developed a *post hoc* GATK-SV *de novo* workflow to eliminate erroneous *de novo* events

- Sample phenotypes and ascertainment can have a huge effect on # of *de novo* SVs per cohort

GATK-SV *De Novo* Workflow



Variant Level

Unfiltered VCF
~4,000 *de novo* SV per person

Remove
mCNVs and BNDs

Exclude SV with >0.5% VF
in parents

Variant Filtered
~100 de novo SV per person

Stringent Filtering

Depth support in
overlapping parental SV

Require raw
algorithm support

PE/SR support
on both sides

Variable copy number
or repetitive site

Sample outliers
excluded

ROC GQ filters

- **TP: HQ rare inherited events with raw support**
- **FP: De novo variants no raw support or PE/SR support on both sides**
- **Both parent and child cutoffs investigated**
- **Depth & PE/SR assessed separately**
- **Variant level filtering for depth events derived from similar training set**

High Quality *de novo* SV
8.1% of samples with ≥1 *de novo* SV

Less Stringent Filtering

Include filtered *de novo*
SV from outliers

Outlier *de novo SV*
for future review

1. Reduce GQ filters for Private SV
2. CNV >500 KB
3. Recurrent Genomic Disorder SV
4. Repetitive Regions (i.e segdups)

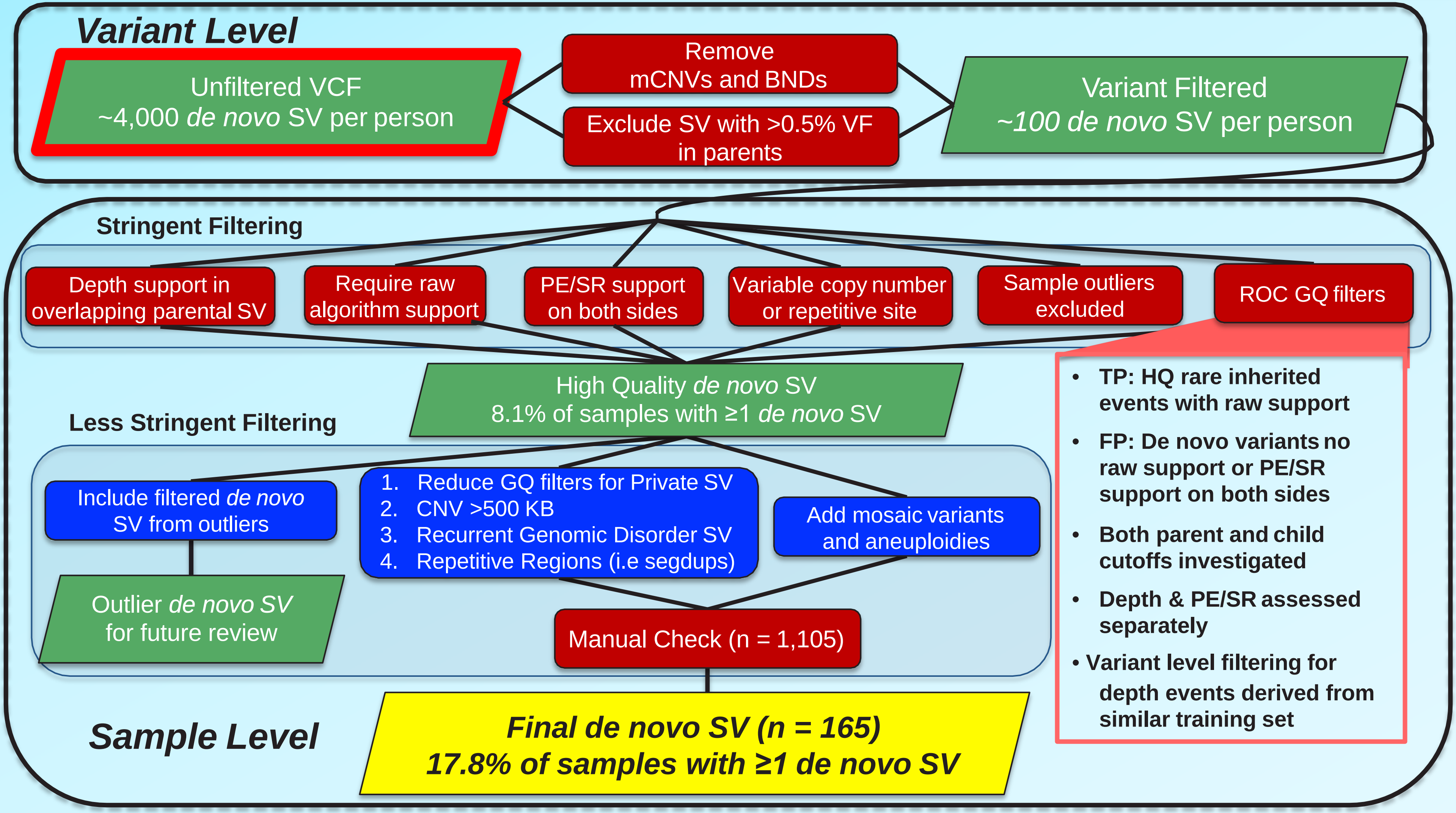
Add mosaic variants
and aneuploidies

Manual Check (n = 1,105)

Sample Level

Final de novo SV (n = 165)
17.8% of samples with ≥1 de novo SV

GATK-SV *De Novo* Workflow



Variant Level

Unfiltered VCF
~4,000 *de novo* SV per person

Remove
mCNVs and BNDs

Exclude SV with >0.5% VF
in parents

Variant Filtered
~100 de novo SV per person

Stringent Filtering

Depth support in overlapping parental SV

Require raw algorithm support

PE/SR support on both sides

Variable copy number or repetitive site

Sample outliers excluded

ROC GQ filters

High Quality *de novo* SV
8.1% of samples with ≥1 *de novo* SV

Less Stringent Filtering

Include filtered *de novo* SV from outliers

Outlier *de novo SV* for future review

1. Reduce GQ filters for Private SV
2. CNV >500 KB
3. Recurrent Genomic Disorder SV
4. Repetitive Regions (i.e segdups)

Add mosaic variants and aneuploidies

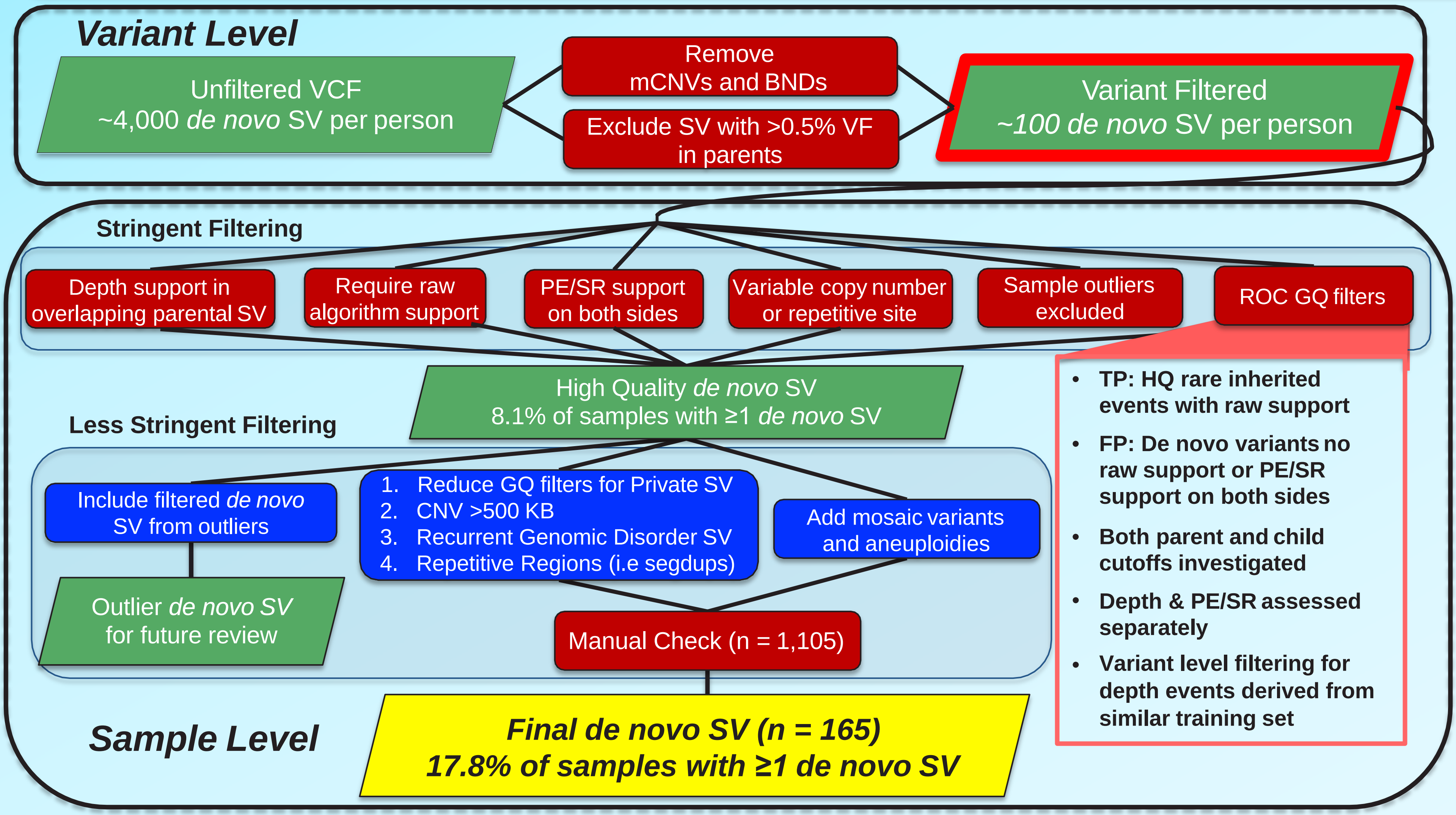
Manual Check (n = 1,105)

Sample Level

Final de novo SV (n = 165)
17.8% of samples with ≥1 de novo SV

- **TP: HQ rare inherited events with raw support**
- **FP: De novo variants no raw support or PE/SR support on both sides**
- **Both parent and child cutoffs investigated**
- **Depth & PE/SR assessed separately**
- **Variant level filtering for depth events derived from similar training set**

GATK-SV *De Novo* Workflow



Variant Level

Unfiltered VCF
~4,000 *de novo* SV per person

Remove
mCNVs and BNDs

Exclude SV with >0.5% VF
in parents

Variant Filtered
~100 de novo SV per person

Stringent Filtering

Depth support in
overlapping parental SV

Require raw
algorithm support

PE/SR support
on both sides

Variable copy number
or repetitive site

Sample outliers
excluded

ROC GQ filters

- **TP: HQ rare inherited events with raw support**
- **FP: De novo variants no raw support or PE/SR support on both sides**
- **Both parent and child cutoffs investigated**
- **Depth & PE/SR assessed separately**
- **Variant level filtering for depth events derived from similar training set**

High Quality *de novo* SV
8.1% of samples with ≥1 *de novo* SV

Less Stringent Filtering

Include filtered *de novo*
SV from outliers

Outlier *de novo SV*
for future review

1. Reduce GQ filters for Private SV
2. CNV >500 KB
3. Recurrent Genomic Disorder SV
4. Repetitive Regions (i.e segdups)

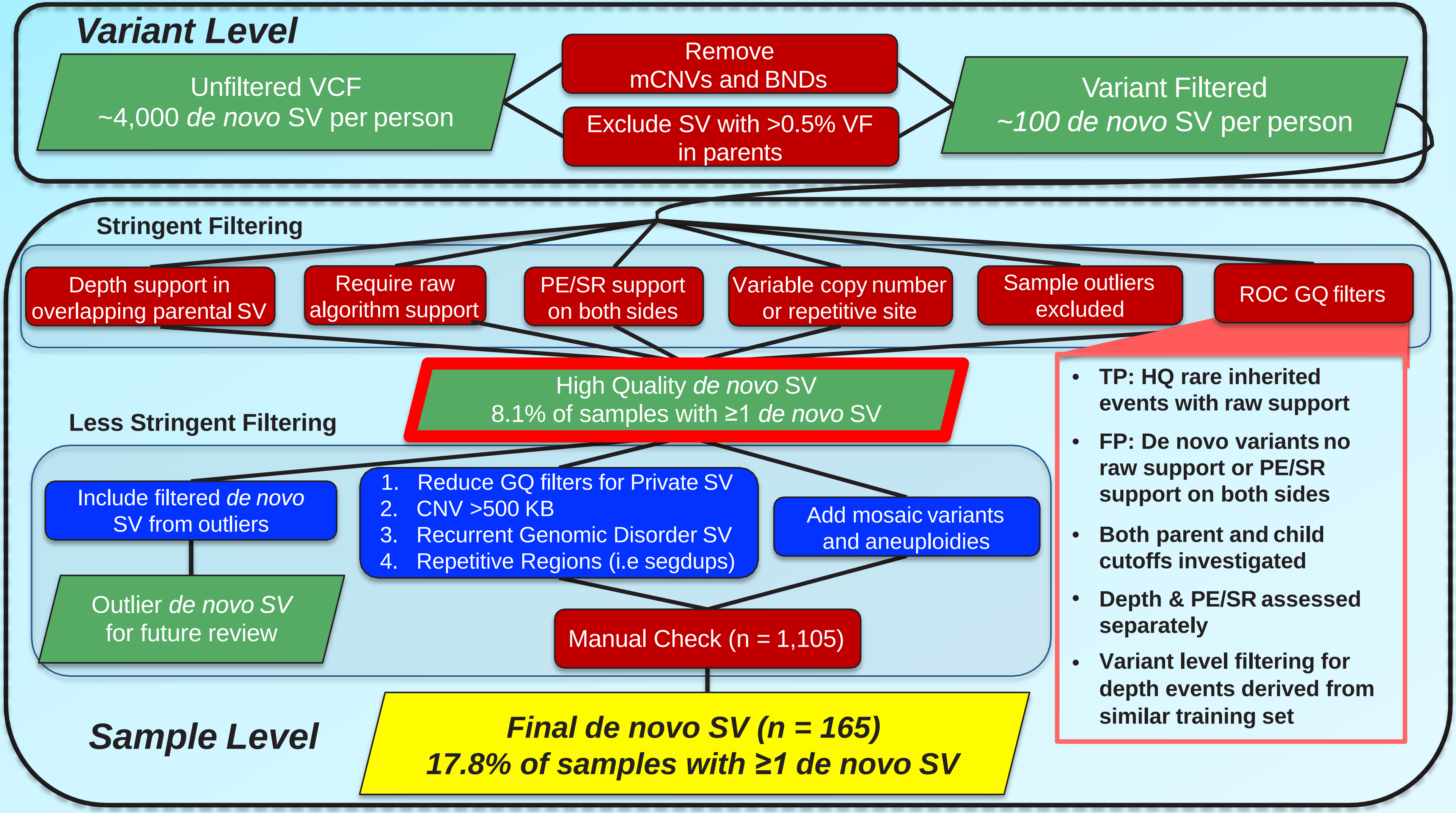
Add mosaic variants
and aneuploidies

Manual Check (n = 1,105)

Sample Level

Final de novo SV (n = 165)
17.8% of samples with ≥1 de novo SV

GATK-SV *De Novo* Workflow



Variant Level

Unfiltered VCF
~4,000 *de novo* SV per person

Remove
mCNVs and BNDs

Exclude SV with >0.5% VF
in parents

Variant Filtered
~100 de novo SV per person

Stringent Filtering

Depth support in
overlapping parental SV

Require raw
algorithm support

PE/SR support
on both sides

Variable copy number
or repetitive site

Sample outliers
excluded

ROC GQ filters

- **TP: HQ rare inherited events with raw support**
- **FP: De novo variants no raw support or PE/SR support on both sides**
- **Both parent and child cutoffs investigated**
- **Depth & PE/SR assessed separately**
- **Variant level filtering for depth events derived from similar training set**

High Quality *de novo* SV
8.1% of samples with ≥1 *de novo* SV

Less Stringent Filtering

Include filtered *de novo*
SV from outliers

Outlier *de novo SV*
for future review

1. Reduce GQ filters for Private SV
2. CNV >500 KB
3. Recurrent Genomic Disorder SV
4. Repetitive Regions (i.e segdups)

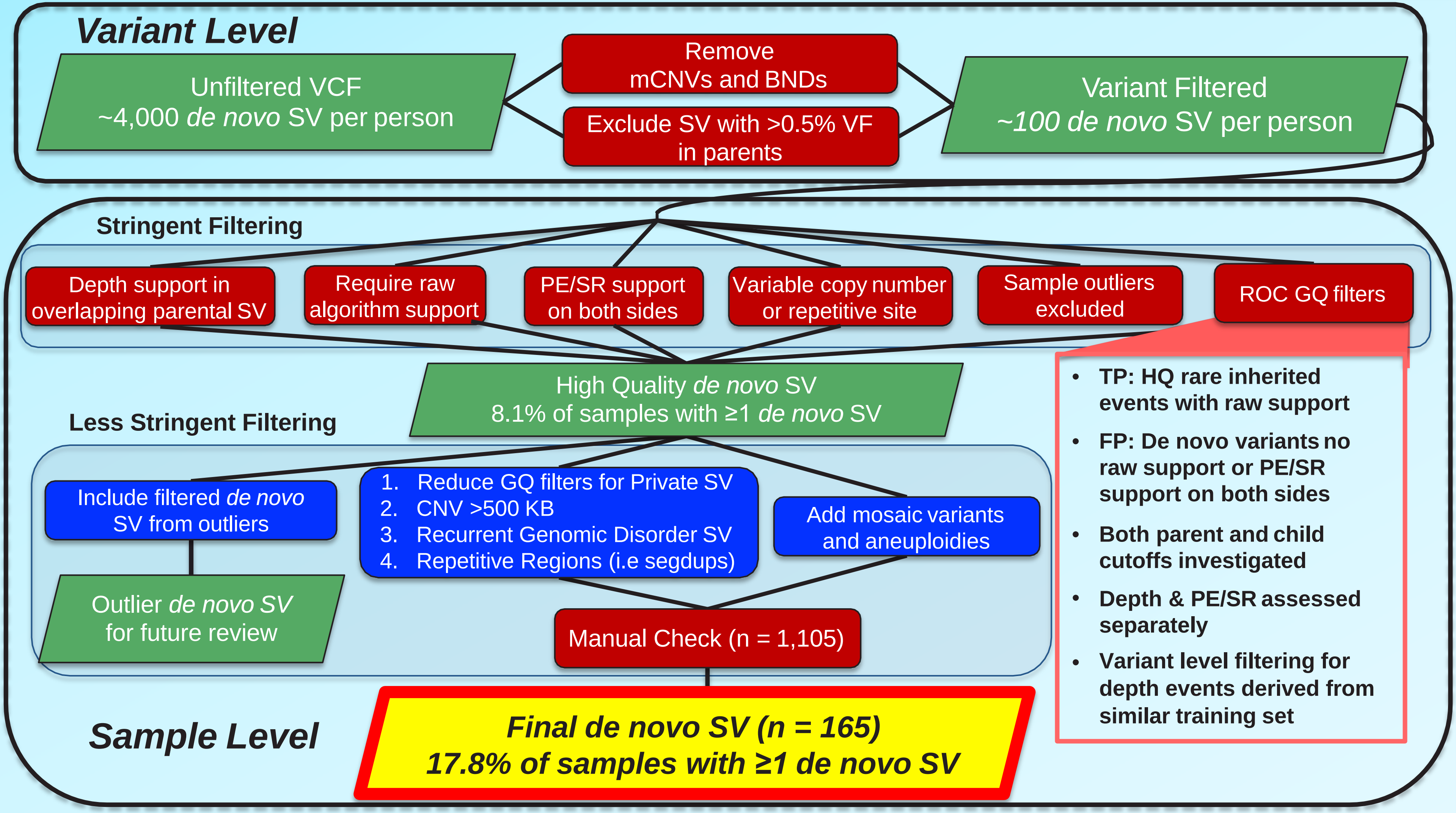
Add mosaic variants
and aneuploidies

Manual Check (n = 1,105)

Sample Level

Final de novo SV (n = 165)
17.8% of samples with ≥1 de novo SV

GATK-SV *De Novo* Workflow



Variant Level

Unfiltered VCF
~4,000 *de novo* SV per person

Remove
mCNVs and BNDs

Exclude SV with >0.5% VF
in parents

Variant Filtered
~100 de novo SV per person

Stringent Filtering

Depth support in
overlapping parental SV

Require raw
algorithm support

PE/SR support
on both sides

Variable copy number
or repetitive site

Sample outliers
excluded

ROC GQ filters

- **TP: HQ rare inherited events with raw support**
- **FP: De novo variants no raw support or PE/SR support on both sides**
- **Both parent and child cutoffs investigated**
- **Depth & PE/SR assessed separately**
- **Variant level filtering for depth events derived from similar training set**

High Quality *de novo* SV
8.1% of samples with ≥1 *de novo* SV

Less Stringent Filtering

Include filtered *de novo*
SV from outliers

Outlier *de novo SV*
for future review

1. Reduce GQ filters for Private SV
2. CNV >500 KB
3. Recurrent Genomic Disorder SV
4. Repetitive Regions (i.e segdups)

Add mosaic variants
and aneuploidies

Manual Check (n = 1,105)

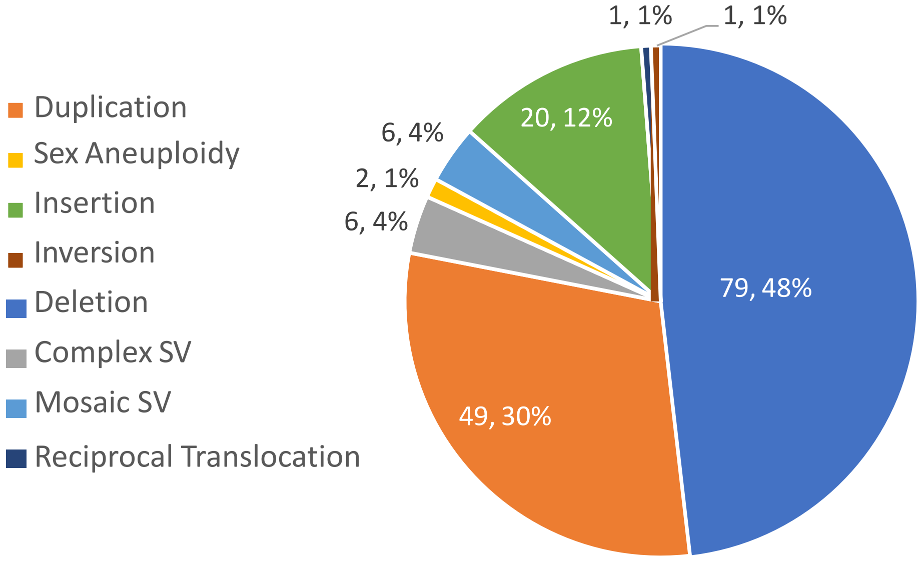
Sample Level

Final de novo SV (n = 165)
17.8% of samples with ≥1 de novo SV

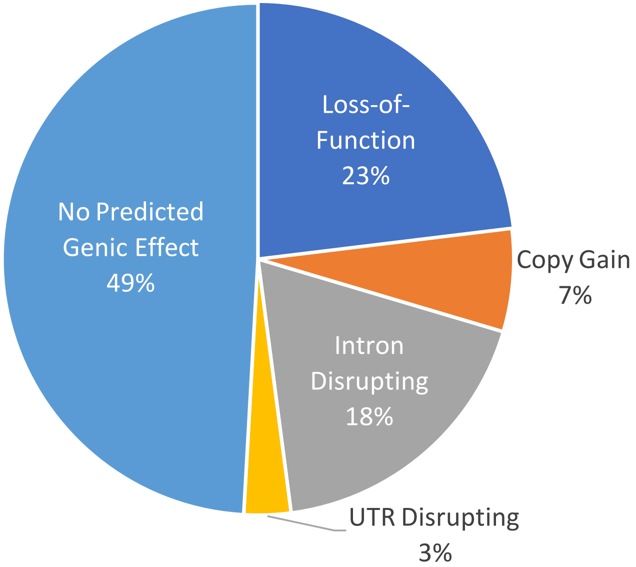
Summary of *De Novo* SVs

- After filtering, we observe 165 *de novo* SVs in 138 probands
- 17.8% of probands have at least one de novo event after accounting for 63 outlier samples that failed the de novo pipeline

SV Type



Functional Impact

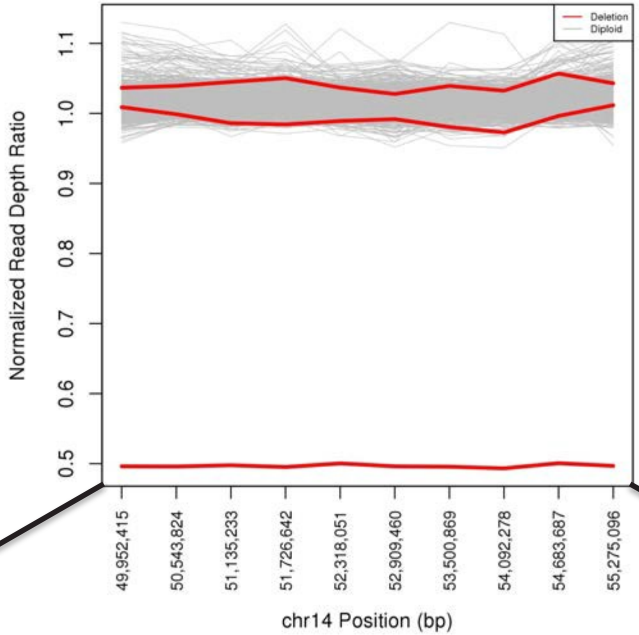


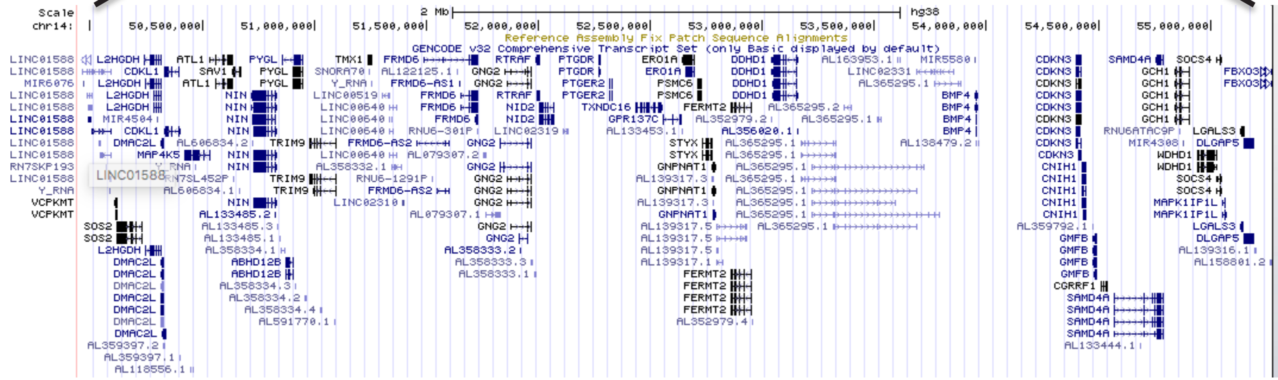
Highlighted De Novo SV

De Novo SV in Recurrent Genomic Disorder Regions (n = 14 in cohort)

Syndrome	Size	N	OFC Reported in Syndrome	Ethnicity	Case Phenotype
1q21.1 proximal dup	195 kb	1	Yes	Asian	CL/P
7q11.23 dup	1.4 Mb	2	Yes	Asian African	CL/P CL/P
8p23.1 dup	3.6 Mb	1	Yes	Latino	CL/P
15q11.2 del (BP1-BP2)	290 kb	1	Yes	Latino	CL/P
15q11.2 dup (BP1-BP2)	290 kb	1	No	Caucasian	CL
16p11.2 distal del	224 kb	1	Yes	Asian	CL/P
16p11.2 distal dup	224 kb	1	No	Caucasian	CP
22q11.2 del	2.6 Mb	1	Yes	Caucasian	CL/P
22q11.2 dup	2.6 Mb	1	Yes	Caucasian	CL/P
22q11.2 distal deletion	1.7 Mb	2	Yes	Caucasian African	CL/P CL/P
22q11.2 distal del	1.7 Mb	1	Yes	Asian	CL/P
Xp22.31 del (female)	1.6 Mb	1	No	Caucasian	CL/P

Other Large *De Novo* CNVs >1 Mb (n = 4 in cohort)

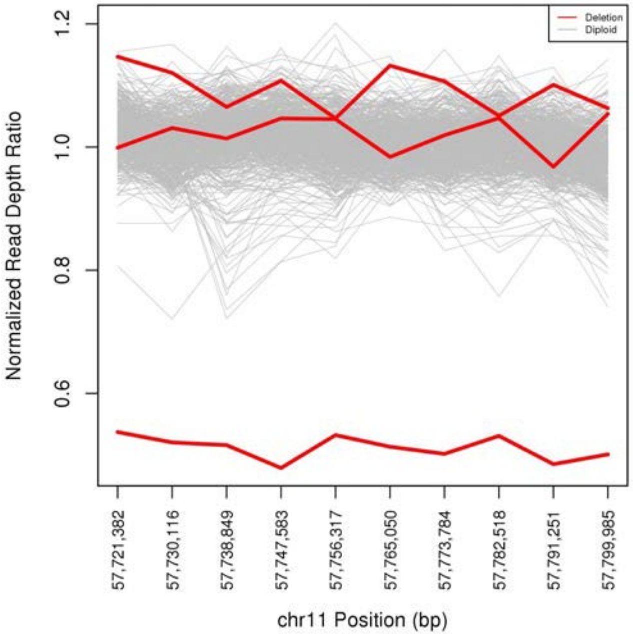




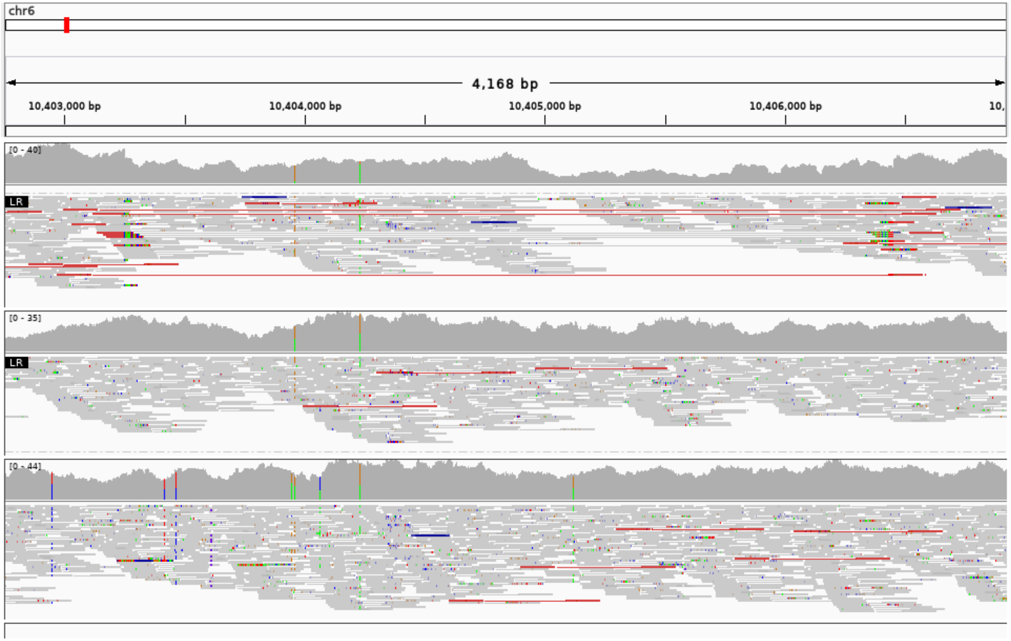
5.3 mb deletion
(chr4:49,952,415-55,275,096)

- Disrupts >30 protein coding genes
- Difficult to pinpoint single causative gene

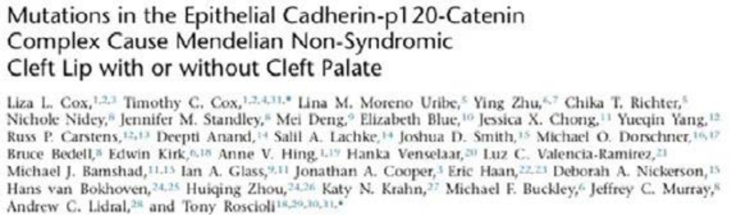
De Novo SV Disrupting Established OFC Genes (n = 2 in cohort)



78.6 kb deletion disrupts *BTBD18*, ***CTNND1***, *SELENOH*, *TMX2*

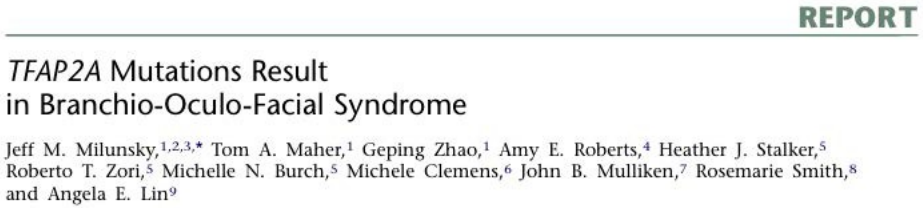


3.2 kb deletion disrupts ***TFAP2A***



Mutations in the Epithelial Cadherin-p120-Catenin Complex Cause Mendelian Non-Syndromic Cleft Lip with or without Cleft Palate

Liza L. Cox,[1,2,3] Timothy C. Cox,[1,2,4,31,*] Lina M. Moreno Uribe,[5] Ying Zhu,[6,7] Chika T. Richter,[5] Nichole Nidey,[8] Jennifer M. Standley,[8] Mei Deng,[9] Elizabeth Blue,[10] Jessica X. Chong,[11] Yueqin Yang,[12] Russ P. Carstens,[12,13] Deepti Anand,[14] Salil A. Lachke,[14] Joshua D. Smith,[15] Michael O. Dorschner,[16,17] Bruce Bedell,[8] Edwin Kirk,[6,18] Anne V. Hing,[1,19] Hanka Venselaar,[20] Luz C. Valencia-Ramirez,[21] Michael J. Bamshad,[11,15] Ian A. Glass,[9,11] Jonathan A. Cooper,[3] Eric Haan,[22,23] Deborah A. Nickerson,[15] Hans van Bokhoven,[24,25] Huiqing Zhou,[24,26] Katy N. Krahn,[27] Michael F. Buckley,[6] Jeffrey C. Murray,[8] Andrew C. Lidral,[28] and Tony Roscioli[18,29,30,31,*]

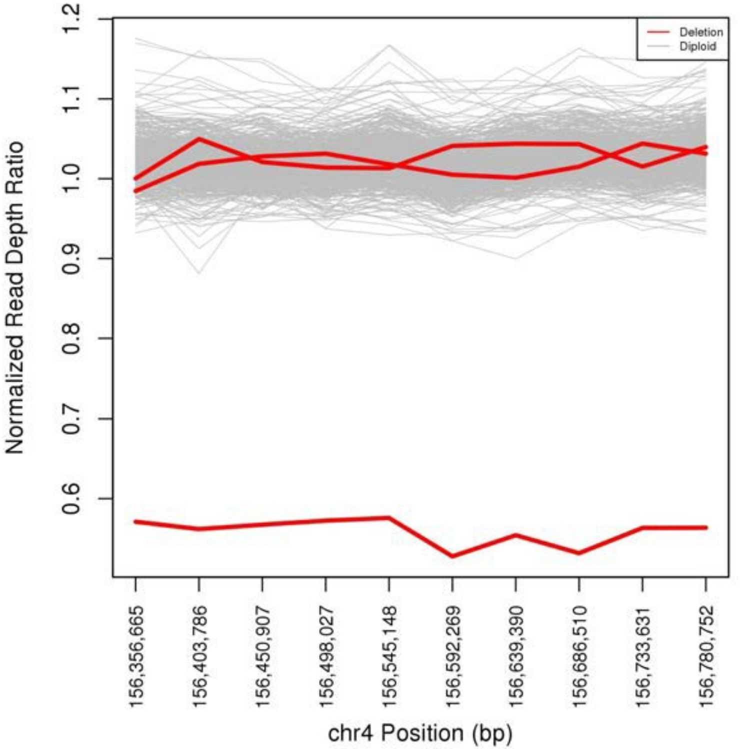


REPORT

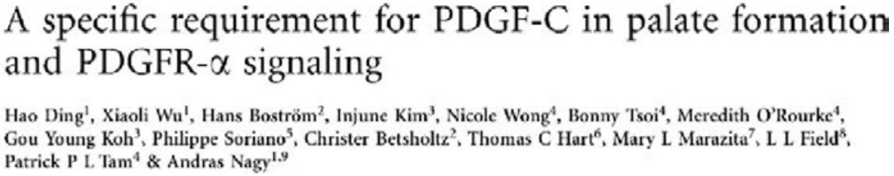
TFAP2A Mutations Result in Branchio-Oculo-Facial Syndrome

Jeff M. Milunsky,[1,2,3,*] Tom A. Maher,[1] Geping Zhao,[1] Amy E. Roberts,[4] Heather J. Stalker,[5] Roberto T. Zori,[5] Michelle N. Burch,[5] Michele Clemens,[6] John B. Mulliken,[7] Rosemarie Smith,[8] and Angela E. Lin[9]

De Novo SVs with Support in Animal Models (n = 2 in cohort)

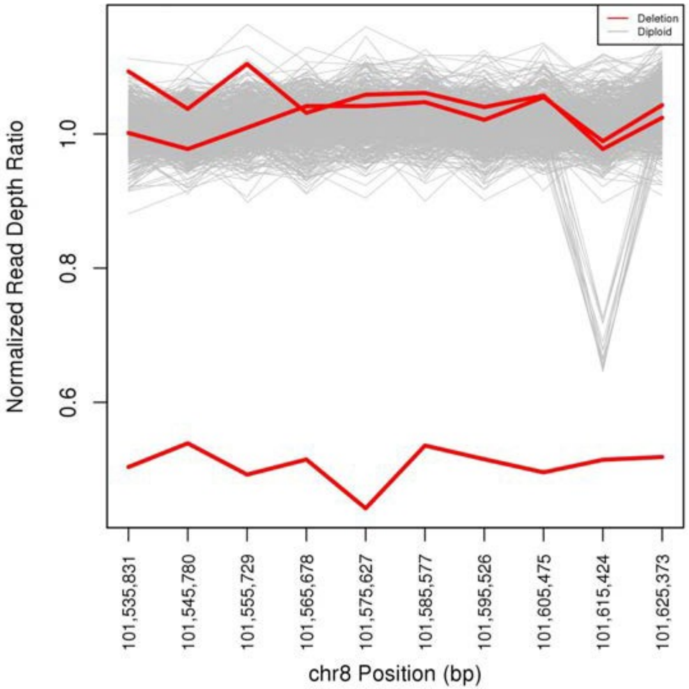


424.1 kb deletion disrupts ***PDGFC***

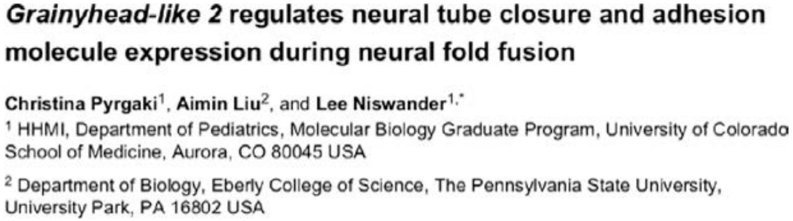


A specific requirement for PDGF-C in palate formation and PDGFR-α signaling

Hao Ding[1], Xiaoli Wu[1], Hans Boström[2], Injune Kim[3], Nicole Wong[4], Bonny Tsoi[4], Meredith O'Rourke[4], Gou Young Koh[3], Philippe Soriano[5], Christer Betsholtz[2], Thomas C Hart[6], Mary L Marazita[7], L L Field[8], Patrick P L Tam[4] & Andras Nagy[1,9]



9.0 kb deletion disrupts ***GRHL2***



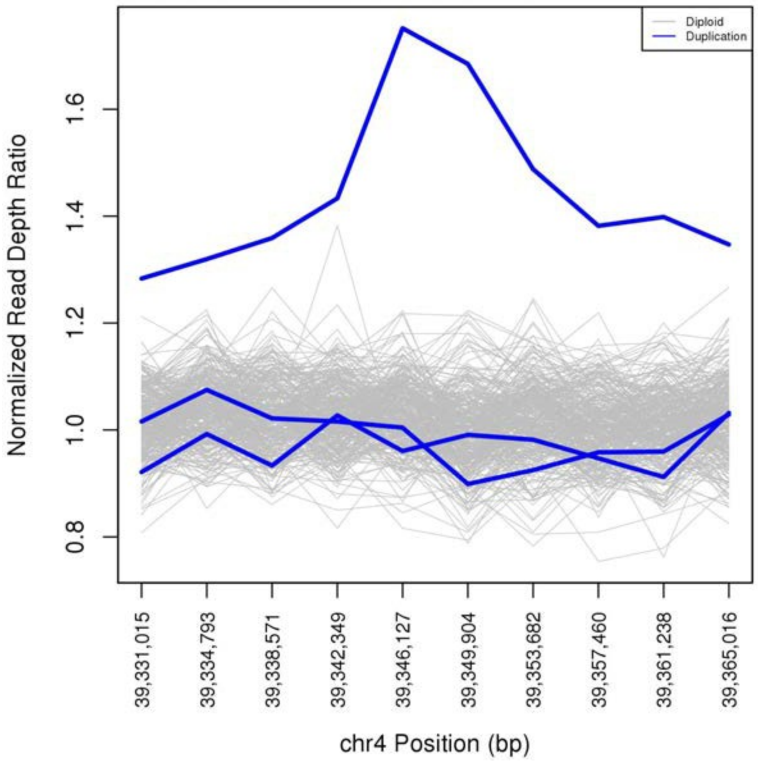
Grainyhead-like 2 **regulates neural tube closure and adhesion molecule expression during neural fold fusion**

Christina Pygaki[1], **Aimin Liu**[2], and **Lee Niswander**[1,*]

[1] HHMI, Department of Pediatrics, Molecular Biology Graduate Program, University of Colorado School of Medicine, Aurora, CO 80045 USA

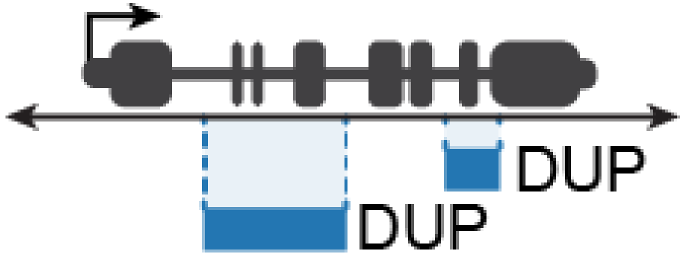
[2] Department of Biology, Eberly College of Science, The Pennsylvania State University, University Park, PA 16802 USA

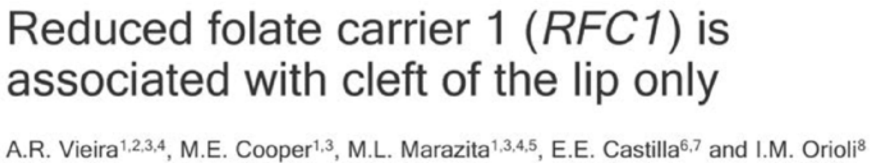
De Novo SVs Disrupting Previous Candidate Gene (n = 2 in cohort)



34 kb duplication disrupts *RFC1*

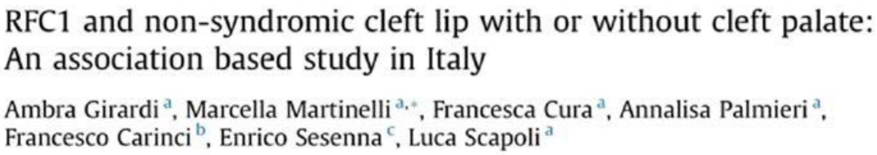
Intragenic Exonic Duplication





Reduced folate carrier 1 (*RFC1*) is associated with cleft of the lip only

A.R. Vieira[1,2,3,4], M.E. Cooper[1,3], M.L. Marazita[1,3,4,5], E.E. Castilla[6,7] and I.M. Orioli[8]

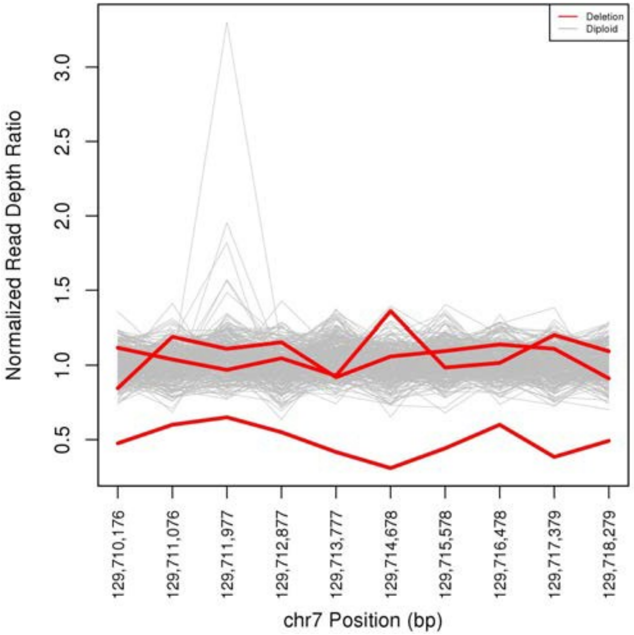


RFC1 and non-syndromic cleft lip with or without cleft palate: An association based study in Italy

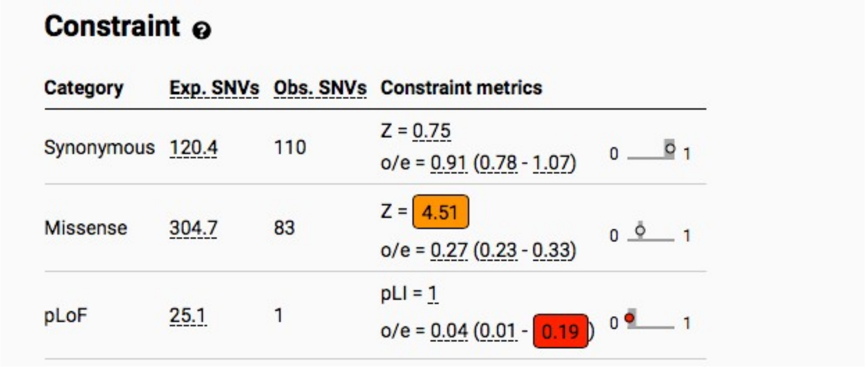
Ambra Girardi[a], Marcella Martinelli[a,*], Francesca Cura[a], Annalisa Palmieri[a], Francesco Carinci[b], Enrico Sesenna[c], Luca Scapoli[a]

De Novo SV Disrupting Novel Constrained Genes (n = 7 in cohort)

Constrained gene defined as gnomAD LOEUF < 0.4



8.1 kb deletion disrupts *NRF1*



Constraint

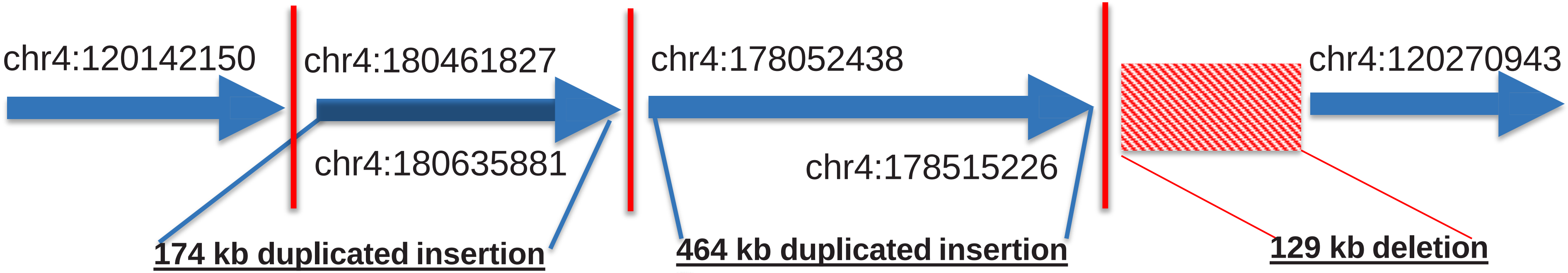
Category	Exp. SNVs	Obs. SNVs	Constraint metrics	
Synonymous	120.4	110	Z = 0.75 o/e = 0.91 (0.78 - 1.07)	0 1
Missense	304.7	83	Z = 4.51 o/e = 0.27 (0.23 - 0.33)	0 1
pLoF	25.1	1	pLI = 1 o/e = 0.04 (0.01 - 0.19)	0 1

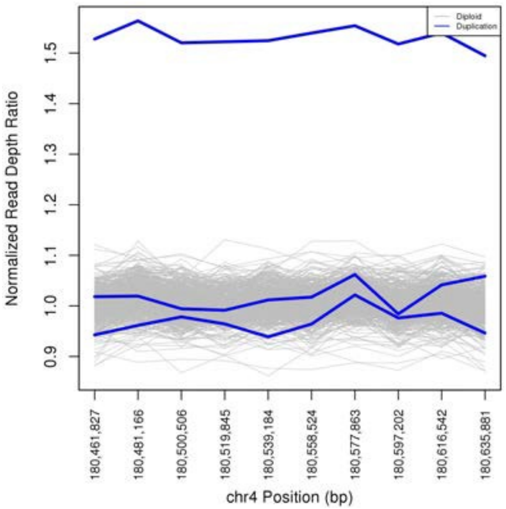
https://gnomad.broadinstitute.org

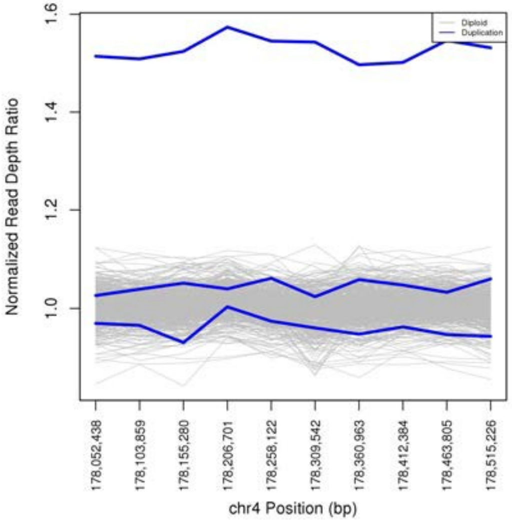
- Mutations in *NRF1* not yet associated with OFC
- Transcription factor involved in several pathways that could contribute to OFC

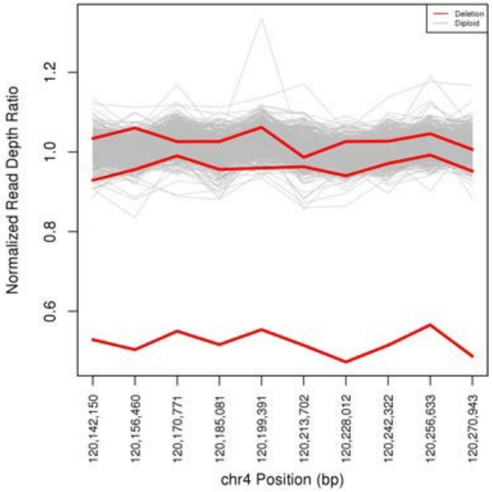
De Novo Complex Events (n = 6 in cohort)

Complex Event on Chromosome 4



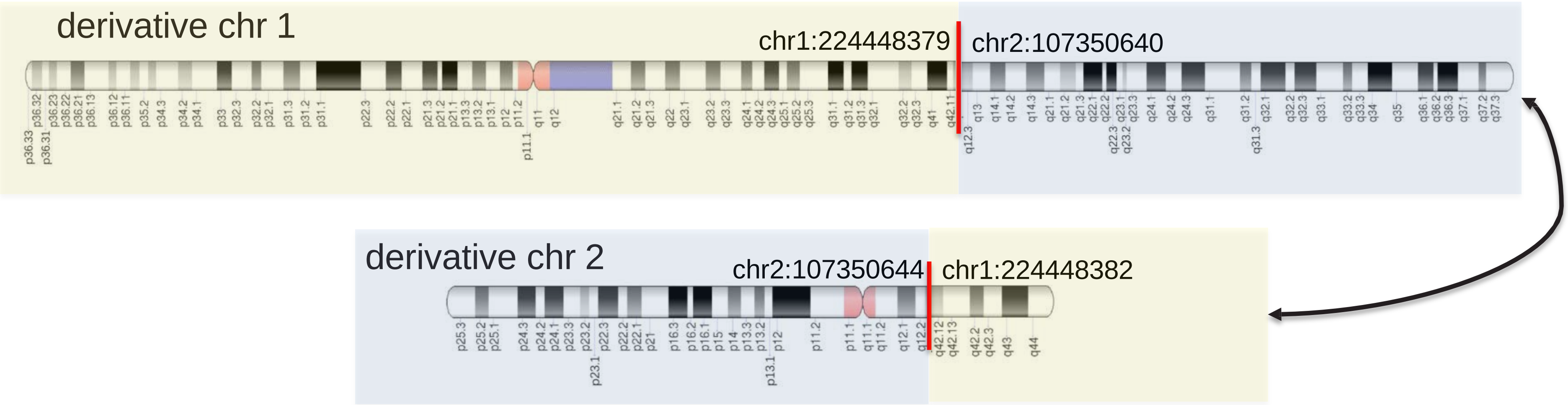




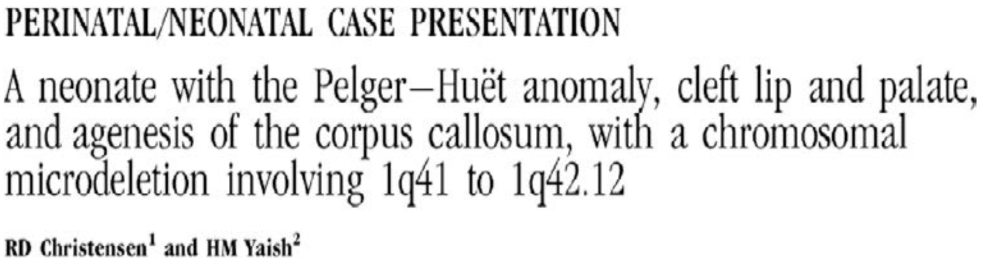


De Novo Reciprocal Translocations (n = 1 in cohort)

46,XY,t(1;2)(q42.12;q12.3) -   Disrupts *CNIH3*



Breakpoint overlaps edge of
1q41-q42 deletion syndrome

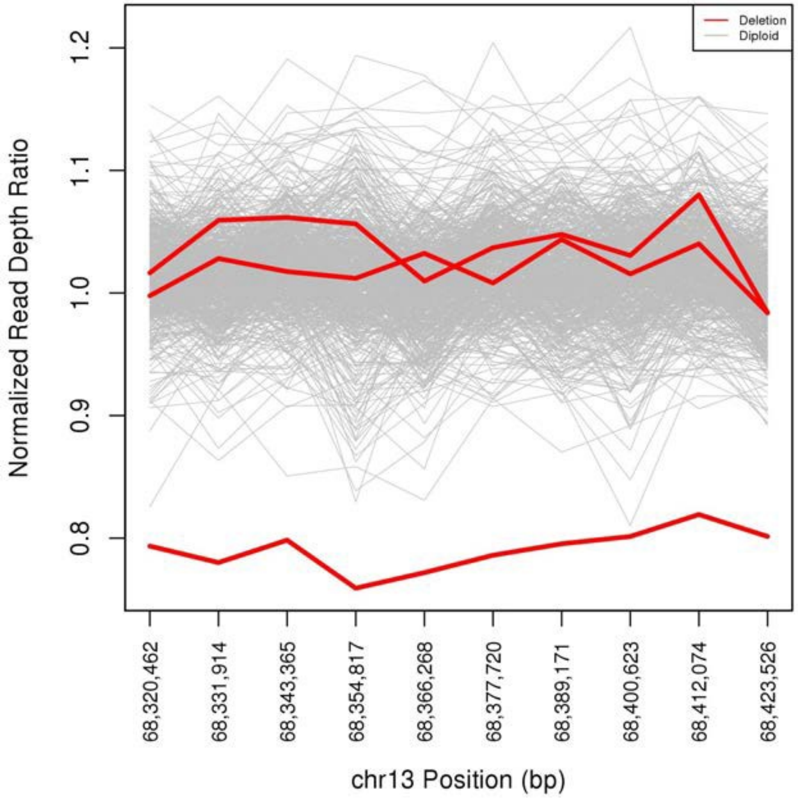

PERINATAL/NEONATAL CASE PRESENTATION

A neonate with the Pelger–Huët anomaly, cleft lip and palate, and agenesis of the corpus callosum, with a chromosomal microdeletion involving 1q41 to 1q42.12

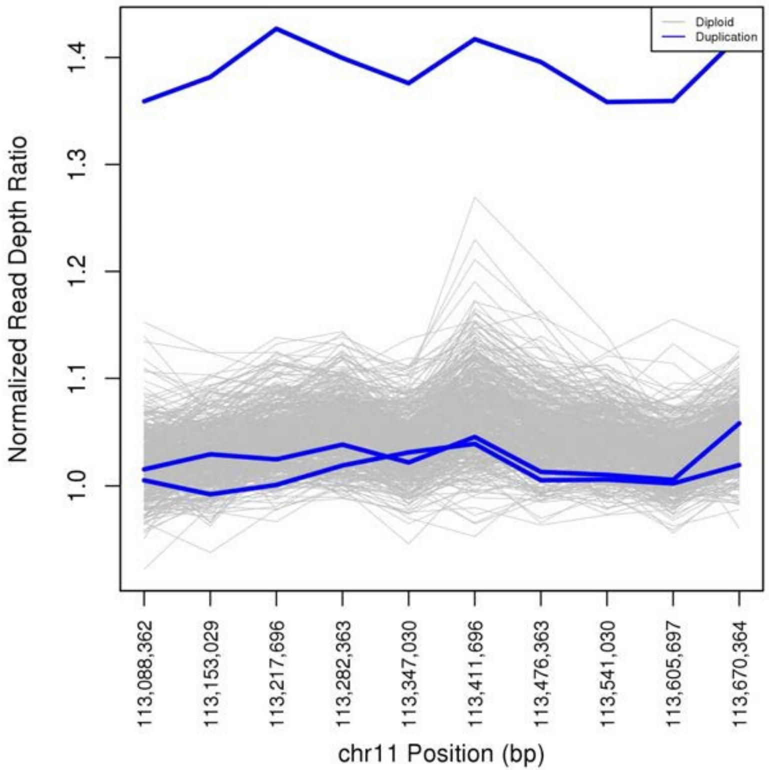
RD Christensen[1] and HM Yaish[2]

Ideograms modified from https://en.wikipedia.org/

Mosaic *De Novo* SV (n = 6 in cohort)



103 kb mosaic deletion predicted to be **present in 40%** of white blood cells



582 kb mosaic duplication predicted to be **present in 70%** of white blood cells

Conclusions

- Application of GATK-SV was able to discover a diverse set of SV in the OFC samples

- Adjudication with additional *de novo* filtering identified 165 *de novo* SV in 17.8% of probands

- We find both established OFC genes disrupted and novel candidate genes for further follow-up

- WGS has the resolution to detect complex SV and balanced SV not easily detectable by exome sequencing or microarrays

Future Directions

- Exploration of rare inherited SVs
- Examination of noncoding SV
- Integration of results with the SNV/Indel callset presented by Elizabeth Leslie
- Investigation of recessive and compound heterozygous variation
- Applying GATK-SV in additional GMKF cohorts to build an aggregated SV map of congenital birth defects

Acknowledgments

Broad GMKF & Broad-SV Sequencing and Analysis Teams

Harrison Brand
Michael Talkowski
Stacey Gabriel
Daniel MacArthur
Xuefang Zhao*
Stacey Mano
Ben Weisburd
Ryan Collins
Harold Wang
Mark Walker
Chris Wheelan
Candace Patterson

Gabriella Miller Kids First
Pediatric Research Program

NIH National Institute of Dental and Craniofacial Research

GMKF OFC Working Group

University of Iowa
Azeez Butali
Jeff Murray
Lina Moreno
Luz Consuelo Valencia-Ramirez
George Wehby
Andrew Lidral

Emory University
Elizabeth J. Leslie
Madison Bishop
Pankaj Chopra
Michael Mortillo
Dave Cutler
Michael Epstein

University of Pittsburgh
Mary L. Marazita
Seth M. Weinberg
Eleanor Feingold
Nandita Mukhopadhyay

Johns Hopkins University
Terri Beaty
Ingo Ruczinski
Margaret Taub
Alan Scott
Jacqueline Hetmanski
Debashree Ray

Other
Jacqueline Hecht
Andrew Czeizel
Yah-Huei Wu Chou
Frederic Deleyiannis
Adebowale Adeyemo
Mauricio Arcos-Burgos
Lord Gowans
Peter Mossey
Lanre Adeyemo
Philip Chen

Financial Support: National Institute of Dental & Craniofacial Research (K99DE026824, R00DE026824)

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

Allison Heath, PhD
Children's Hospital of Philadelphia
Co-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

Gabriella Miller Kids First Data Resource Center

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Kids First Data:

By the Numbers

1 2 3 4 5 6 7 8 9

Kids First Data - By the Numbers . . .

Studies
10

Kids First Data - By the Numbers . . .

Studies
10

Participants
10,560

Kids First Data - By the Numbers . . .

Studies
10

Participants
10,560

Families
3,679

Kids First Data - By the Numbers . . .

Studies	Participants	Families	Diagnoses
10	10,560	3,679	7,502

Kids First Data - By the Numbers . . .

Studies
10

Participants
10,560

Families
3,679

Diagnoses
7,502

Phenotypes
70,916

Kids First Data - By the Numbers . . .

Studies	10
Participants	10,560
Families	3,679
Diagnoses	7,502
Phenotypes	70,916
Genomes	10,901

Kids First Data - By the Numbers . . .

Studies
10

Participants
10,560

Families
3,679

Diagnoses
7,502

Phenotypes
70,916

Genomes
10,901

Size
~1 PB

New KFDRC Portal Features

Outline

Two new major Kids First DRC portal feature developments

- Enhanced ontology data model and search tool
- The germline variant data warehouse

Next steps/future directions

Ontologies within Kids First DRC

Kids First DRC makes extensive use of ontologies

Human Phenotype Ontology (HPO), Mondo, NCIT, SNOMED

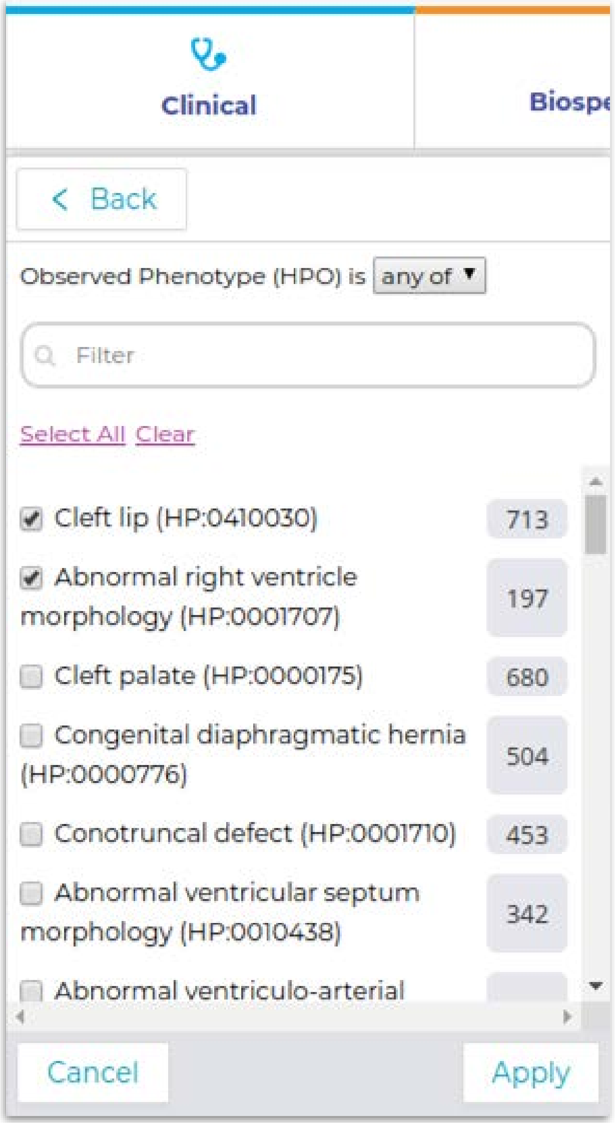
Ontologies provide both controlled vocabularies and "parent-child" relationships

E.g. *Oral cleft* **(HP:0000202) IS AN** *Abnormal oral cavity morphology* (HP:0000163)

New The portal now integrates relationships in participant search queries

Users can now find participants with a specific term ***and*** **all its descendant**

E.g. Searching for participants with *Abnormal oral cavity morphology* **will return** *Oral cleft* **participants**



Observed Phenotype (HPO) is any of ▾

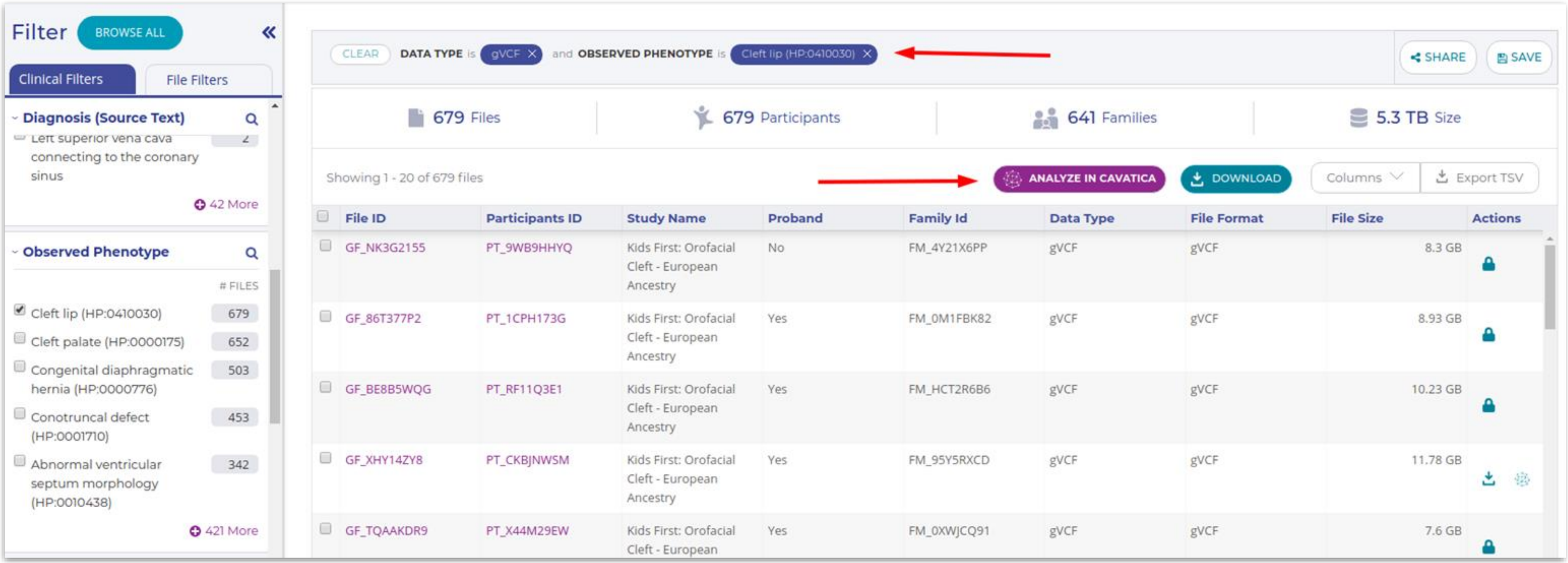
☑ Cleft lip (HP:0410030)	713
☑ Abnormal right ventricle morphology (HP:0001707)	197
☐ Cleft palate (HP:0000175)	680
☐ Congenital diaphragmatic hernia (HP:0000776)	504
☐ Conotruncal defect (HP:0001710)	453
☐ Abnormal ventricular septum morphology (HP:0010438)	342
☐ Abnormal ventriculo-arterial	

https://portal.kidsfirstdrc.org/explore

Variant Data within Kids First DRC

Currently available in gVCF files:

- Files can be searched using the portal's File Repository
- Selected files can be pushed to Cavatica for in-depth analyses



Filter BROWSE ALL «

Clinical Filters | File Filters

Diagnosis (Source Text)

□ Left superior vena cava connecting to the coronary sinus	2

⊕ 42 More

Observed Phenotype

	# FILES
☑ Cleft lip (HP:0410030)	679
□ Cleft palate (HP:0000175)	652
□ Congenital diaphragmatic hernia (HP:0000776)	503
□ Conotruncal defect (HP:0001710)	453
□ Abnormal ventricular septum morphology (HP:0010438)	342

⊕ 421 More

CLEAR **DATA TYPE** is gVCF ✕ and **OBSERVED PHENOTYPE** is Cleft lip (HP:0410030) ✕

SHARE SAVE

679 Files | 679 Participants | 641 Families | 5.3 TB Size

Showing 1 - 20 of 679 files

ANALYZE IN CAVATICA DOWNLOAD Columns Export TSV

□	File ID	Participants ID	Study Name	Proband	Family Id	Data Type	File Format	File Size	Actions
□	GF_NK3G2155	PT_9WB9HHYQ	Kids First: Orofacial Cleft - European Ancestry	No	FM_4Y21X6PP	gVCF	gVCF	8.3 GB	🔒
□	GF_86T377P2	PT_1CPH173G	Kids First: Orofacial Cleft - European Ancestry	Yes	FM_0M1FBK82	gVCF	gVCF	8.93 GB	🔒
□	GF_BE8B5WQG	PT_RF11Q3E1	Kids First: Orofacial Cleft - European Ancestry	Yes	FM_HCT2R6B6	gVCF	gVCF	10.23 GB	🔒
□	GF_XHY14ZY8	PT_CKBJNWSM	Kids First: Orofacial Cleft - European Ancestry	Yes	FM_95Y5RXCD	gVCF	gVCF	11.78 GB	⤓
□	GF_TQAAKDR9	PT_X44M29EW	Kids First: Orofacial Cleft - European	Yes	FM_0XWJCQ91	gVCF	gVCF	7.6 GB	🔒

New

The KFDRC Variant Data Warehouse Workspace Environment

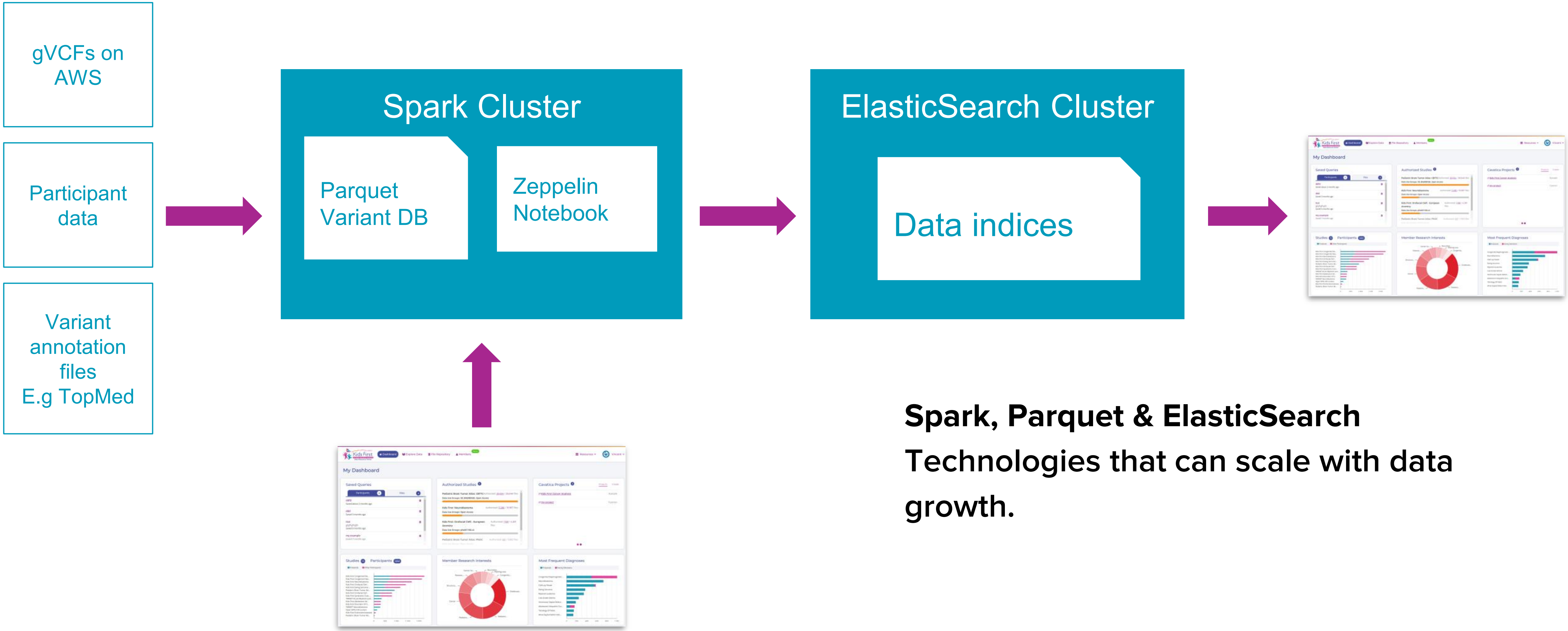
A performant and scalable variant database that can be queried directly from the portal

- Comprehensive set of variant annotations
  - Genes, allele frequencies, gene panels, inheritance, functional impact predictions, pathways, external references, etc.
- Individual-level clinical data integration to enable multi-dimensional queries
  - E.g. find all **rare missense** variants with **high functional impact** in **low grade glioma** patients affected by any **cardiovascular abnormalities**
- Web-based variant data analytics and visualisation tools
- Security and privacy rules enforcement
  - Users can only access variant datasets they have been authorized to

A *Big Data* Challenge

- High number of germline variants to process from whole genomes
- Current version
    - 8 studies, 8,100 participants, **251,801,242** unique variants, **42,513,213,093** occurrences
- For comparison/context
    - NCI Genomic Data Commons (GDC): 3.1 M somatic variants for ~10,000 cases
    - International Cancer Genome Consortium (ICGC): 82 M somatic variants for 19,700 cases
- Challenge: Complex data to query through **responsive** web interfaces
    - Link to extended individual-level clinical data
    - Integrate rich variant annotations

KFDRC Variant Data Processing Workflow



gVCFs on AWS

Participant data

Variant annotation files E.g TopMed

Spark Cluster

Parquet Variant DB

Zeppelin Notebook

ElasticSearch Cluster

Data indices

Spark, Parquet & ElasticSearch

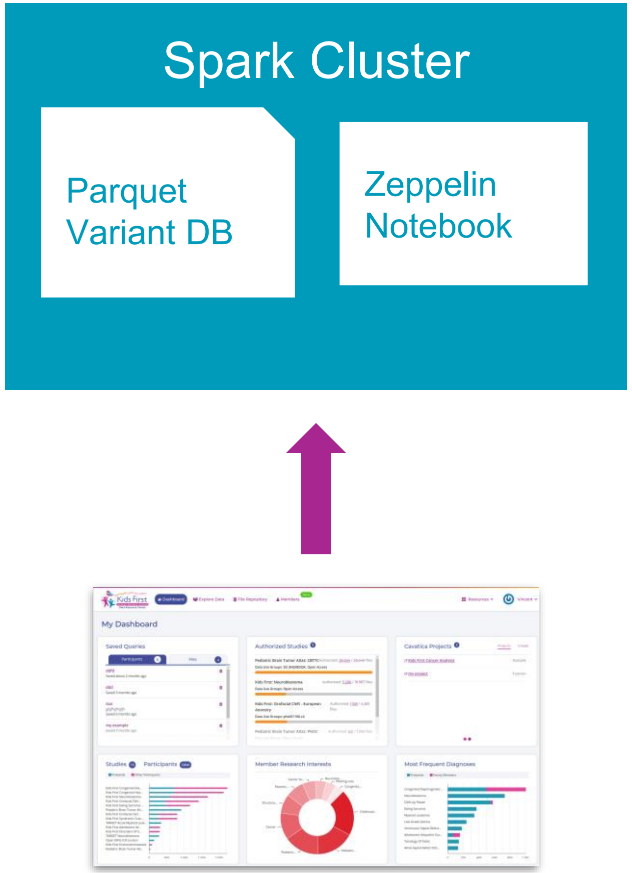
Technologies that can scale with data growth.

Phase I (First Release, Beta)

Foundation & Zeppelin notebooks

Objectives

- Build and deploy the foundational infrastructure of the KFDRC variant warehouse database
- Implement the data extraction, annotation and loading workflow
- Annotate variants with a limited (initially) set of annotations
- Implement the data security framework
- Provide researchers with the Zeppelin data analytic environment for querying and analysing the variant database
- Link the variant data analytic environment to the Cohort Builder, enabling researchers to analyse variants from their virtual patient cohorts

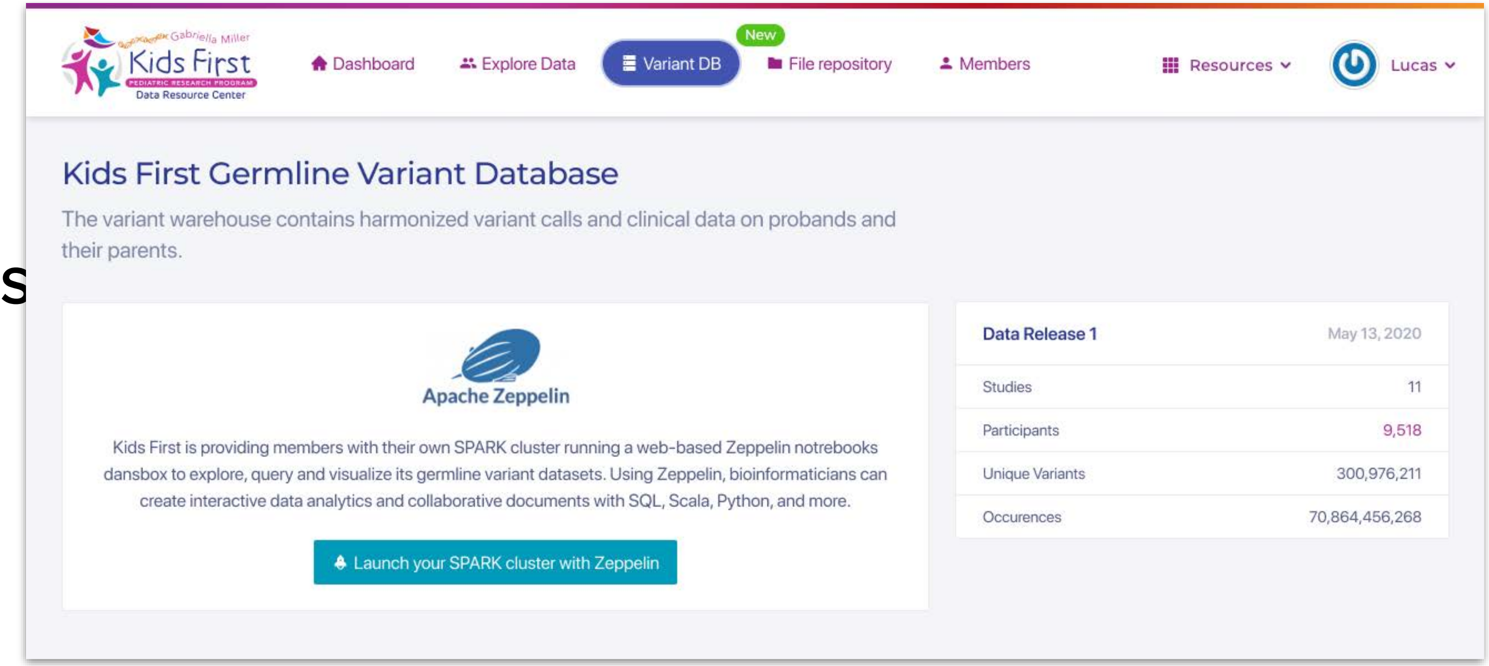


The Zeppelin Data Analytic Environment

Provides programmatic access to the variant database from web browsers

- Accessible from the Portal
- User notebook workspace
- Private/Individual Spark clusters on AWS
- Support for various programmatic languages (SQL, Python, R, Scala)

https://portal.kidsfirstdrc.org/variantDb



Kids First Germline Variant Database

The variant warehouse contains harmonized variant calls and clinical data on probands and their parents.

Kids First is providing members with their own SPARK cluster running a web-based Zeppelin notrebooks dansbox to explore, query and visualize its germline variant datasets. Using Zeppelin, bioinformaticians can create interactive data analytics and collaborative documents with SQL, Scala, Python, and more.

Data Release 1	May 13, 2020
Studies	11
Participants	9,518
Unique Variants	300,976,211
Occurences	70,864,456,268

Demo

Next Steps

- *Performance tests and data quality control (QC)*
- *Beta release to KFDRC user groups and X01 Investigators*
- *User testing and feedback integrations*

Additional Short Term Development Road map

- Indexing variant data warehouse using Elasticsearch
- Build data querying interfaces within the portal (integrating notebook use cases)
  - GA4GH-like Beacon service (return yes/no answers on variant occurrences)
  - Gnomad-like Summary interface (mainly allele frequencies)
  - Direct integration within the Cohort Builder allowing complex queries that return *both* participant and variant lists
- More annotations supporting variant prioritization

Special Thanks To

CHU Ste-Justine Research Center

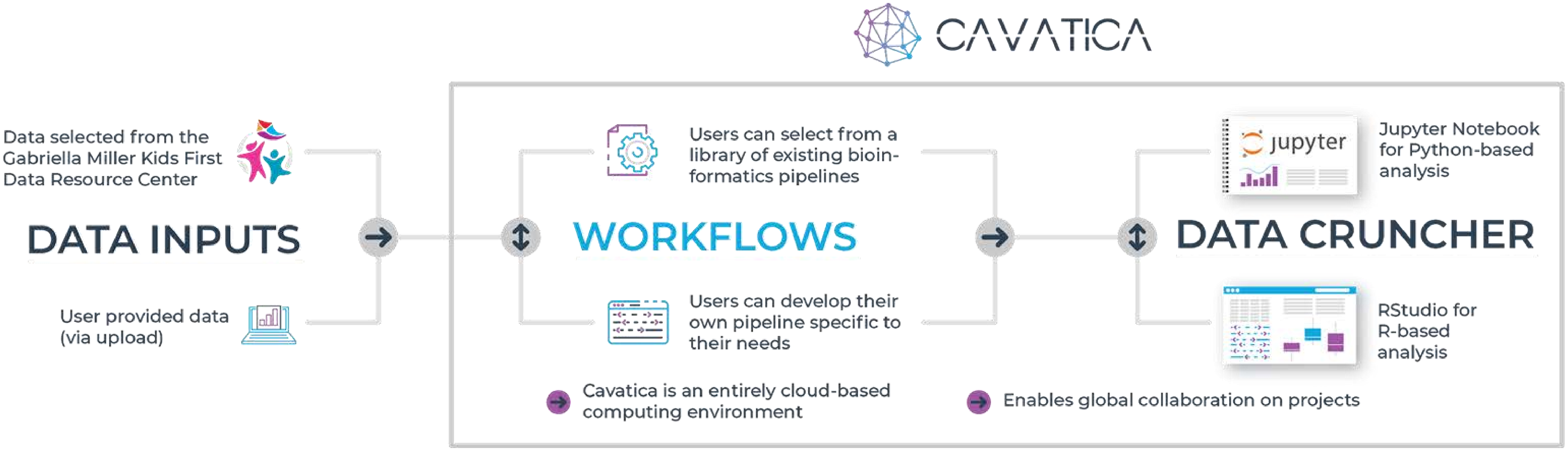
- **Jeremy Costanza**, Lead software architect and developer
- Developers
  - Adrian Paul
  - Evans Girard
  - Francis Lavoie
- UX
  - Lucas Lemonnier

CHOP

- DevOps Lead
  - Alex Lubneuski
- Bioinformatics Lead
  - Yuankun Zhu

CAVATICA: Cloud User Workspace Introduction

CAVATICA: Integrated Cloud-Based Workspace



CAVATICA

Data selected from the Gabriella Miller Kids First Data Resource Center

DATA INPUTS

User provided data (via upload)

Users can select from a library of existing bioinformatics pipelines

WORKFLOWS

Users can develop their own pipeline specific to their needs

Jupyter Notebook for Python-based analysis

DATA CRUNCHER

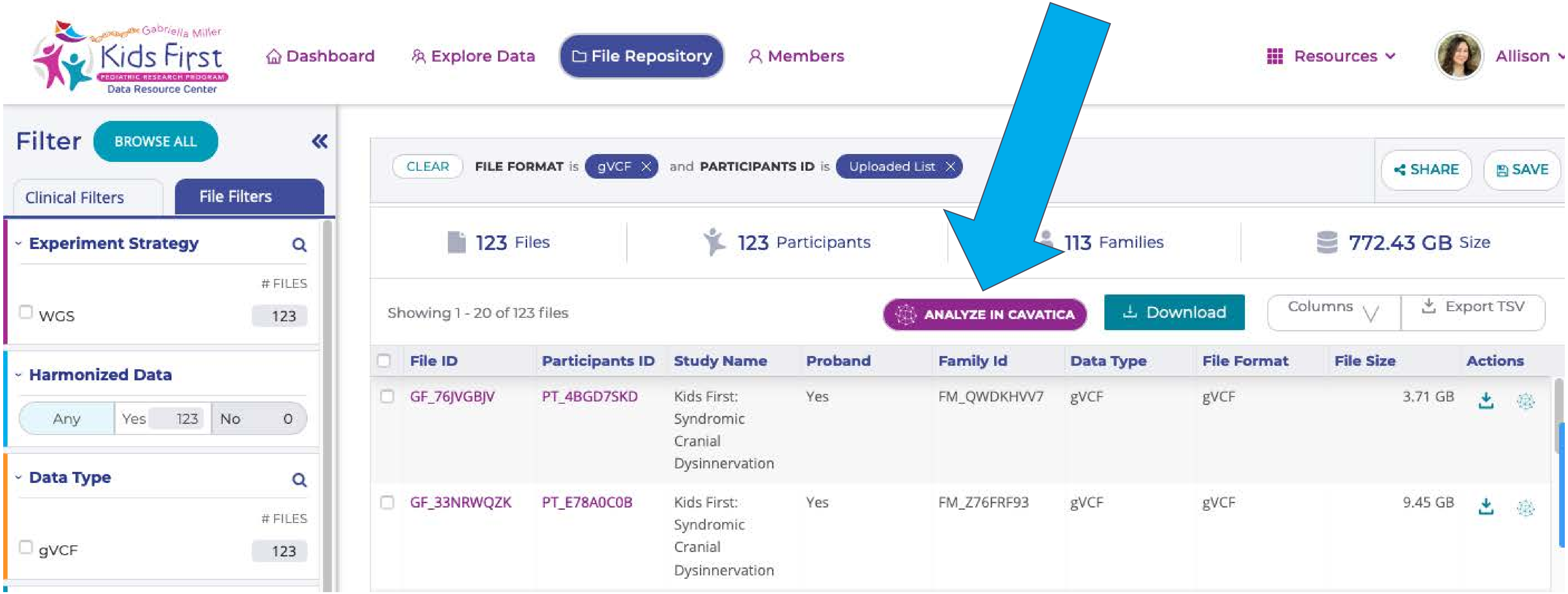
RStudio for R-based analysis

Cavatica is an entirely cloud-based computing environment

Enables global collaboration on projects

Data Inputs - Kids First

Immediately bring over files you're authorized to use into CAVATICA



Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

Experiment Strategy

	# FILES
□ WGS	123

Harmonized Data

Any | Yes 123 | No 0

Data Type

	# FILES
□ gVCF	123

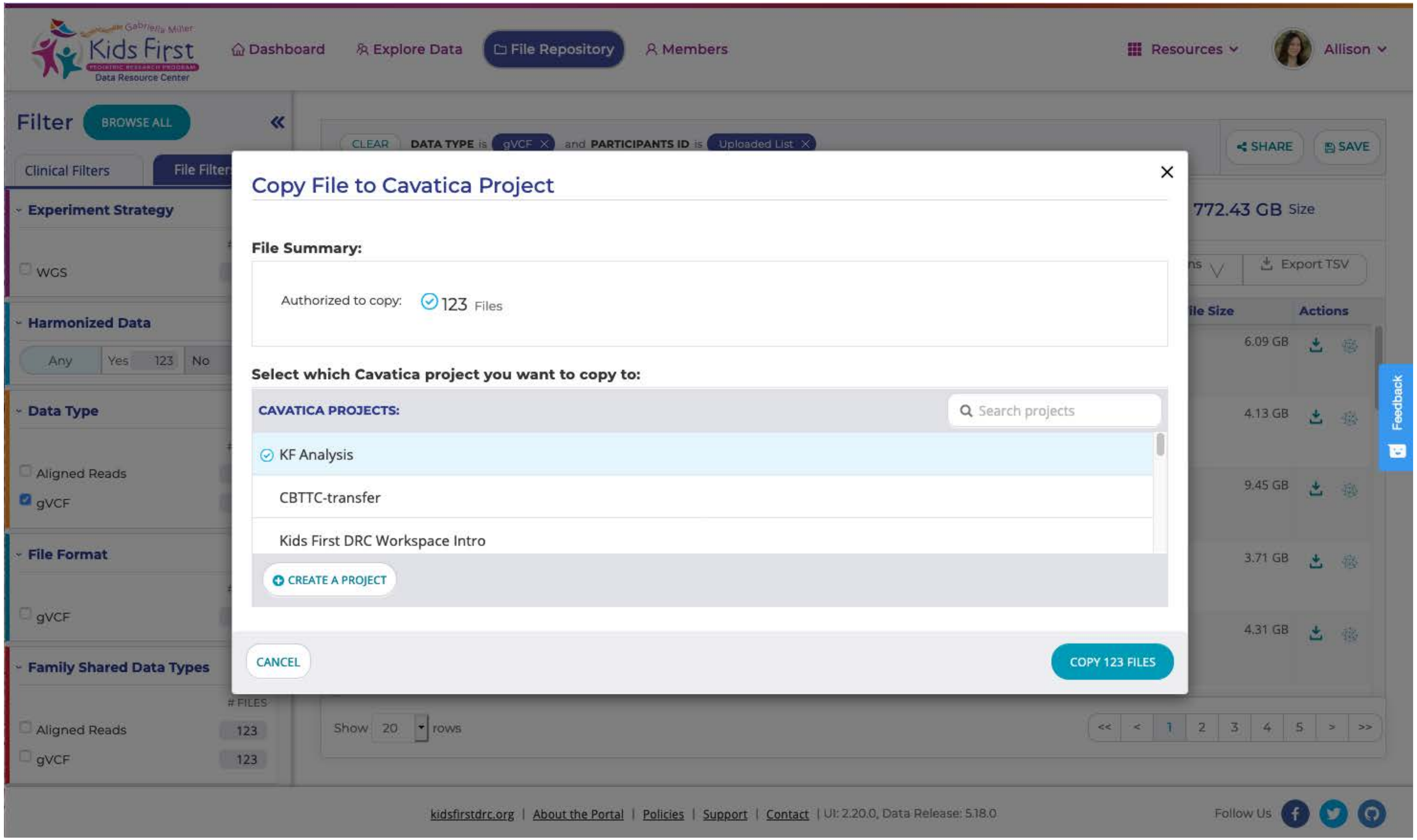
CLEAR **FILE FORMAT** is gVCF and **PARTICIPANTS ID** is Uploaded List

123 Files | 123 Participants | 113 Families | 772.43 GB Size

Showing 1 - 20 of 123 files

□	File ID	Participants ID	Study Name	Proband	Family Id	Data Type	File Format	File Size	Actions
□	GF_76JVGBJV	PT_4BGD7SKD	Kids First: Syndromic Cranial Dysinnervation	Yes	FM_QWDKHVV7	gVCF	gVCF	3.71 GB	
□	GF_33NRWQZK	PT_E78A0C0B	Kids First: Syndromic Cranial Dysinnervation	Yes	FM_Z76FRF93	gVCF	gVCF	9.45 GB	

Data Inputs - Kids First



Copy File to Cavatica Project

File Summary:

Authorized to copy: ✓ 123 Files

Select which Cavatica project you want to copy to:

CAVATICA PROJECTS:

Search projects

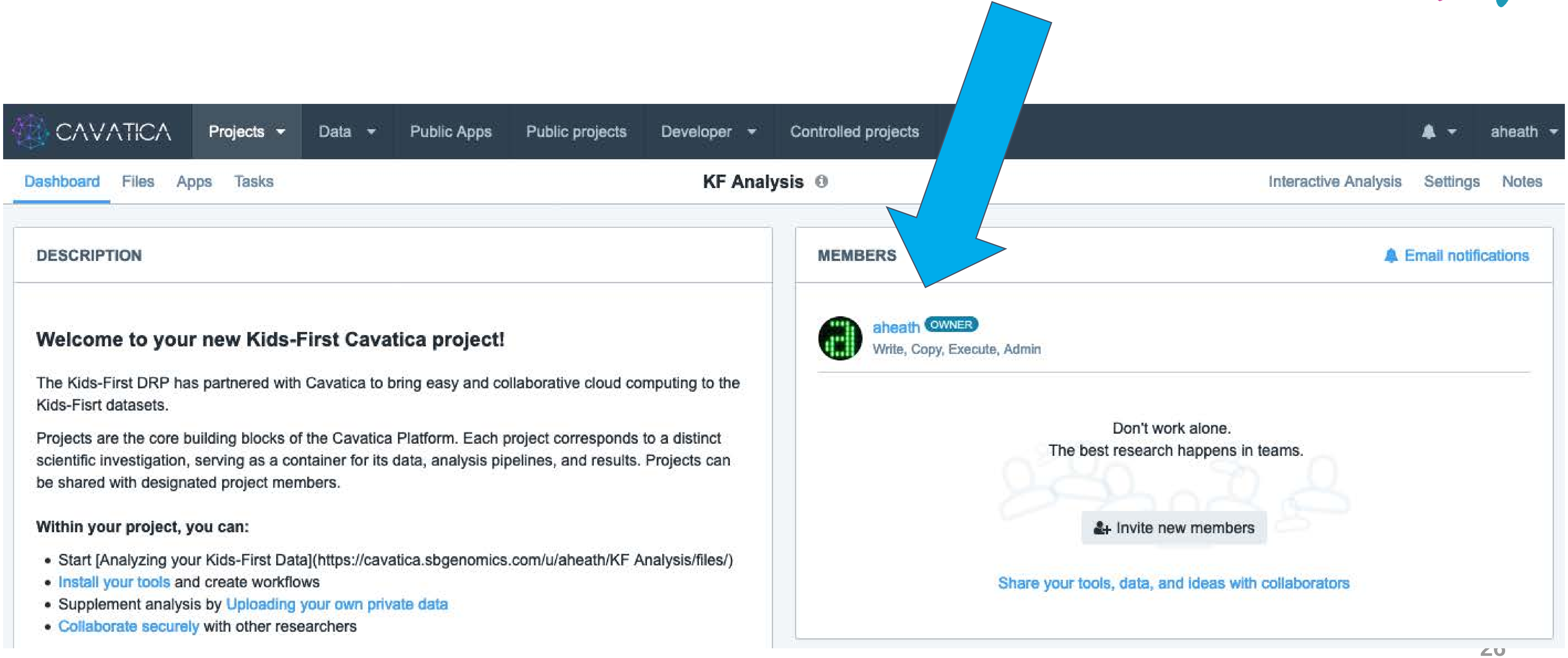
✓ KF Analysis

CBTTC-transfer

Kids First DRC Workspace Intro

Data Inputs - Kids First

New projects created in CAVATICA start as private



Welcome to your new Kids-First Cavatica project!

The Kids-First DRP has partnered with Cavatica to bring easy and collaborative cloud computing to the Kids-Fisrt datasets.

Projects are the core building blocks of the Cavatica Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects can be shared with designated project members.

Within your project, you can:

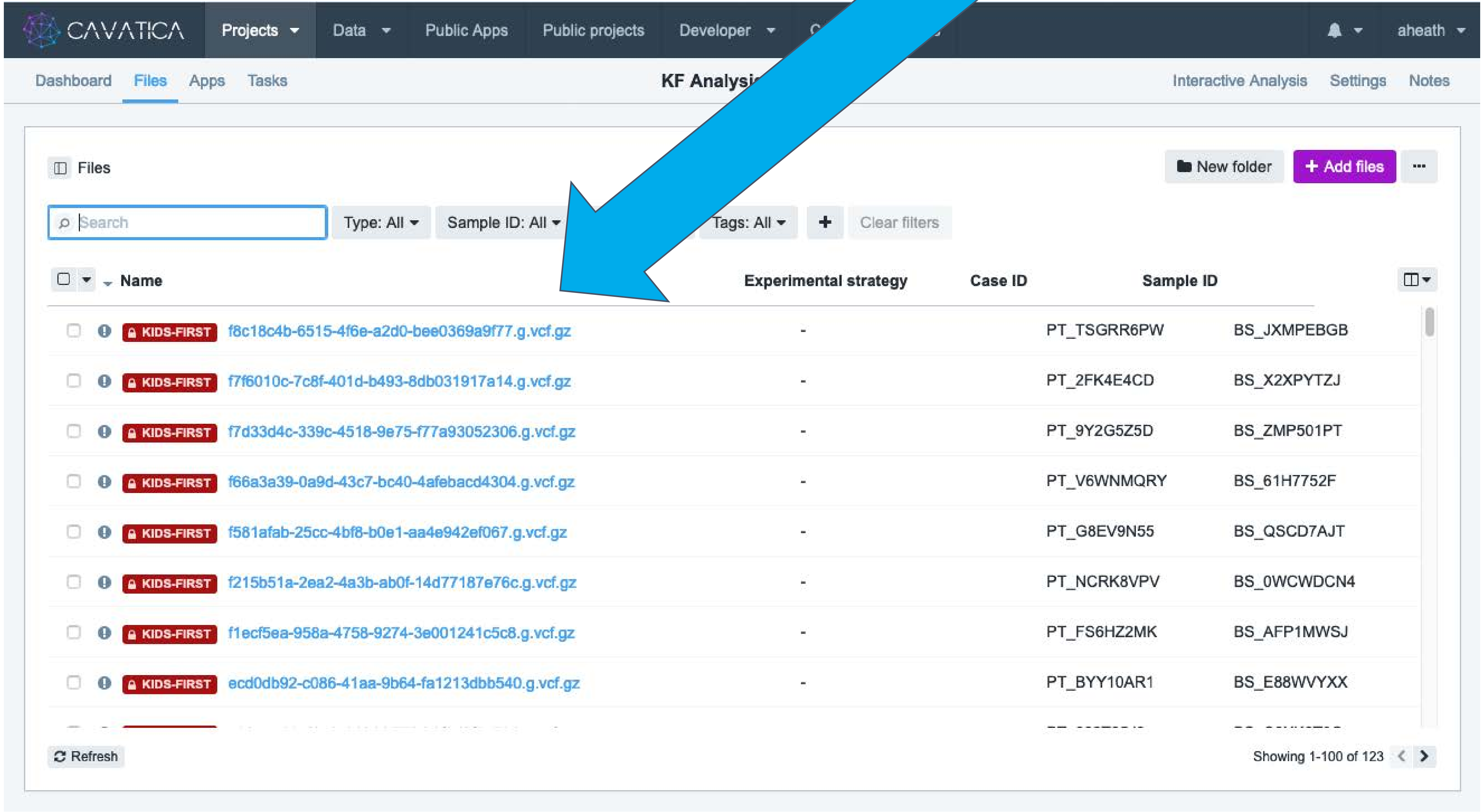
- Start [Analyzing your Kids-First Data](https://cavatica.sbgenomics.com/u/aheath/KF Analysis/files/)
- Install your tools and create workflows
- Supplement analysis by Uploading your own private data
- Collaborate securely with other researchers

Don't work alone.
The best research happens in teams.

Share your tools, data, and ideas with collaborators

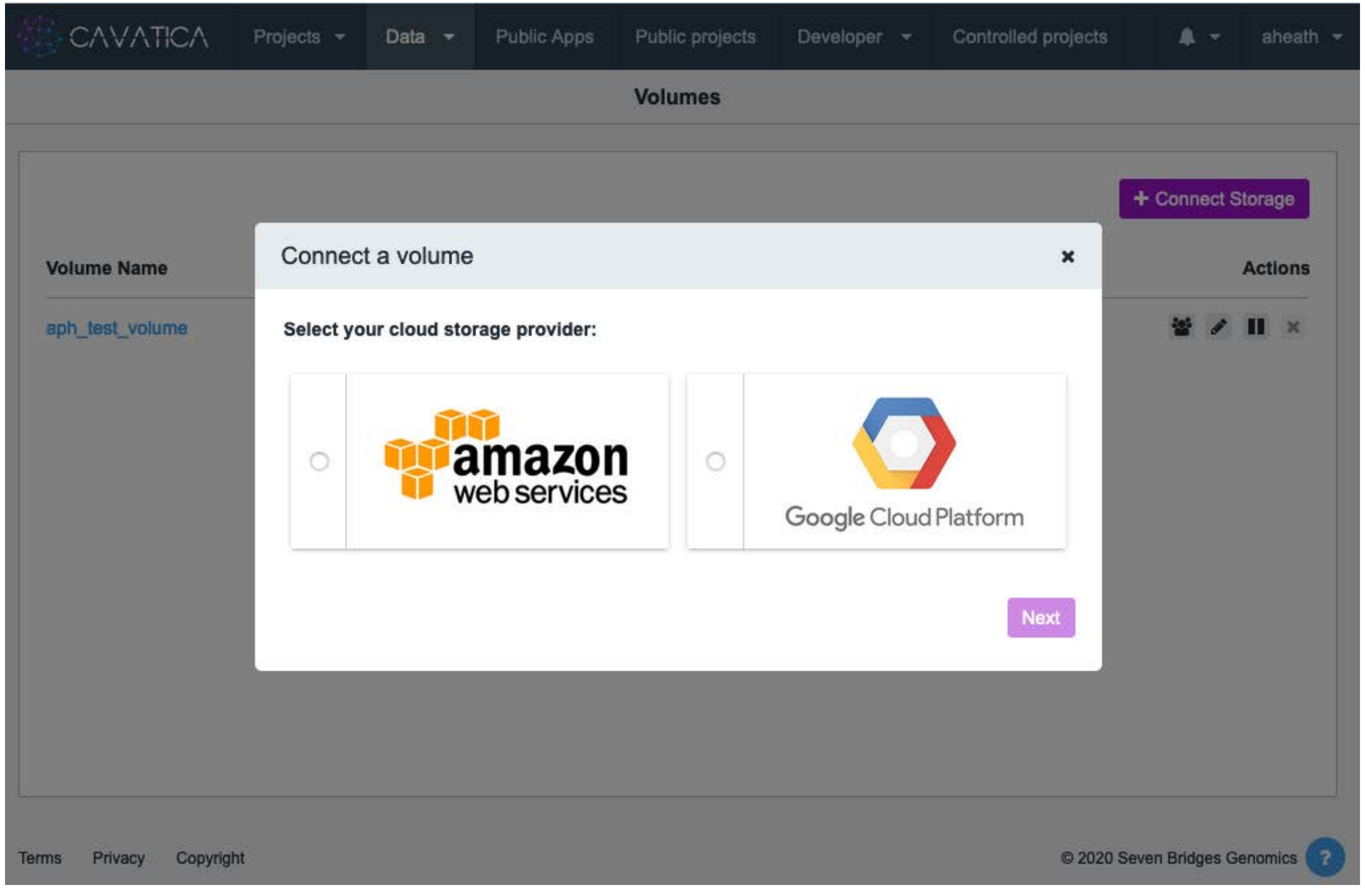
Data Inputs - Kids First

CAVATICA continues to check permissions for integrated datasets



	Name	Experimental strategy	Case ID	Sample ID
KIDS-FIRST	f8c18c4b-6515-4f6e-a2d0-bee0369a9f77.g.vcf.gz	-	PT_TSGRR6PW	BS_JXMPEBGB
KIDS-FIRST	f7f6010c-7c8f-401d-b493-8db031917a14.g.vcf.gz	-	PT_2FK4E4CD	BS_X2XPYTZJ
KIDS-FIRST	f7d33d4c-339c-4518-9e75-f77a93052306.g.vcf.gz	-	PT_9Y2G5Z5D	BS_ZMP501PT
KIDS-FIRST	f66a3a39-0a9d-43c7-bc40-4afebacd4304.g.vcf.gz	-	PT_V6WNMQRY	BS_61H7752F
KIDS-FIRST	f581afab-25cc-4bf8-b0e1-aa4e942ef067.g.vcf.gz	-	PT_G8EV9N55	BS_QSCD7AJT
KIDS-FIRST	f215b51a-2ea2-4a3b-ab0f-14d77187e76c.g.vcf.gz	-	PT_NCRK8VPV	BS_0WCWDCN4
KIDS-FIRST	f1ecf5ea-958a-4758-9274-3e001241c5c8.g.vcf.gz	-	PT_FS6HZ2MK	BS_AFP1MWSJ
KIDS-FIRST	ecd0db92-c086-41aa-9b64-fa1213dbb540.g.vcf.gz	-	PT_BYY10AR1	BS_E88WVYXX

Data Inputs - Own Data on AWS or GCP



Connect a volume ×

Select your cloud storage provider:

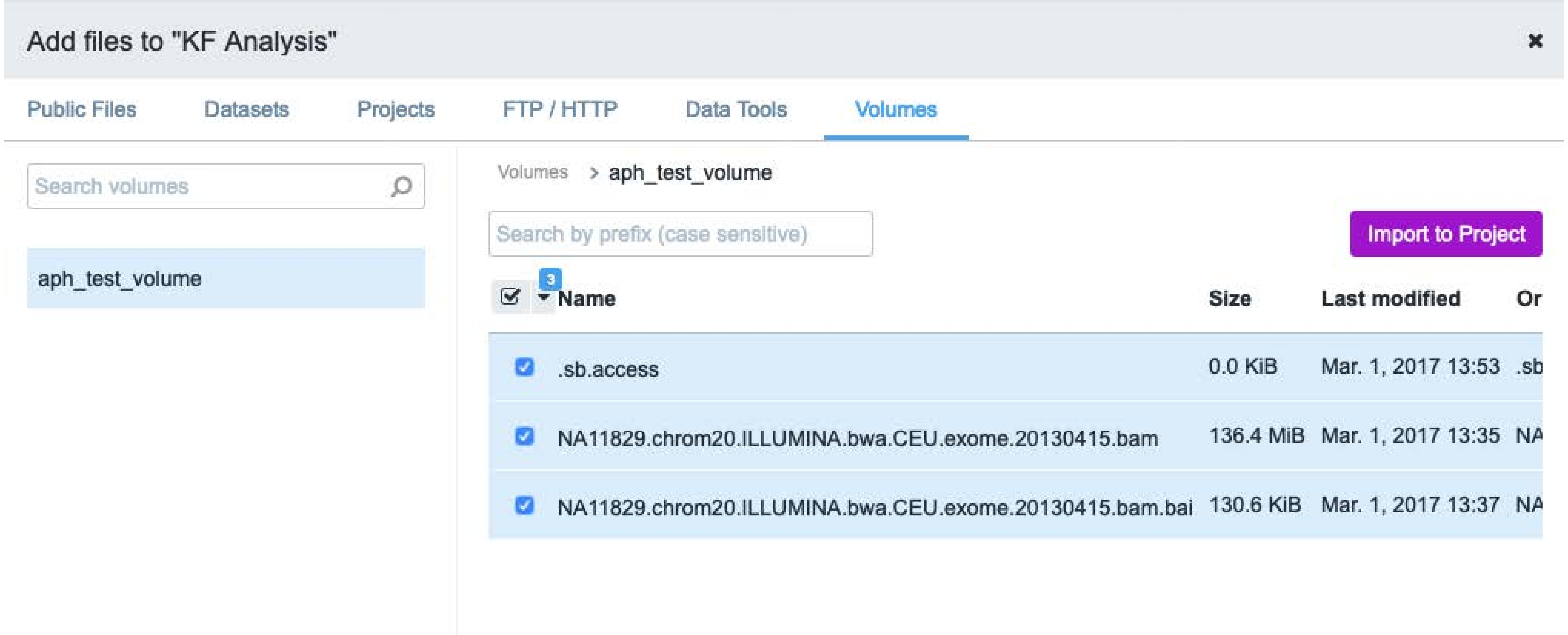
○ amazon web services

○ Google Cloud Platform

Next

Via the Data->Volumes Features

Data Inputs - Own Data on AWS or GCP

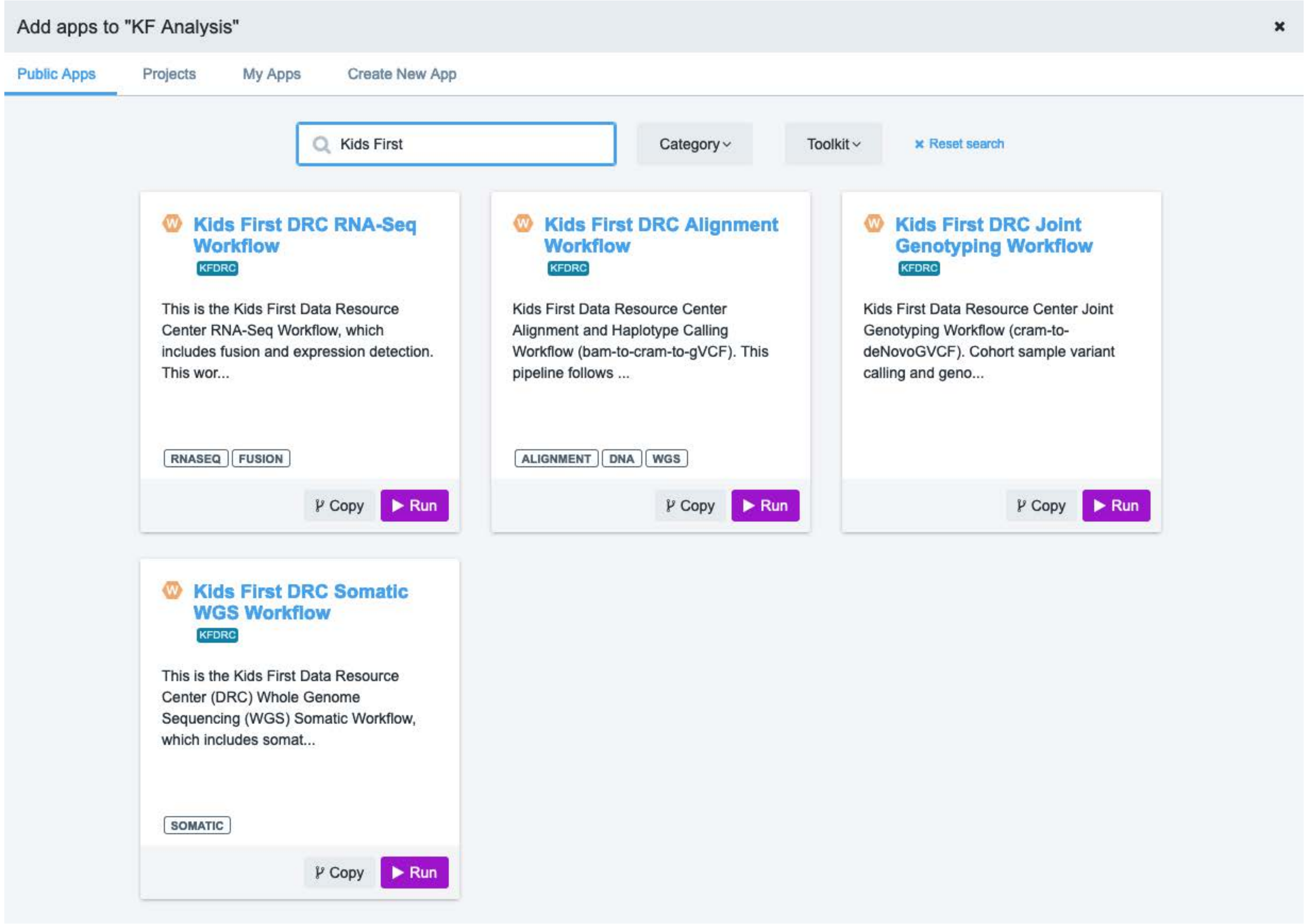


Add files to "KF Analysis" ✖

☑ 3	Name	Size	Last modified	Or
☑	.sb.access	0.0 KiB	Mar. 1, 2017 13:53	.sb
☑	NA11829.chrom20.ILLUMINA.bwa.CEU.exome.20130415.bam	136.4 MiB	Mar. 1, 2017 13:35	NA
☑	NA11829.chrom20.ILLUMINA.bwa.CEU.exome.20130415.bam.bai	130.6 KiB	Mar. 1, 2017 13:37	NA

After adding own files to same private project with Kids First files, can now utilize existing or bring your own workflows

Workflows - Existing Examples



Add apps to "KF Analysis"

Kids First DRC RNA-Seq Workflow

KFDRC

This is the Kids First Data Resource Center RNA-Seq Workflow, which includes fusion and expression detection. This wor...

RNASEQ FUSION

Kids First DRC Alignment Workflow

KFDRC

Kids First Data Resource Center Alignment and Haplotype Calling Workflow (bam-to-cram-to-gVCF). This pipeline follows ...

ALIGNMENT DNA WGS

Kids First DRC Joint Genotyping Workflow

KFDRC

Kids First Data Resource Center Joint Genotyping Workflow (cram-to-deNovoGVCF). Cohort sample variant calling and geno...

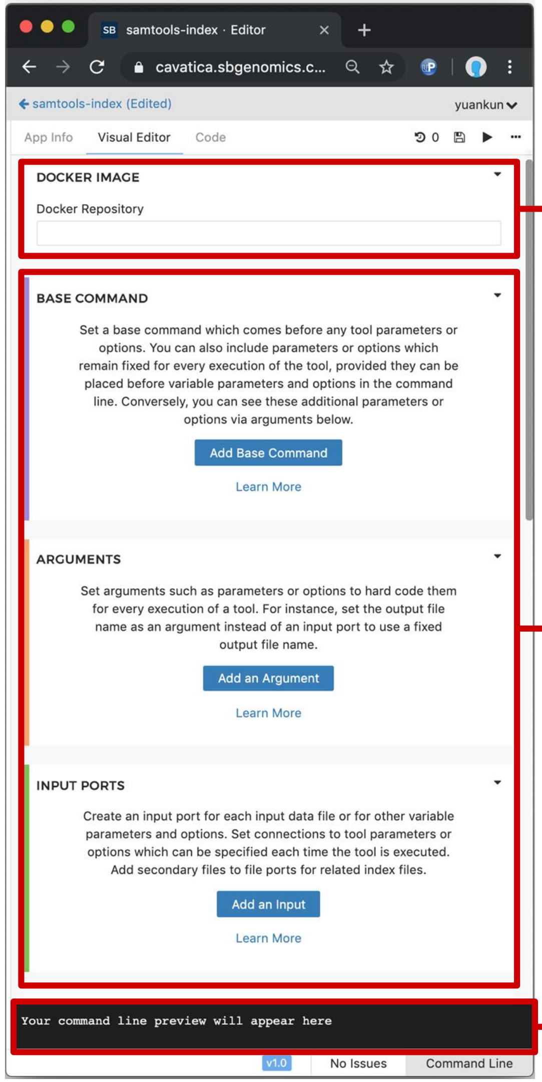
Kids First DRC Somatic WGS Workflow

KFDRC

This is the Kids First Data Resource Center (DRC) Whole Genome Sequencing (WGS) Somatic Workflow, which includes somat...

SOMATIC

Workflows - Port Your Own



DOCKER IMAGE

Docker Repository

BASE COMMAND

Set a base command which comes before any tool parameters or options. You can also include parameters or options which remain fixed for every execution of the tool, provided they can be placed before variable parameters and options in the command line. Conversely, you can see these additional parameters or options via arguments below.

ARGUMENTS

Set arguments such as parameters or options to hard code them for every execution of a tool. For instance, set the output file name as an argument instead of an input port to use a fixed output file name.

INPUT PORTS

Create an input port for each input data file or for other variable parameters and options. Set connections to tool parameters or options which can be specified each time the tool is executed. Add secondary files to file ports for related index files.

```
Your command line preview will appear here
```

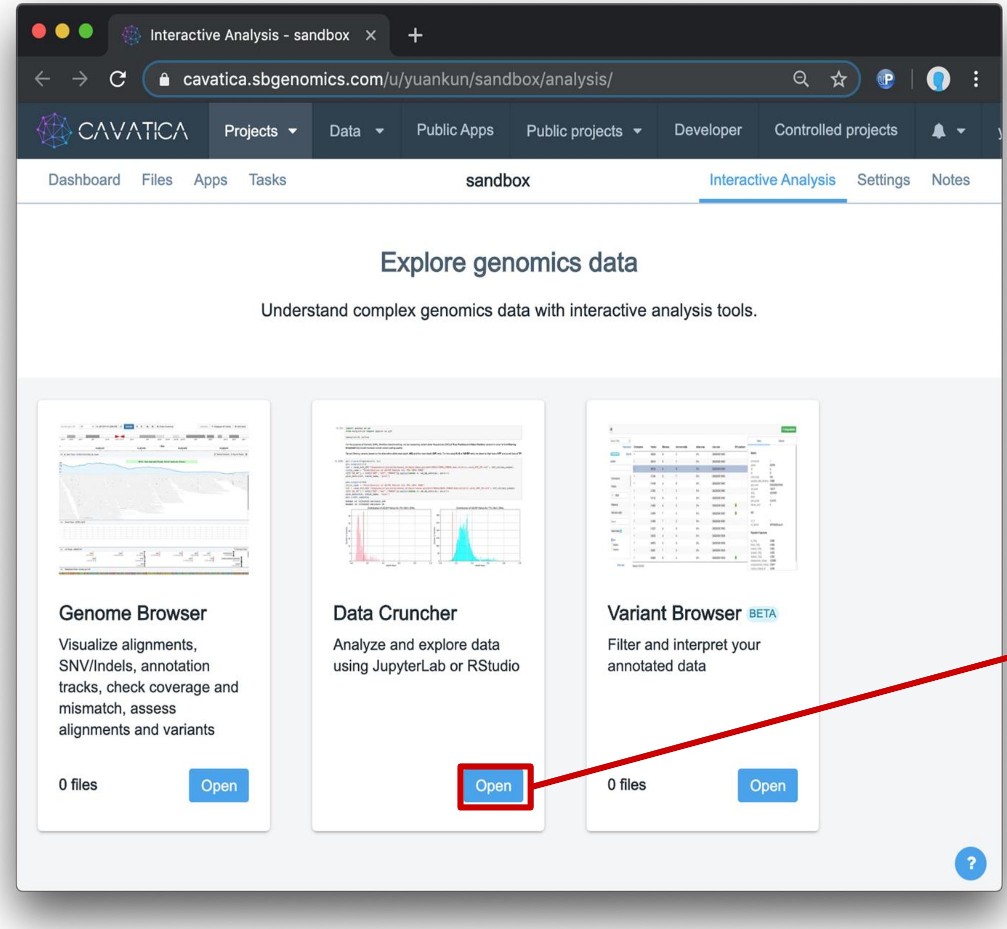
Tool docker image

Tool details, basic command line, input/output arguments, parameters etc

Final command line preview

COMMON WORKFLOW LANGUAGE

Data Cruncher - Interactive Analysis



Interactive Analysis - sandbox

cavatica.sbgenomics.com/u/yuankun/sandbox/analysis/

Explore genomics data

Understand complex genomics data with interactive analysis tools.

Genome Browser

Visualize alignments, SNV/Indels, annotation tracks, check coverage and mismatch, assess alignments and variants

0 files — Open

Data Cruncher

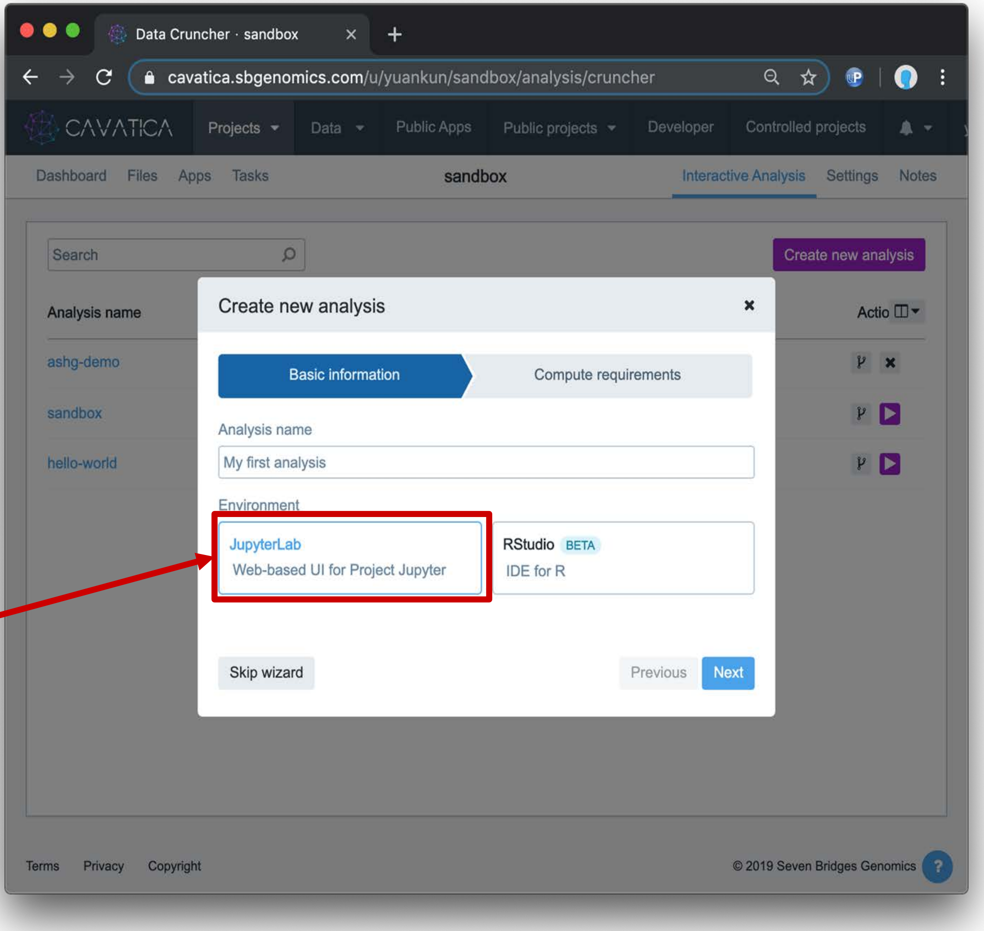
Analyze and explore data using JupyterLab or RStudio

Open

Variant Browser BETA

Filter and interpret your annotated data

0 files — Open



Data Cruncher · sandbox

cavatica.sbgenomics.com/u/yuankun/sandbox/analysis/cruncher

Create new analysis

Basic information | Compute requirements

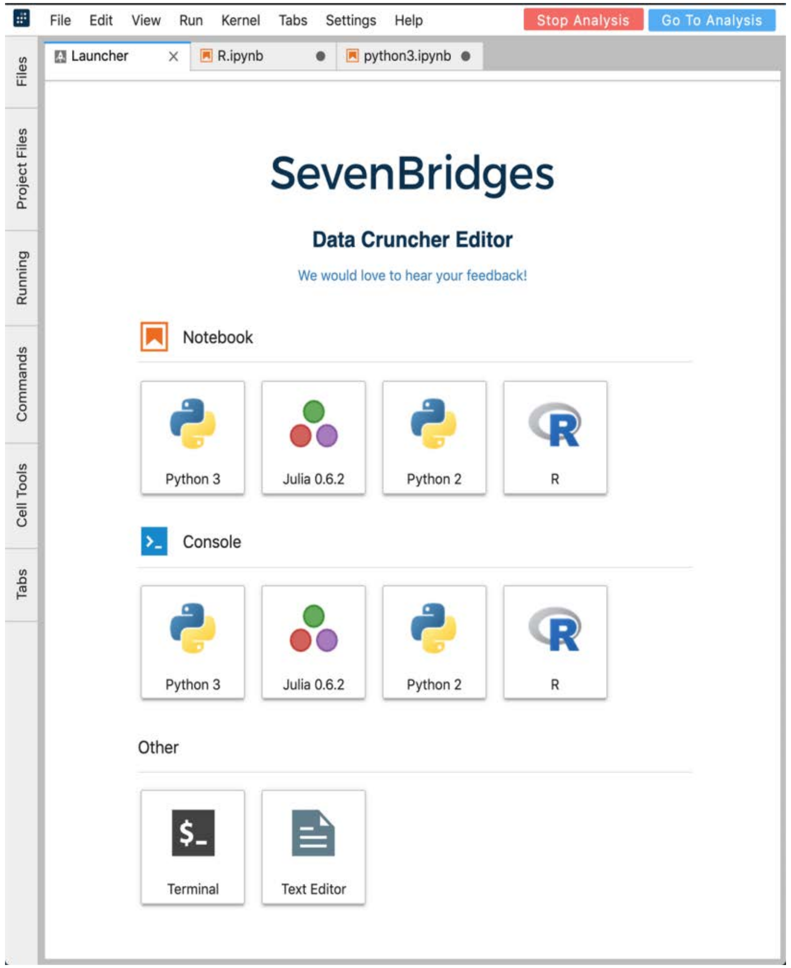
Analysis name: My first analysis

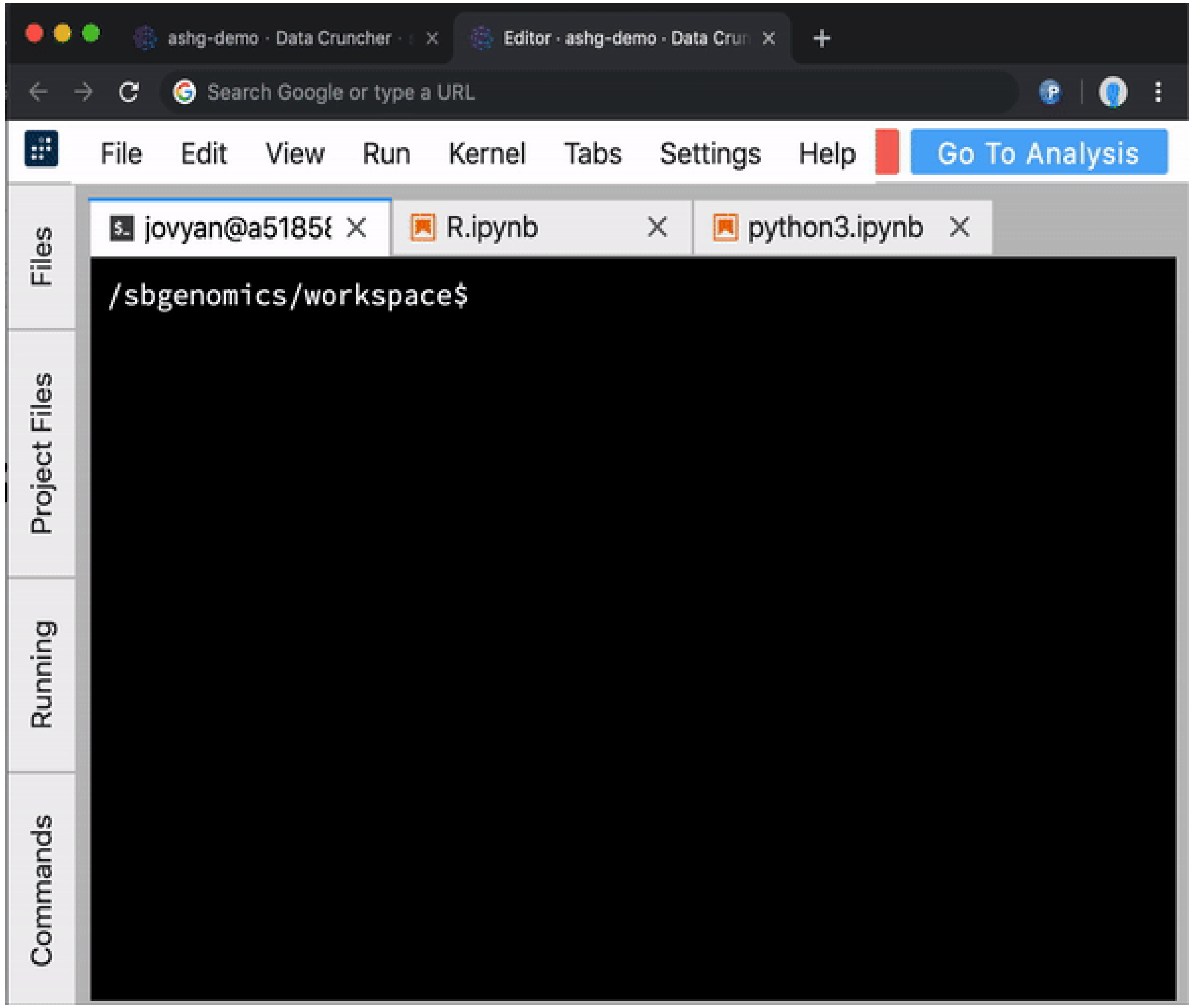
Environment:

- JupyterLab — Web-based UI for Project Jupyter
- RStudio BETA — IDE for R

Skip wizard — Previous — Next

Data Cruncher - Interactive Analysis

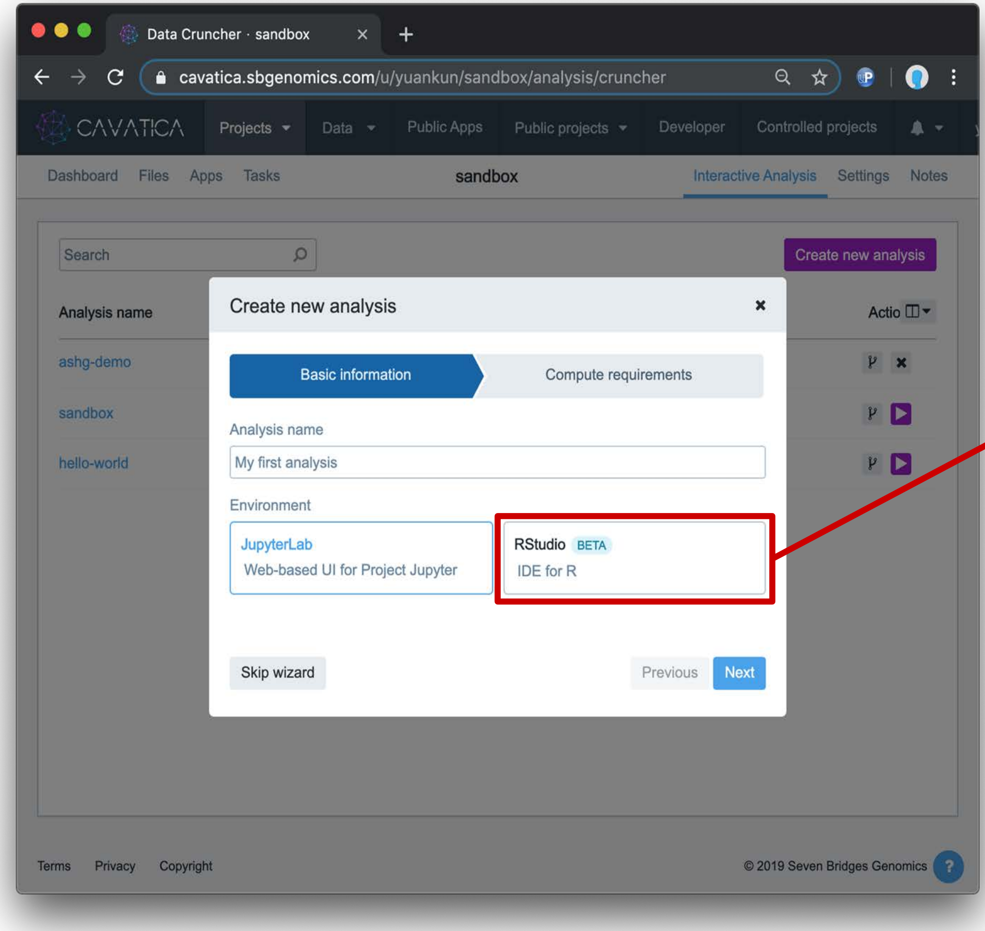




ashg-demo · Data Cruncher
Editor · ashg-demo · Data Crun
Search Google or type a URL
File Edit View Run Kernel Tabs Settings Help
Go To Analysis
Files
Project Files
Running
Commands
jovyan@a5185
R.ipynb
python3.ipynb

```
/sbgenomics/workspace$
```

Data Cruncher - RStudio and Shiny Apps



Create new analysis

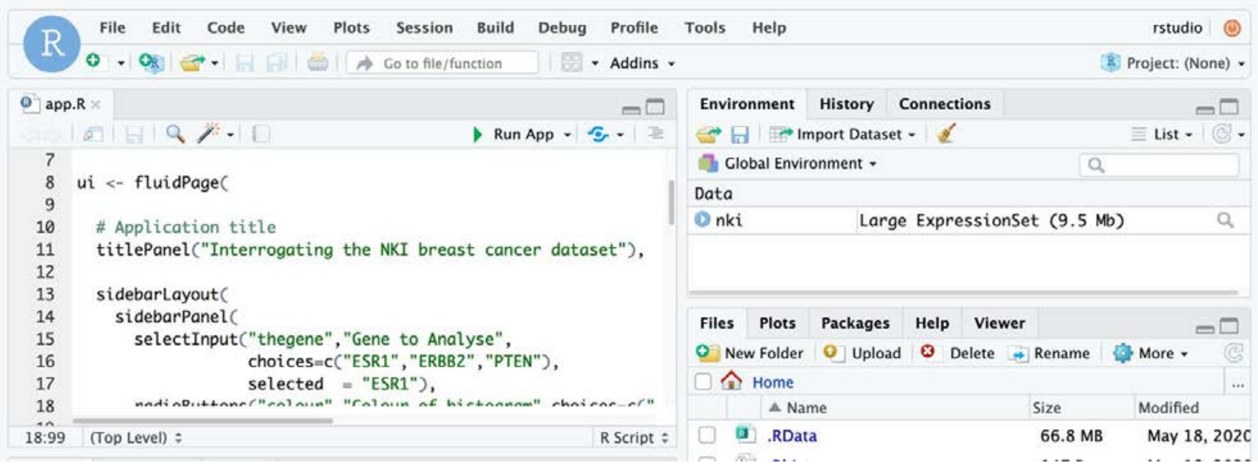
Basic information | Compute requirements

Analysis name: My first analysis

Environment

- JupyterLab — Web-based UI for Project Jupyter
- RStudio BETA — IDE for R

Skip wizard | Previous | Next



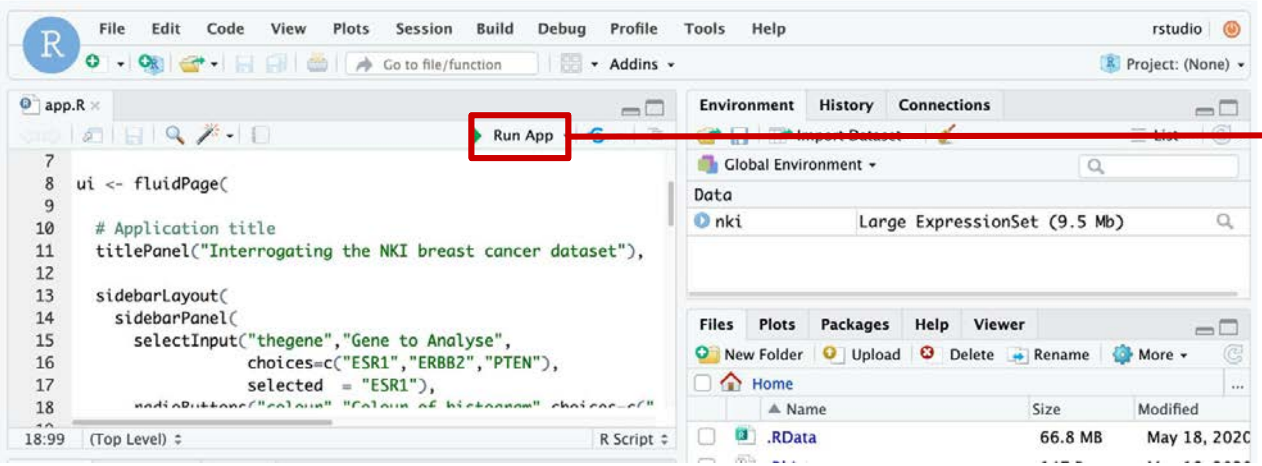
```r

ui <- fluidPage(

  # Application title
  titlePanel("Interrogating the NKI breast cancer dataset"),

  sidebarLayout(
    sidebarPanel(
      selectInput("thegene","Gene to Analyse",
                  choices=c("ESR1","ERBB2","PTEN"),
                  selected  = "ESR1"),
      radioButtons("colour","Colour of histogram",choices=c("
```

Data

nki   Large ExpressionSet (9.5 Mb)

	Name	Size	Modified
□	.RData	66.8 MB	May 18, 202C

Data Cruncher - RStudio and Shiny Apps

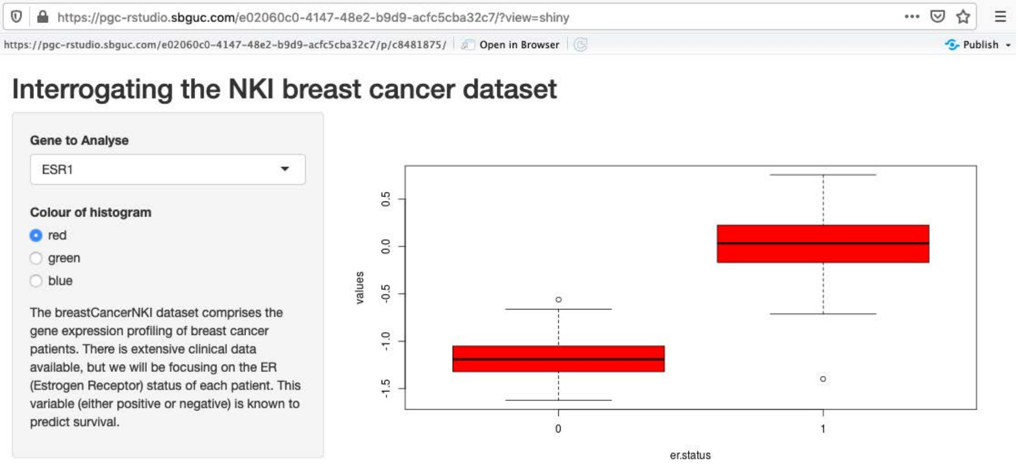


```r
ui <- fluidPage(

  # Application title
  titlePanel("Interrogating the NKI breast cancer dataset"),

  sidebarLayout(
    sidebarPanel(
      selectInput("thegene","Gene to Analyse",
                  choices=c("ESR1","ERBB2","PTEN"),
                  selected  = "ESR1"),
      radioButtons("colour","Colour of histogram",choices=c("
```

Data

nki	Large ExpressionSet (9.5 Mb)

Name	Size	Modified
.RData	66.8 MB	May 18, 2020


https://pgc-rstudio.sbguc.com/e02060c0-4147-48e2-b9d9-acfc5cba32c7/?view=shiny
https://pgc-rstudio.sbguc.com/e02060c0-4147-48e2-b9d9-acfc5cba32c7/p/c8481875/
Open in Browser
Publish
0.5
0.0
-0.5
-1.0
-1.5
values
0
1
er.status

Interrogating the NKI breast cancer dataset

Gene to Analyse

ESR1

Colour of histogram

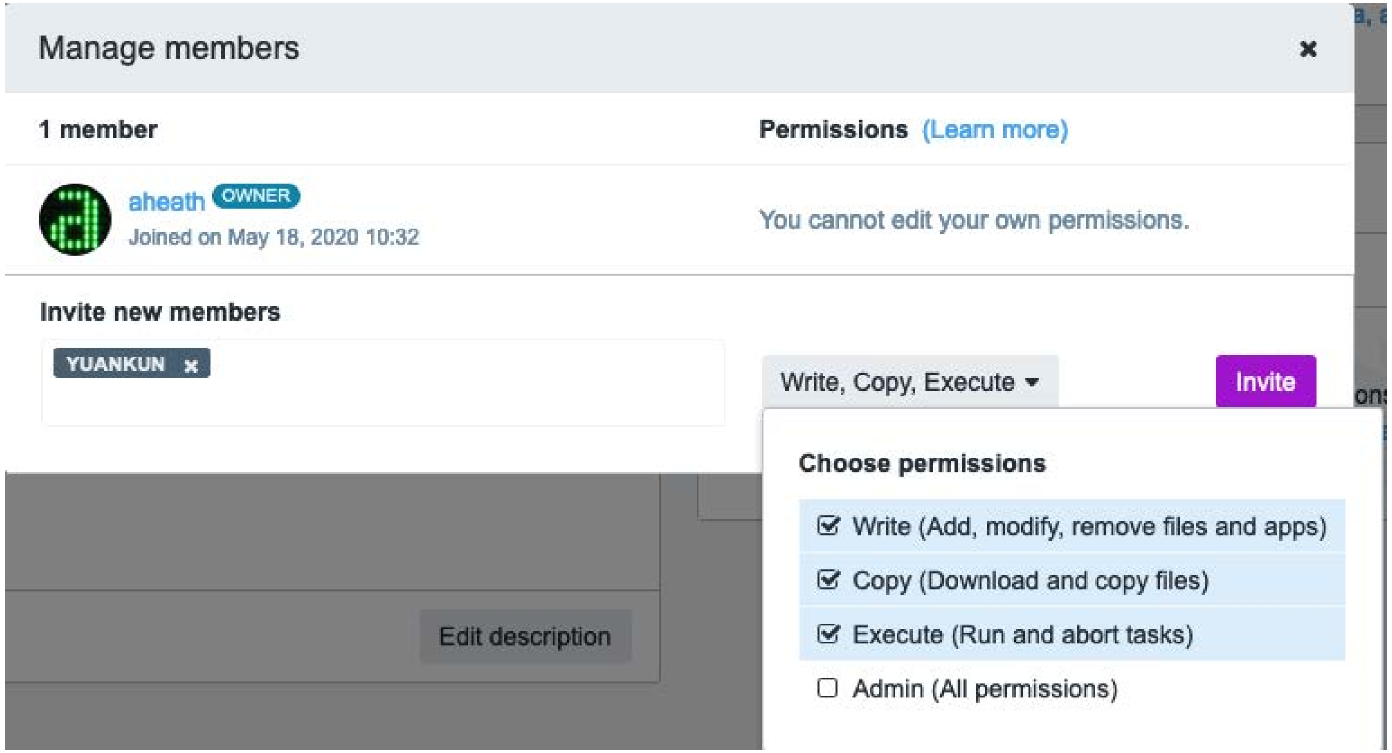
● red
○ green
○ blue

The breastCancerNKI dataset comprises the gene expression profiling of breast cancer patients. There is extensive clinical data available, but we will be focusing on the ER (Estrogen Receptor) status of each patient. This variable (either positive or negative) is known to predict survival.

Beta Feature!

All Features are Collaborative

You maintain control of access to your projects and data.



Manage members ×

1 member

Permissions (Learn more)

aheath OWNER
Joined on May 18, 2020 10:32

You cannot edit your own permissions.

Invite new members

YUANKUN ×

Write, Copy, Execute ▾

Invite

Choose permissions

☑ Write (Add, modify, remove files and apps)

☑ Copy (Download and copy files)

☑ Execute (Run and abort tasks)

☐ Admin (All permissions)

User Workspace Demonstration: Owen Hirschi, Baylor College of Medicine

Whole Genome Sequencing Analysis of the BASIC3 Childhood Cancer Cohort

May 18th, 2020

Owen Hirschi

Dr. Sharon Plon's Lab

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

Texas Children's Cancer Center

Baylor College of Medicine

CAVATICA

Germline mutations in cancer susceptibility genes occur in 8-10% of pediatric cancers

Germline susceptibility genes include: RB1, NF1, WT1 etc.

Susceptibility genes and Wilms Tumor

1991

WT1 first cloned

multiple other susceptibility genes discovered

2019

4 new genes discovered from WES

Not all cancer predisposition or susceptibility genes have been identified

Probands from BASIC3 have undergone germline and somatic WES

NIH BASIC3

BCM Advancing Sequencing
Into Childhood Cancer Care

Goal: characterize the diagnostic yield of combined tumor and germline WES for 287 children with solid tumors

- Not enriched for specific cancer type between CNS and non-CNS tumors

- Found pathogenic variants in
  - Genes with associated with specific cancers
  - Genes not previously associated with specific cancers

D.W. Parsons et al., 2016
dbGaP: phs001026.v2.p1

120 probands-parents from BASIC3 selected for germline WGS

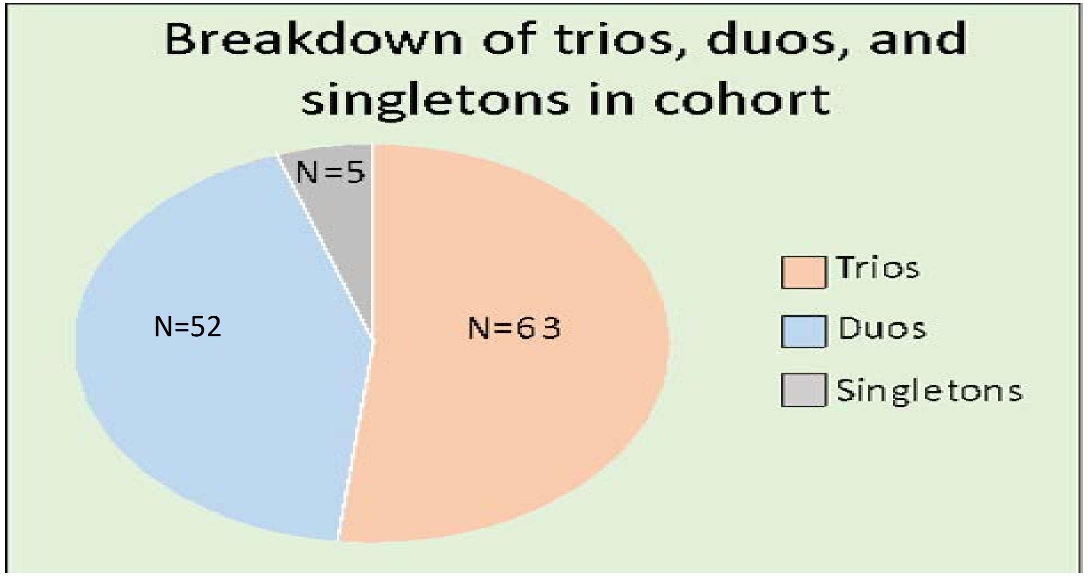
Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

BASIC^3

BCM Advancing Sequencing Into Childhood Cancer Care

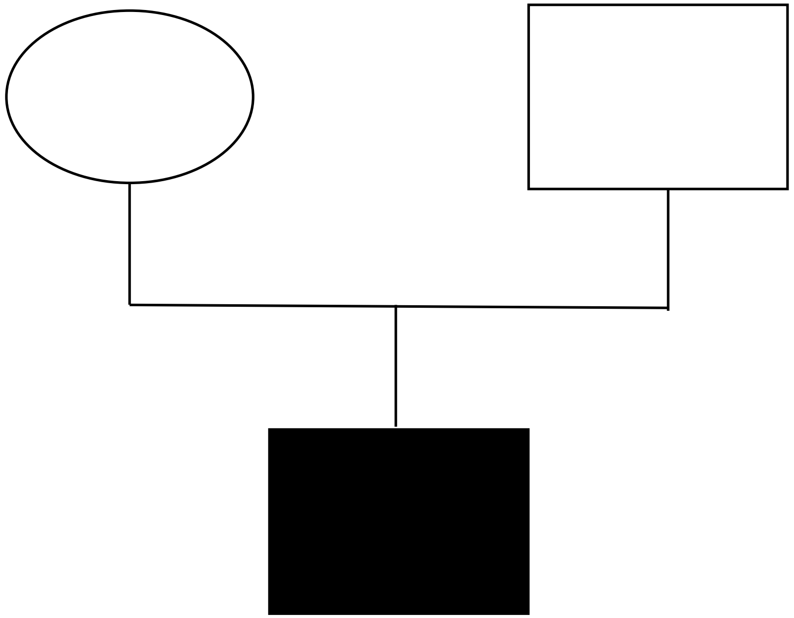
Goal 1: Identify *de novo* Single Nucleotide Variants (SNVs) and Structural Variants (SVs)

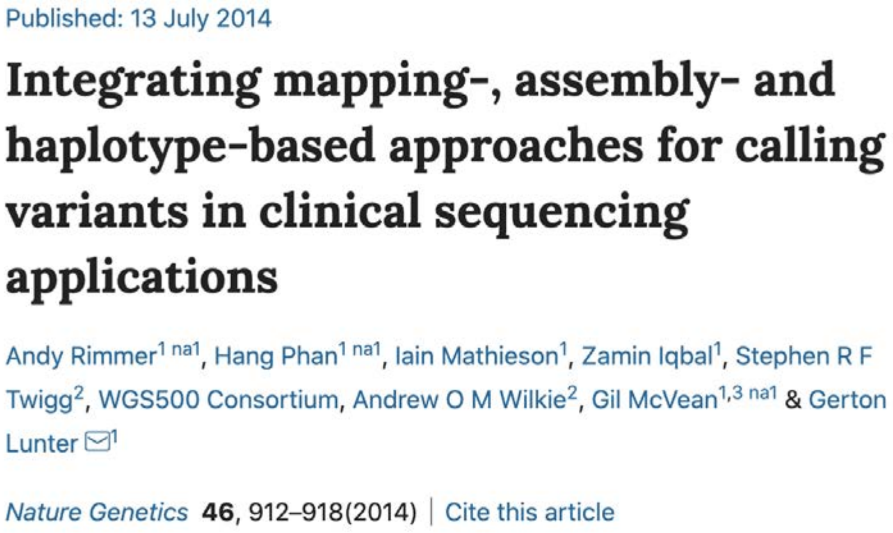
Goal 2: Identify putative pathogenic variants in known cancer genes that may have been missed by WES



CAVATICA

Use of Platypus for *de novo* SNV calling on Cavatica



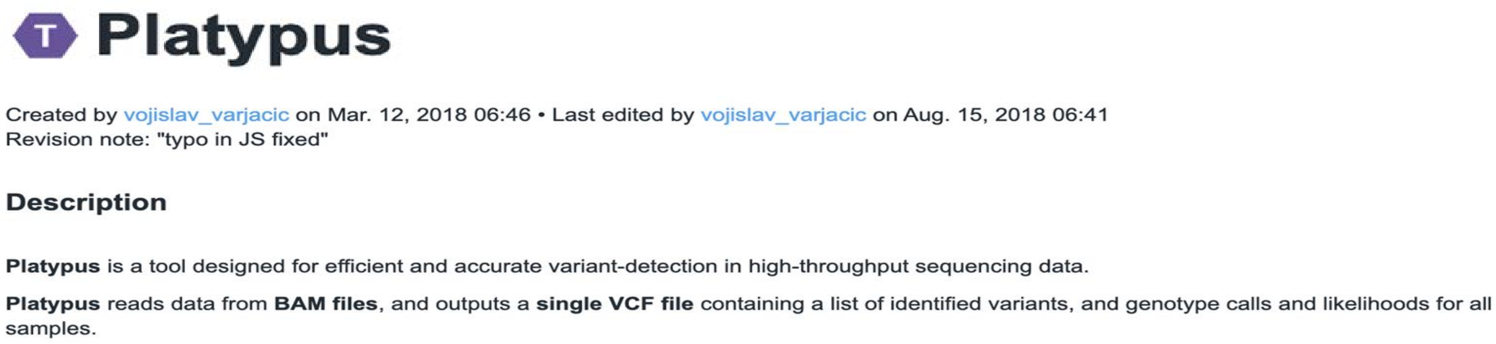


Published: 13 July 2014

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



Platypus

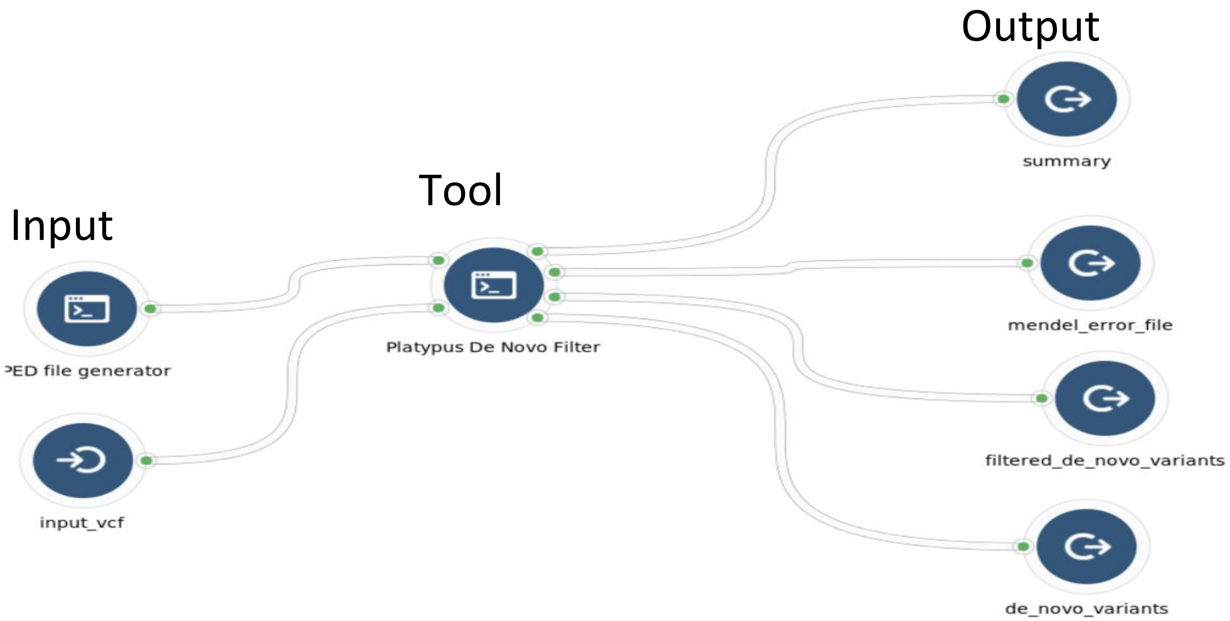
Created by vojislav_varjacic on Mar. 12, 2018 06:46 • Last edited by vojislav_varjacic on Aug. 15, 2018 06:41
Revision note: "typo in JS fixed"

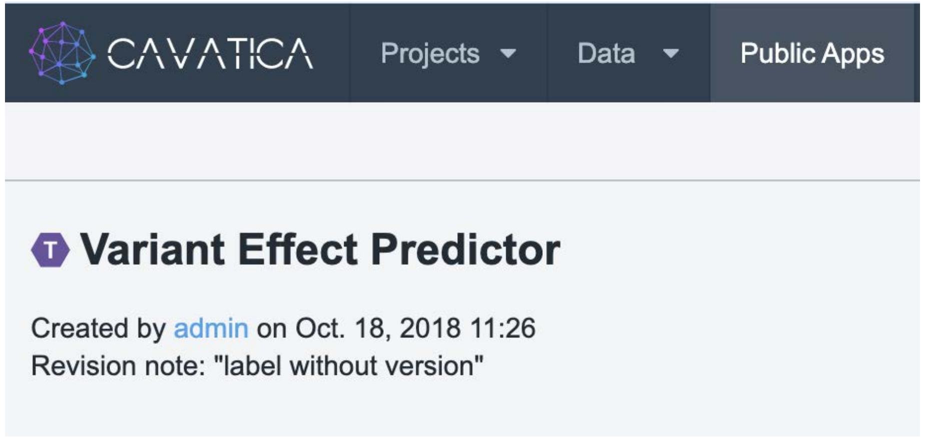
Description

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.

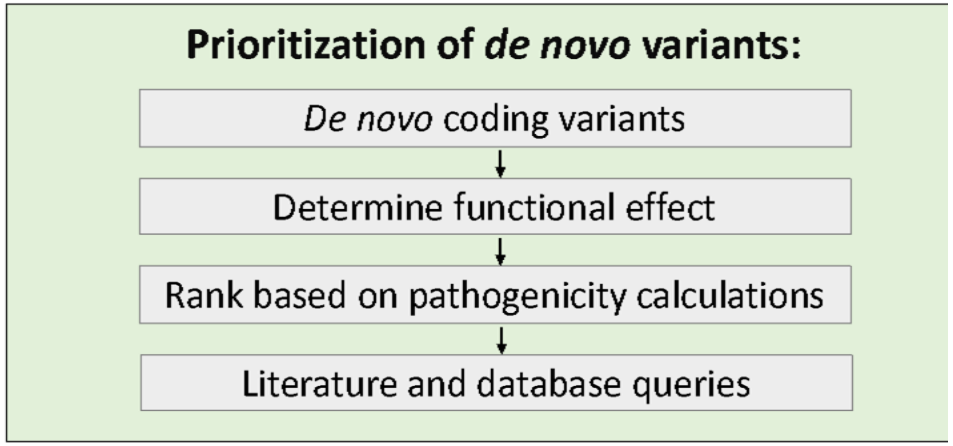
Analysis on Cavatica expedited *de novo* variant discovery





Variant Effect Predictor

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



Prioritization of *de novo* variants:

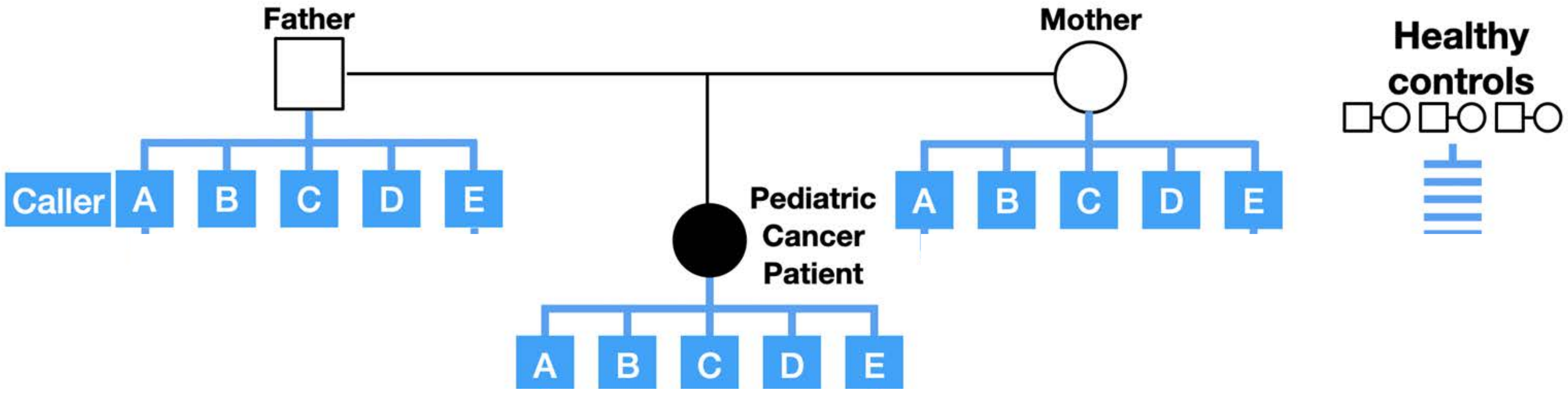
De novo coding variants
↓
Determine functional effect
↓
Rank based on pathogenicity calculations
↓
Literature and database queries

Outcome:

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

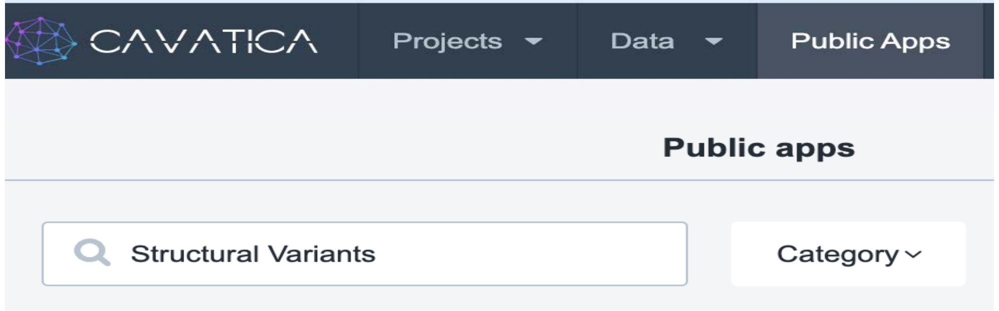
Variant Type	Frequency
Genome-wide *de novo*	60 to 190
Coding *de novo*	0 to 4

De novo SV analysis on Cavatica

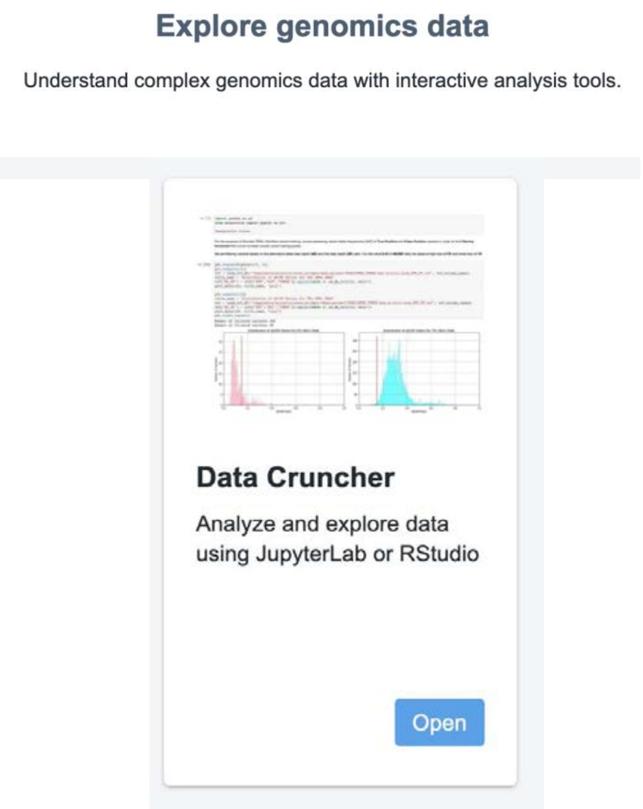


**Caller A, B, C, D, & E:
Lumpy, Manta, Delly,
Breakdancer, & CNVnator**

Analysis of SV on Cavatica requires multiple features of the platform



Public apps

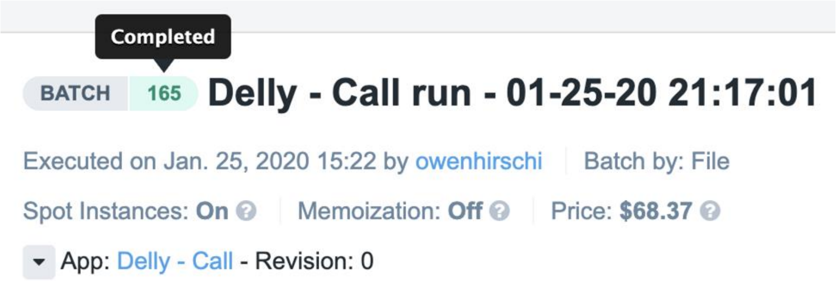


Explore genomics data

Understand complex genomics data with interactive analysis tools.

Data Cruncher

Analyze and explore data using JupyterLab or RStudio

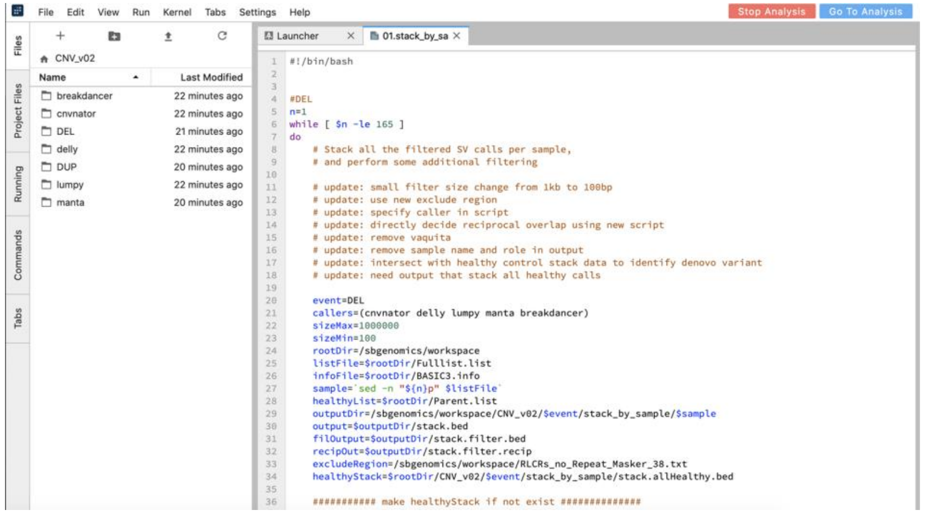


BATCH 165 Delly - Call run - 01-25-20 21:17:01

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

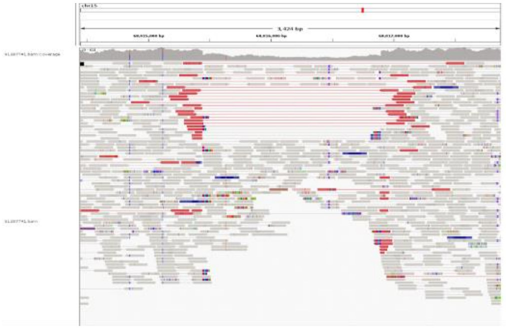


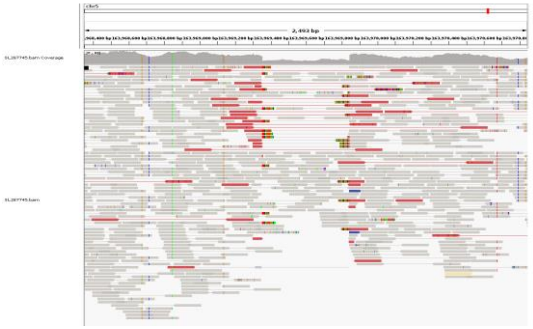
CNV_v02

Name	Last Modified
breakdancer	22 minutes ago
cnvnator	22 minutes ago
DEL	21 minutes ago
delly	22 minutes ago
DUP	20 minutes ago
lumpy	22 minutes ago
manta	20 minutes ago

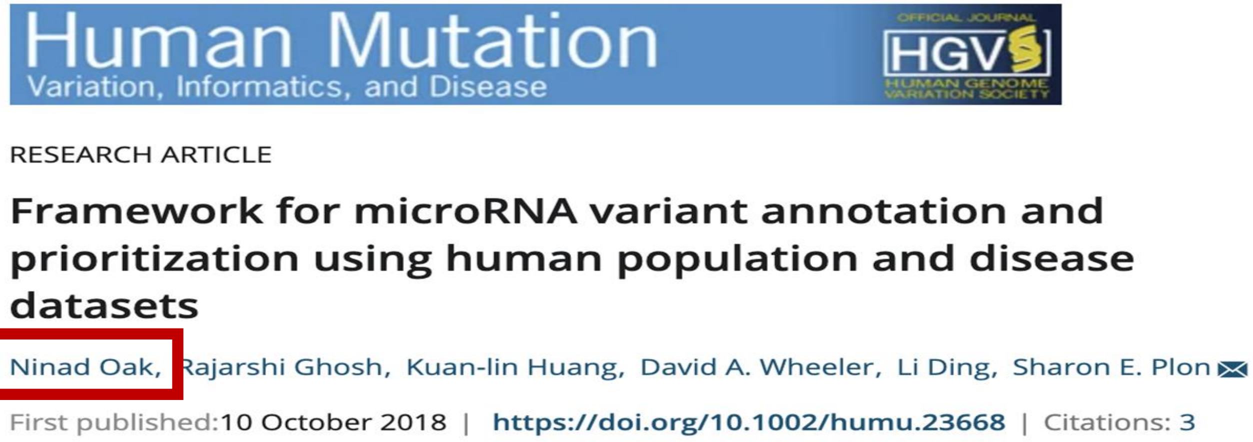
```bash
#!/bin/bash


#DEL
n=1
while [ $n -le 165 ]
do
    # Stack all the filtered SV calls per sample,
    # and perform some additional filtering

    # update: small filter size change from 1kb to 100bp
    # update: use new exclude region
    # update: specify caller in script
    # update: directly decide reciprocal overlap using new script
    # update: remove vaquita
    # update: remove sample name and role in output
    # update: intersect with healthy control stack data to identify denovo variant
    # update: need output that stack all healthy calls

    event=DEL
    callers=(cnvnator delly lumpy manta breakdancer)
    sizeMax=1000000
    sizeMin=100
    rootDir=/sbgenomics/workspace
    listFile=$rootDir/Fulllist.list
    infoFile=$rootDir/BASIC3.info
    sample=`sed -n "${n}p" $listFile`
    healthyList=$rootDir/Parent.list
    outputDir=/sbgenomics/workspace/CNV_v02/$event/stack_by_sample/$sample
    output=$outputDir/stack.bed
    filOutput=$outputDir/stack.filter.bed
    recipOut=$outputDir/stack.filter.recip
    excludeRegion=/sbgenomics/workspace/RLCRs_no_Repeat_Masker_38.txt
    healthyStack=$rootDir/CNV_v02/$event/stack_by_sample/stack.allHealthy.bed

    ########### make healthyStack if not exist ##############
```

IGV images





Analysis of miRNA variation on Cavatica



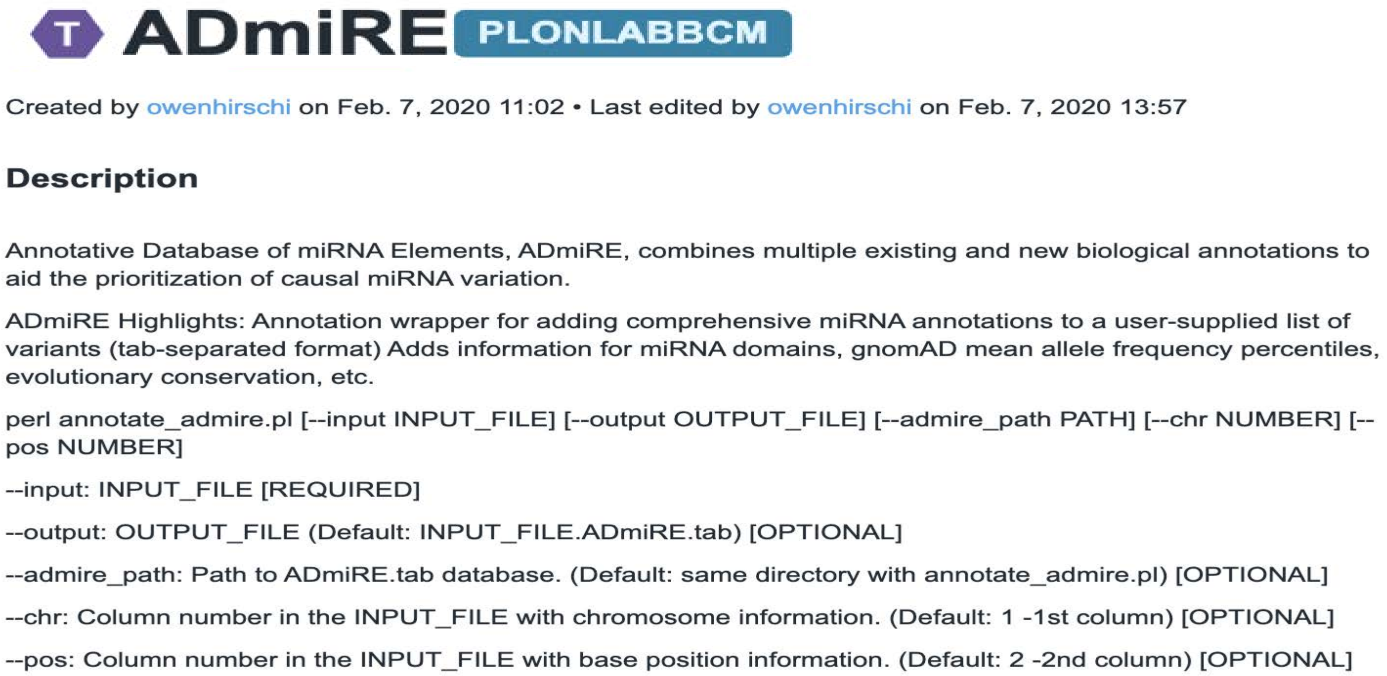
Human Mutation
Variation, Informatics, and Disease

RESEARCH ARTICLE

Framework for microRNA variant annotation and prioritization using human population and disease datasets

Ninad Oak, Rajarshi Ghosh, Kuan-lin Huang, David A. Wheeler, Li Ding, Sharon E. Plon

First published:10 October 2018 | https://doi.org/10.1002/humu.23668 | Citations: 3



T ADmiRE PLONLABBCM

Created by owenhirschi on Feb. 7, 2020 11:02 • Last edited by owenhirschi on Feb. 7, 2020 13:57

Description

Annotative Database of miRNA Elements, ADmiRE, combines multiple existing and new biological annotations to aid the prioritization of causal miRNA variation.

ADmiRE Highlights: Annotation wrapper for adding comprehensive miRNA annotations to a user-supplied list of variants (tab-separated format) Adds information for miRNA domains, gnomAD mean allele frequency percentiles, evolutionary conservation, etc.

perl annotate_admire.pl [--input INPUT_FILE] [--output OUTPUT_FILE] [--admire_path PATH] [--chr NUMBER] [--pos NUMBER]

--input: INPUT_FILE [REQUIRED]

--output: OUTPUT_FILE (Default: INPUT_FILE.ADmiRE.tab) [OPTIONAL]

--admire_path: Path to ADmiRE.tab database. (Default: same directory with annotate_admire.pl) [OPTIONAL]

--chr: Column number in the INPUT_FILE with chromosome information. (Default: 1 -1st column) [OPTIONAL]

--pos: Column number in the INPUT_FILE with base position information. (Default: 2 -2nd column) [OPTIONAL]

Acknowledgments

Texas Children's Cancer Center

HGSC
HUMAN GENOME SEQUENCING CENTER

Plon Lab members:

Sharon Plon, MD, PhD

Saumya Sisoudiya

Adam Weinstein

Deborah Ritter, PhD

Xi Luo, PhD

Ryan Zabriskie

Ninad Oak, PhD- former

Baylor HGSC:

Hurley Li, PhD

BASIC3 Co-PI:

William Parson, MD, PhD

Baylor College of Medicine

Funding:

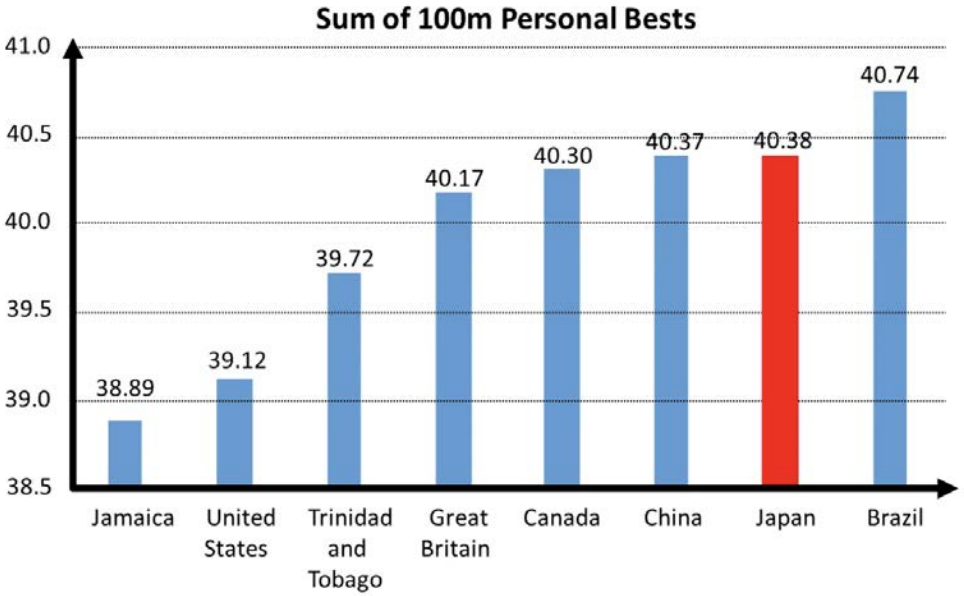
Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

CULLEN FOUNDATION

Kids First DRC Roadmap

Kids First DRC - The Model We Follow

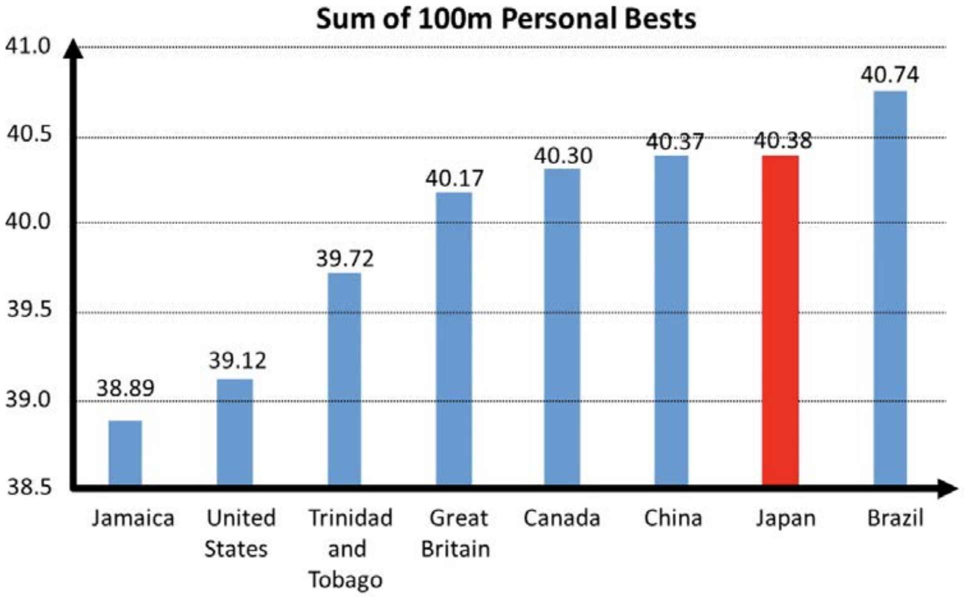
Japan win silver in the 4-x-100-meter relay at the Rio de Janeiro Games



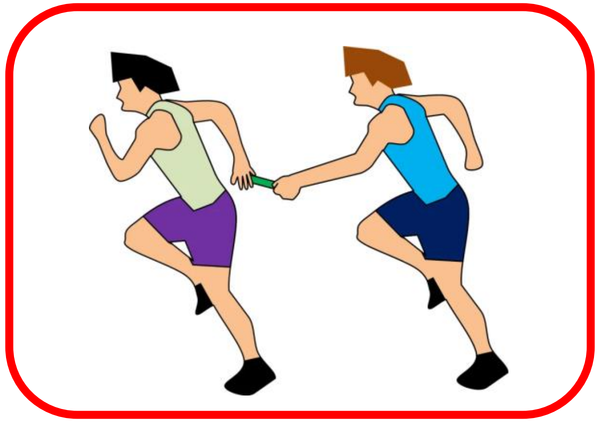
None of their team having ever run 100m in under 10 seconds

Kids First DRC - The Model We Follow

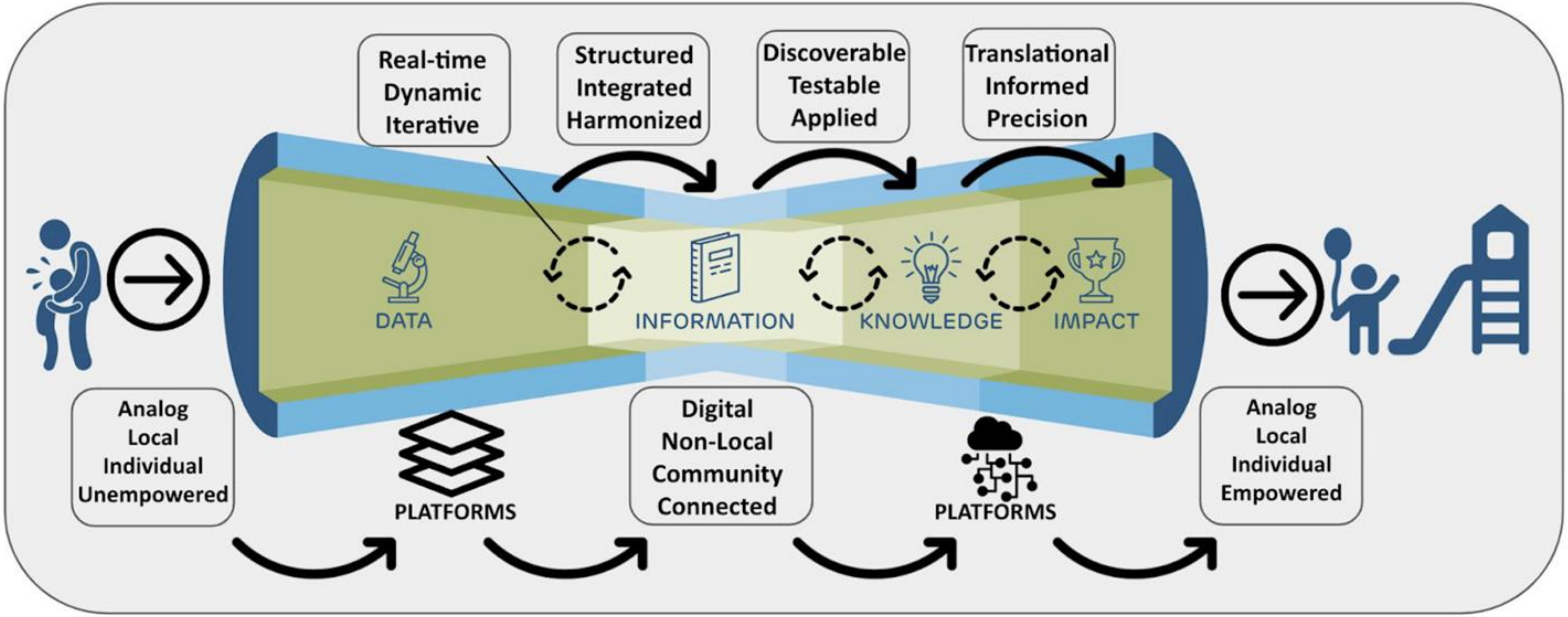
Japan win silver in the 4-x-100-meter relay at the Rio de Janeiro Games



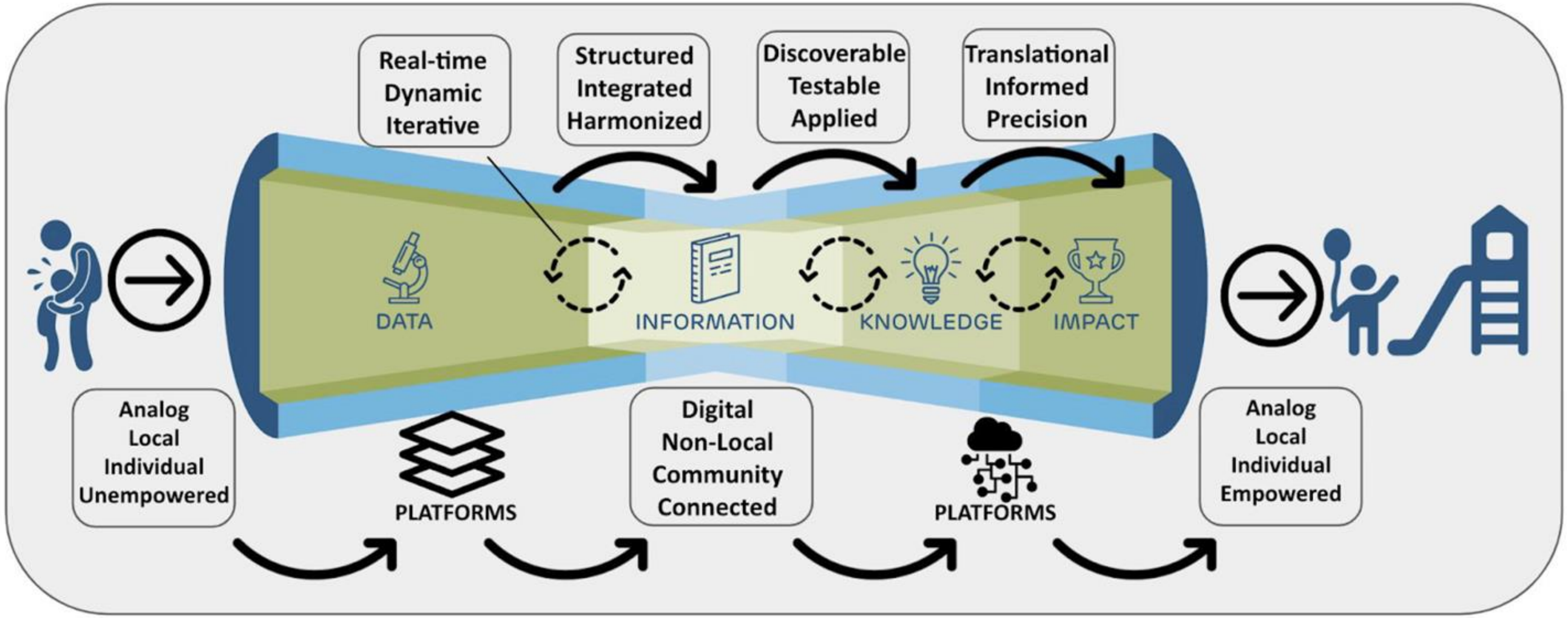
<u>None of their team having ever run 100m in under 10 seconds</u>



Kids First DRC - The Model We Follow



Addressing Scale - A Model



Real-time
Dynamic
Iterative

Structured
Integrated
Harmonized

Discoverable
Testable
Applied

Translational
Informed
Precision

DATA

INFORMATION

KNOWLEDGE

IMPACT

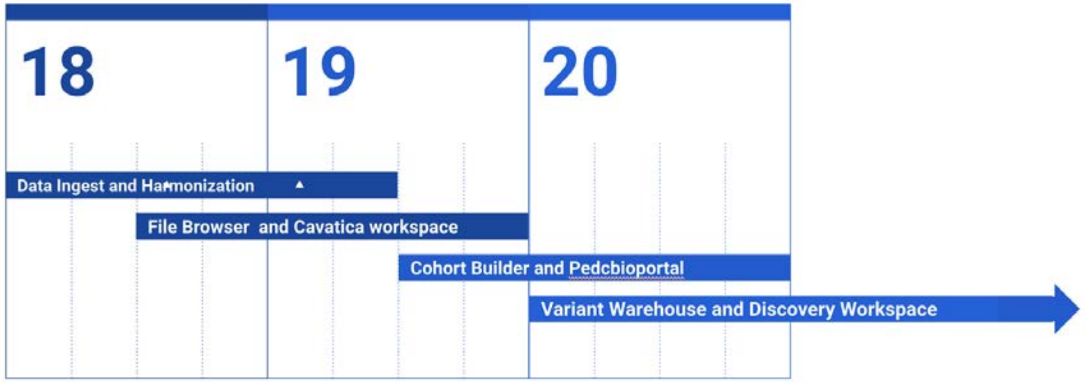
Analog
Local
Individual
Unempowered

PLATFORMS

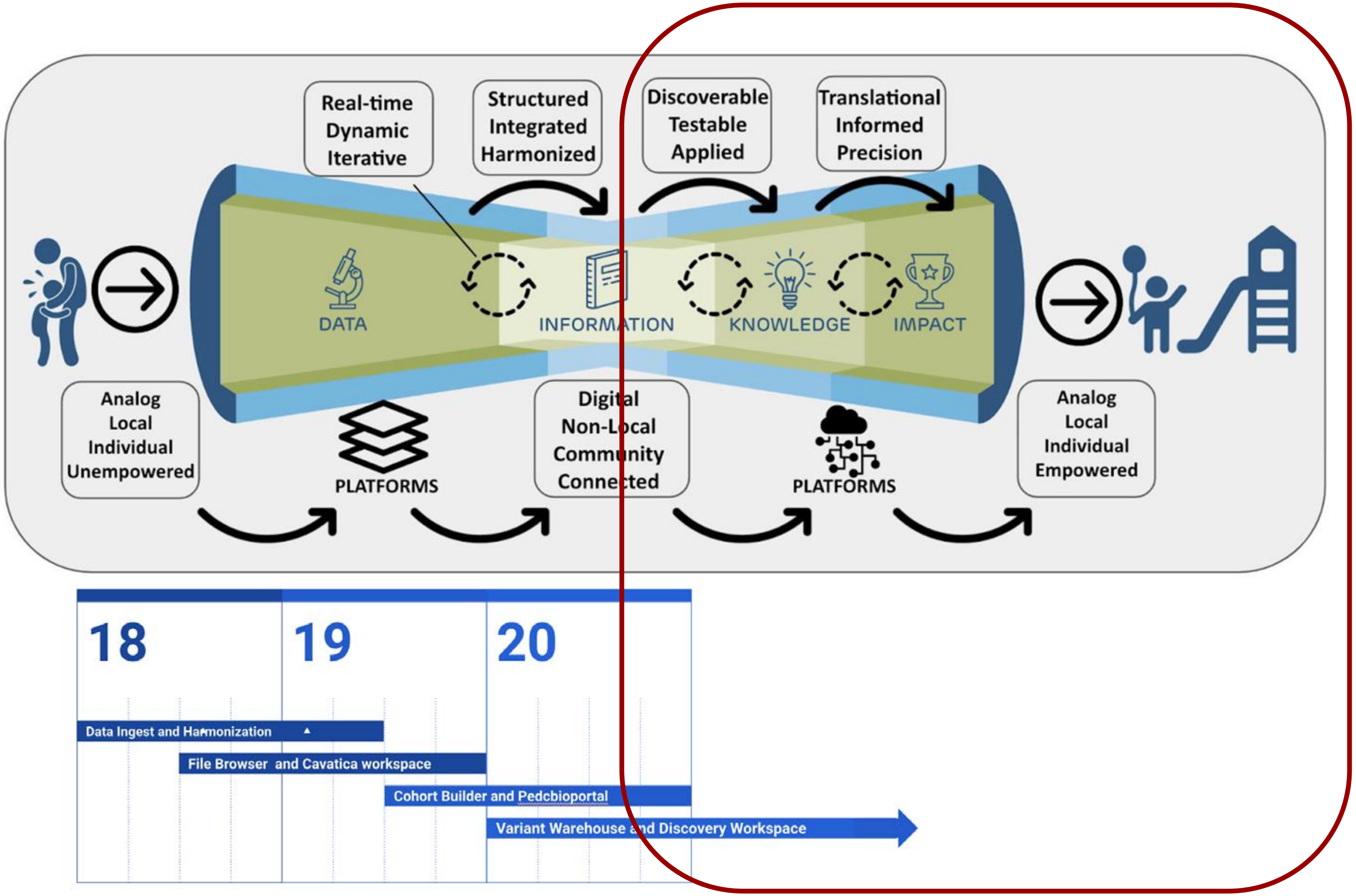
Digital
Non-Local
Community
Connected

PLATFORMS

Analog
Local
Individual
Empowered



Addressing Scale - A Model



Real-time
Dynamic
Iterative

Structured
Integrated
Harmonized

Discoverable
Testable
Applied

Translational
Informed
Precision

DATA

INFORMATION

KNOWLEDGE

IMPACT

Analog
Local
Individual
Unempowered

PLATFORMS

Digital
Non-Local
Community
Connected

PLATFORMS

Analog
Local
Individual
Empowered

18

19

20

Data Ingest and Harmonization

File Browser and Cavatica workspace

Cohort Builder and Pedcbioportal

Variant Warehouse and Discovery Workspace

Gabriella Miller Kids First Pediatric Research Program

Program & Collaboration Updates

NIH

The Common Fund

Gabriella Miller

Kids First

PEDIATRIC RESEARCH PROGRAM

NIH

The Common Fund

Gabriella Miller Kids First

PEDIATRIC RESEARCH PROGRAM

Danyelle Winchester, PhD

Health Specialist
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director, National Institutes of Health (NIH)

10 Released Datasets

➢ Disorders of Sex Development.	PI: Eric Vilain
➢ Congenital Diaphragmatic Hernia	PI: Wendy Chung
➢ Ewing Sarcoma	PI: Joshua Schiffman
➢ Orofacial Clefts: Caucasian families	PI: Mary Marazita
➢ Orofacial Clefts: Latin American families	PI: Mary Marazita
➢ Structural Heart & Other Defects	PI: Christine Seidman (PCGC)
➢ Cranial Dysinnervation Disorders	PI: Elizabeth Engle
➢ Adolescent Idiopathic Scoliosis	PI: Jonathan Rios
➢ Neuroblastoma	PI: John Maris
➢ Enchondromatoses	PI: Nara Sobreira

- **Kids First DRC website:** https://kidsfirstdrc.org/support/studies-and-access/
- **NIH Kids First Umbrella BioProject:** https://www.ncbi.nlm.nih.gov/bioproject/338775 > dbGaP links
- **X01 Abstracts**: https://commonfund.nih.gov/kidsfirst/x01projects

The value of Kids First datasets will be amplified when researchers use and analyze these data to make discoveries that will ultimately improve prevention, diagnostics, and therapeutic interventions for these conditions

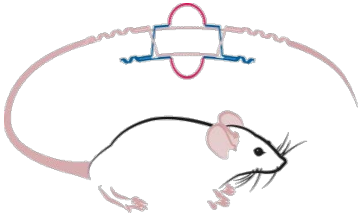
Researchers are using Kids First data to answer new scientific questions

- ***13 awards*** *for R03 for analyses of Kids First data* (PAR-16-348 ; PAR-18-733; PAR-19-069, PAR-19-375)

- ***1 award for NIDCR R03*** (PAR-16-070)

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

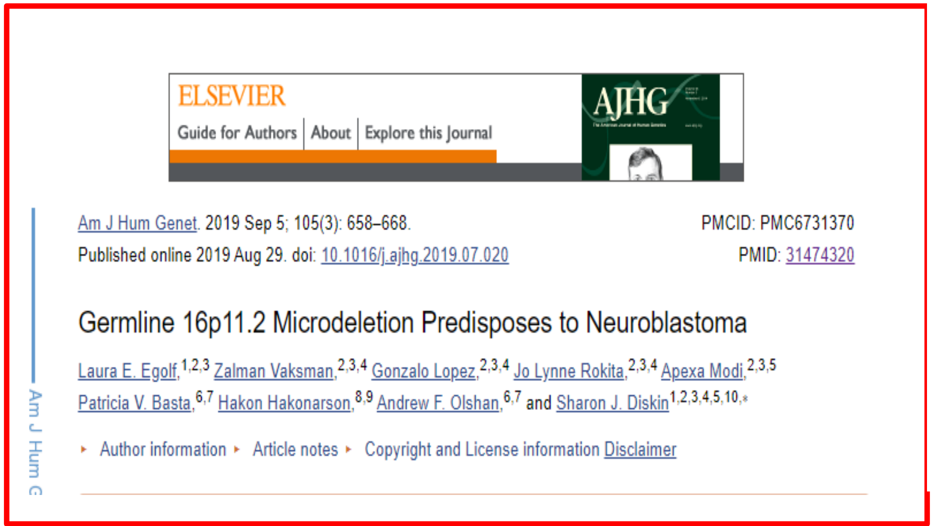
- Spurred **new collaborations** with KOMP2 & INCLUDE



Knockout Mouse Phenotyping Project (KOMP2)

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

Kids First Publications


ELSEVIER

Guide for Authors | About | Explore this Journal

AJHG

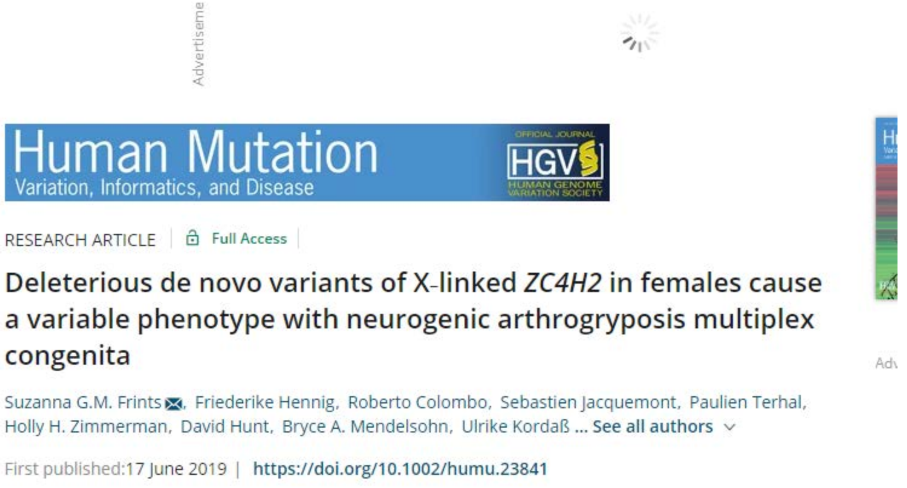
Am J Hum Genet. 2019 Sep 5; 105(3): 658–668. PMCID: PMC6731370

Published online 2019 Aug 29. doi: 10.1016/j.ajhg.2019.07.020 PMID: 31474320

Germline 16p11.2 Microdeletion Predisposes to Neuroblastoma

Laura E. Egolf,[1,2,3] Zalman Vaksman,[2,3,4] Gonzalo Lopez,[2,3,4] Jo Lynne Rokita,[2,3,4] Apexa Modi,[2,3,5] Patricia V. Basta,[6,7] Hakon Hakonarson,[8,9] Andrew F. Olshan,[6,7] and Sharon J. Diskin[1,2,3,4,5,10,*]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer


Human Mutation

Variation, Informatics, and Disease

RESEARCH ARTICLE | Full Access

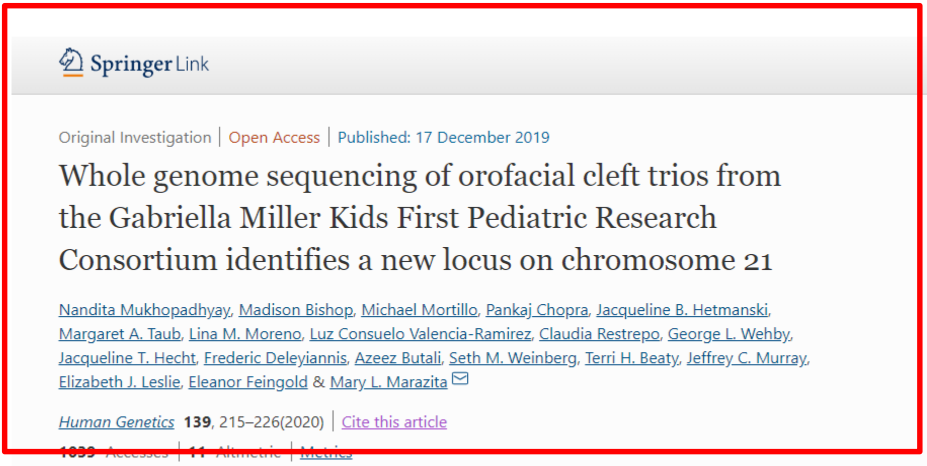
Deleterious de novo variants of X‐linked *ZC4H2* in females cause a variable phenotype with neurogenic arthrogryposis multiplex congenita

Suzanna G.M. Frints, Friederike Hennig, Roberto Colombo, Sebastien Jacquemont, Paulien Terhal, Holly H. Zimmerman, David Hunt, Bryce A. Mendelsohn, Ulrike Kordaß ... See all authors ⌄

First published:17 June 2019 | https://doi.org/10.1002/humu.23841

Total number of publications from Kids First ~13

Average RCR: 1.32


Springer Link

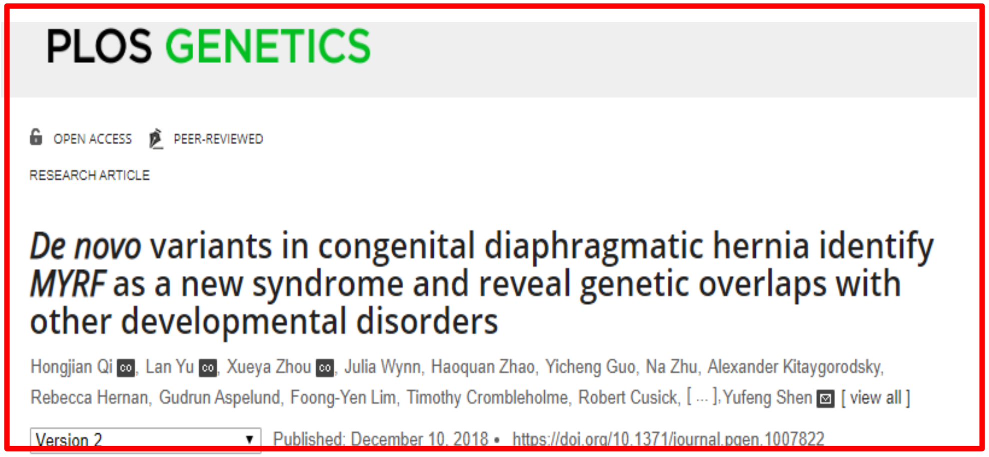
Original Investigation | Open Access | Published: 17 December 2019

Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21

Nandita Mukhopadhyay, Madison Bishop, Michael Mortillo, Pankaj Chopra, Jacqueline B. Hetmanski, Margaret A. Taub, Lina M. Moreno, Luz Consuelo Valencia-Ramirez, Claudia Restrepo, George L. Wehby, Jacqueline T. Hecht, Frederic Deleyiannis, Azeez Butali, Seth M. Weinberg, Terri H. Beaty, Jeffrey C. Murray, Elizabeth J. Leslie, Eleanor Feingold & Mary L. Marazita

Human Genetics **139**, 215–226(2020) | Cite this article

1039 Accesses | 11 Altmetric | Metrics


PLOS GENETICS

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

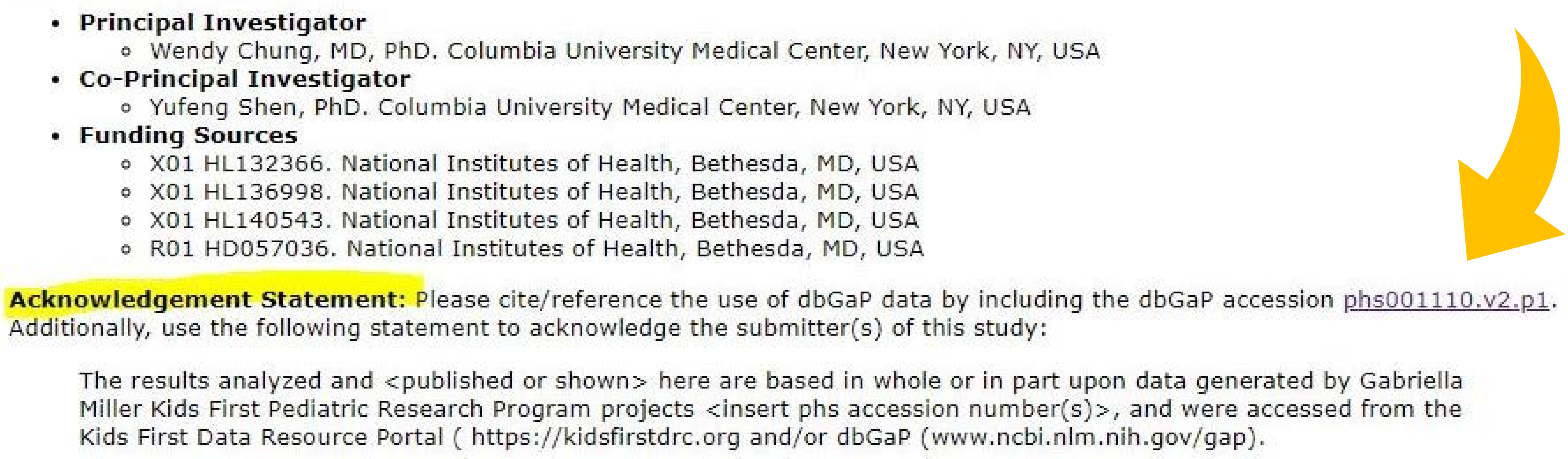
De novo variants in congenital diaphragmatic hernia identify *MYRF* as a new syndrome and reveal genetic overlaps with other developmental disorders

Hongjian Qi, Lan Yu, Xueya Zhou, Julia Wynn, Haoquan Zhao, Yicheng Guo, Na Zhu, Alexander Kitaygorodsky, Rebecca Hernan, Gudrun Aspelund, Foong-Yen Lim, Timothy Crombleholme, Robert Cusick, [...], Yufeng Shen [view all]

Version 2 ▾ Published: December 10, 2018 • https://doi.org/10.1371/journal.pgen.1007822

Kids First Publications: How to Acknowledge Kids First Data

- Secondary users (end users) must acknowledge the dataset(s) they use by listing dbGaP accession numbers and the databases from which the data were accessed (e.g. link to the Kids First Data Resource Center or Portal). The acknowledgement statement can be found at the bottom of the dbGaP study page and in the Data Use Certification.
  - **See Frequently Asked Questions for X01 Cohorts Selected for Sequencing #5**: https://commonfund.nih.gov/kidsfirst/FAQ#X01%20selected



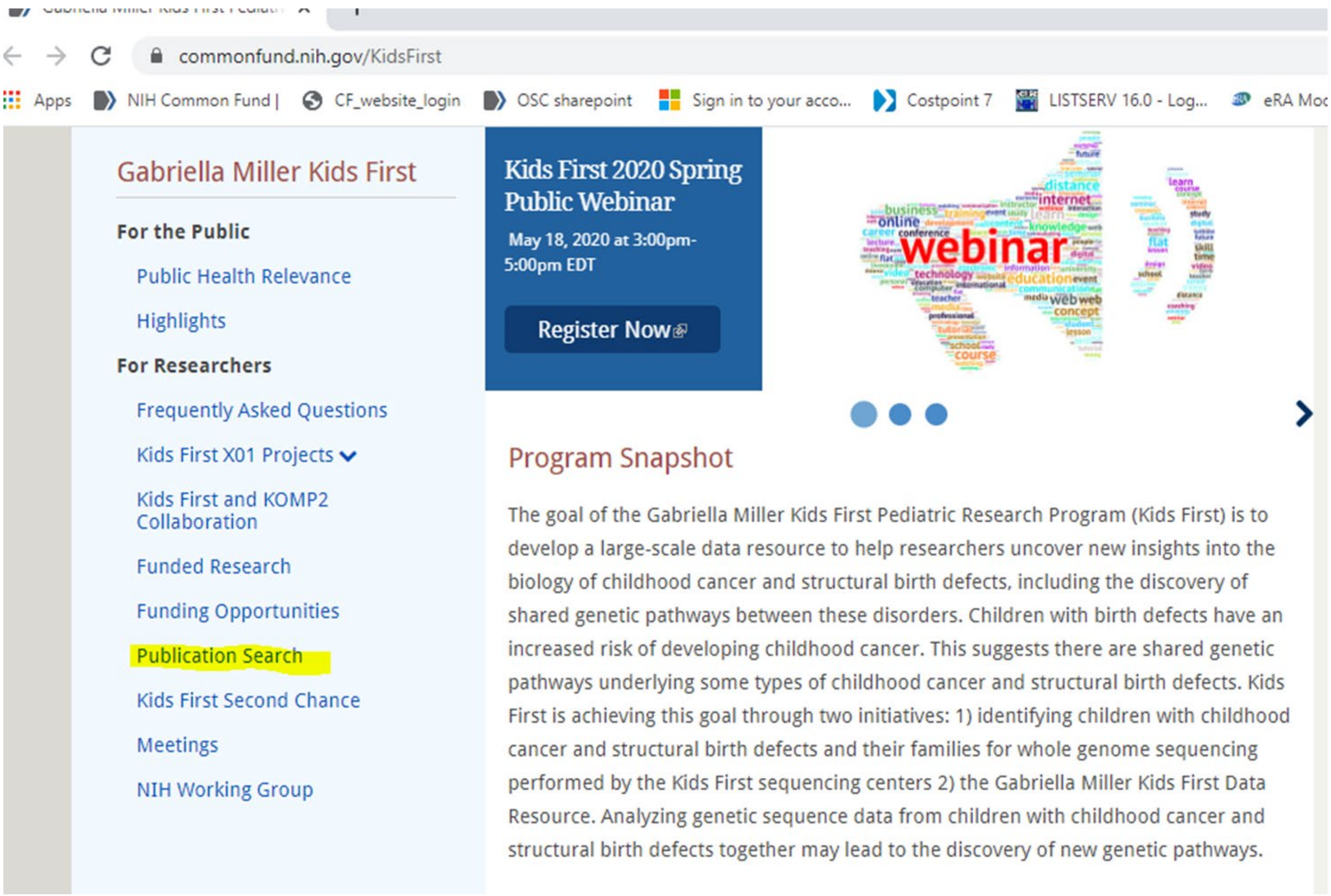
- **Principal Investigator**
  - Wendy Chung, MD, PhD. Columbia University Medical Center, New York, NY, USA
- **Co-Principal Investigator**
  - Yufeng Shen, PhD. Columbia University Medical Center, New York, NY, USA
- **Funding Sources**
  - X01 HL132366. National Institutes of Health, Bethesda, MD, USA
  - X01 HL136998. National Institutes of Health, Bethesda, MD, USA
  - X01 HL140543. National Institutes of Health, Bethesda, MD, USA
  - R01 HD057036. National Institutes of Health, Bethesda, MD, USA

Acknowledgement Statement: Please cite/reference the use of dbGaP data by including the dbGaP accession phs001110.v2.p1. Additionally, use the following statement to acknowledge the submitter(s) of this study:

> The results analyzed and <published or shown> here are based in whole or in part upon data generated by Gabriella Miller Kids First Pediatric Research Program projects <insert phs accession number(s)>, and were accessed from the Kids First Data Resource Portal (https://kidsfirstdrc.org and/or dbGaP (www.ncbi.nlm.nih.gov/gap).

Kids First Publications Search Page

https://commonfund.nih.gov/KidsFirst



Gabriella Miller Kids First

For the Public

- Public Health Relevance
- Highlights

For Researchers

- Frequently Asked Questions
- Kids First X01 Projects
- Kids First and KOMP2 Collaboration
- Funded Research
- Funding Opportunities
- Publication Search
- Kids First Second Chance
- Meetings
- NIH Working Group

Kids First 2020 Spring Public Webinar

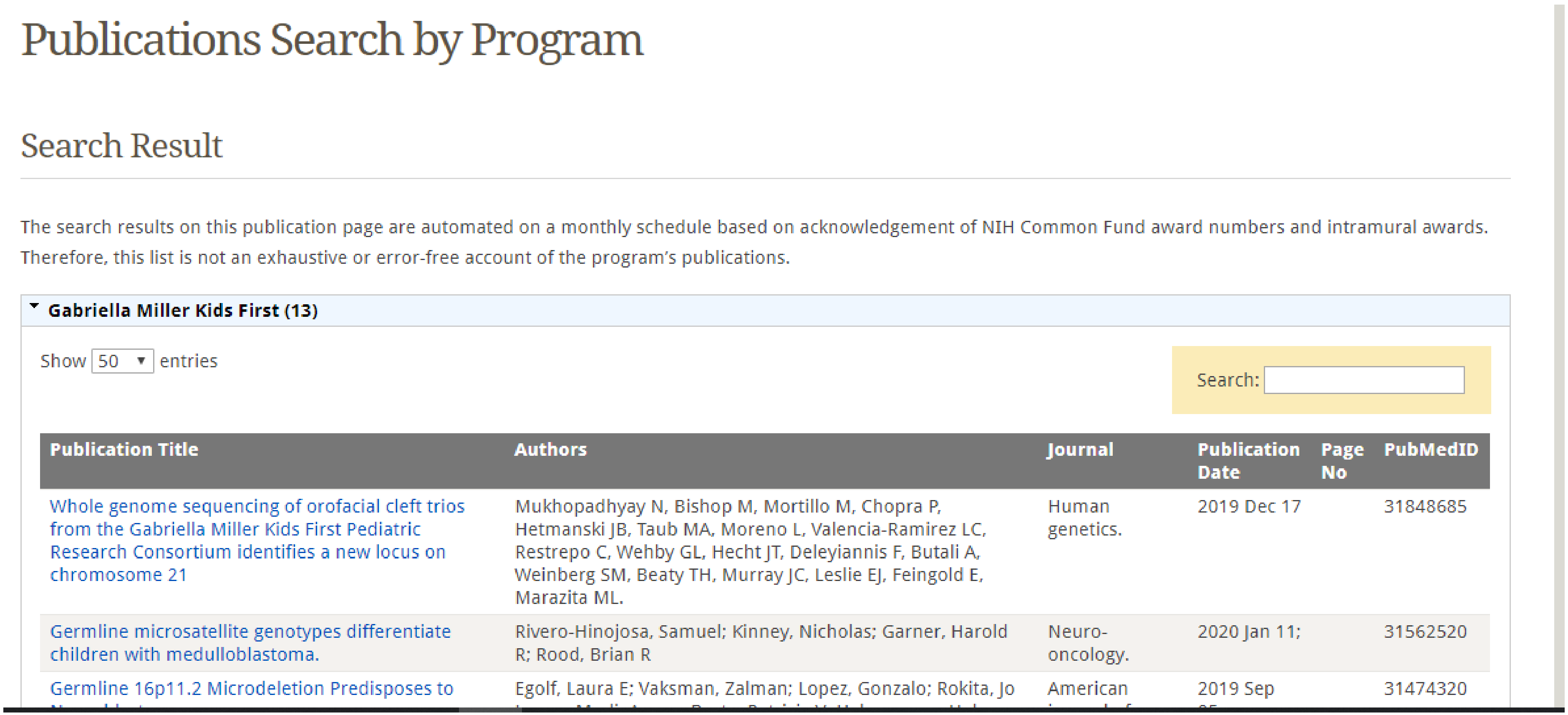
May 18, 2020 at 3:00pm-5:00pm EDT

Program Snapshot

The goal of the Gabriella Miller Kids First Pediatric Research Program (Kids First) is to develop a large-scale data resource to help researchers uncover new insights into the biology of childhood cancer and structural birth defects, including the discovery of shared genetic pathways between these disorders. Children with birth defects have an increased risk of developing childhood cancer. This suggests there are shared genetic pathways underlying some types of childhood cancer and structural birth defects. Kids First is achieving this goal through two initiatives: 1) identifying children with childhood cancer and structural birth defects and their families for whole genome sequencing performed by the Kids First sequencing centers 2) the Gabriella Miller Kids First Data Resource. Analyzing genetic sequence data from children with childhood cancer and structural birth defects together may lead to the discovery of new genetic pathways.

Kids First Publications Search Page

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



Publications Search by Program

Search Result

The search results on this publication page are automated on a monthly schedule based on acknowledgement of NIH Common Fund award numbers and intramural awards. Therefore, this list is not an exhaustive or error-free account of the program's publications.

▼ **Gabriella Miller Kids First (13)**

Show 50 ▼ entries

Search:

Publication Title	Authors	Journal	Publication Date	Page No	PubMedID
Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21	Mukhopadhyay N, Bishop M, Mortillo M, Chopra P, Hetmanski JB, Taub MA, Moreno L, Valencia-Ramirez LC, Restrepo C, Wehby GL, Hecht JT, Deleyiannis F, Butali A, Weinberg SM, Beaty TH, Murray JC, Leslie EJ, Feingold E, Marazita ML.	Human genetics.	2019 Dec 17		31848685
Germline microsatellite genotypes differentiate children with medulloblastoma.	Rivero-Hinojosa, Samuel; Kinney, Nicholas; Garner, Harold R; Rood, Brian R	Neuro-oncology.	2020 Jan 11;		31562520
Germline 16p11.2 Microdeletion Predisposes to	Egolf, Laura E; Vaksman, Zalman; Lopez, Gonzalo; Rokita, Jo	American	2019 Sep		31474320

Kids First Investigators: Past Presentations

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

Strategic Planning

Progress on Addressing Key Challenges

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

7 Consensus Recommendation Themes

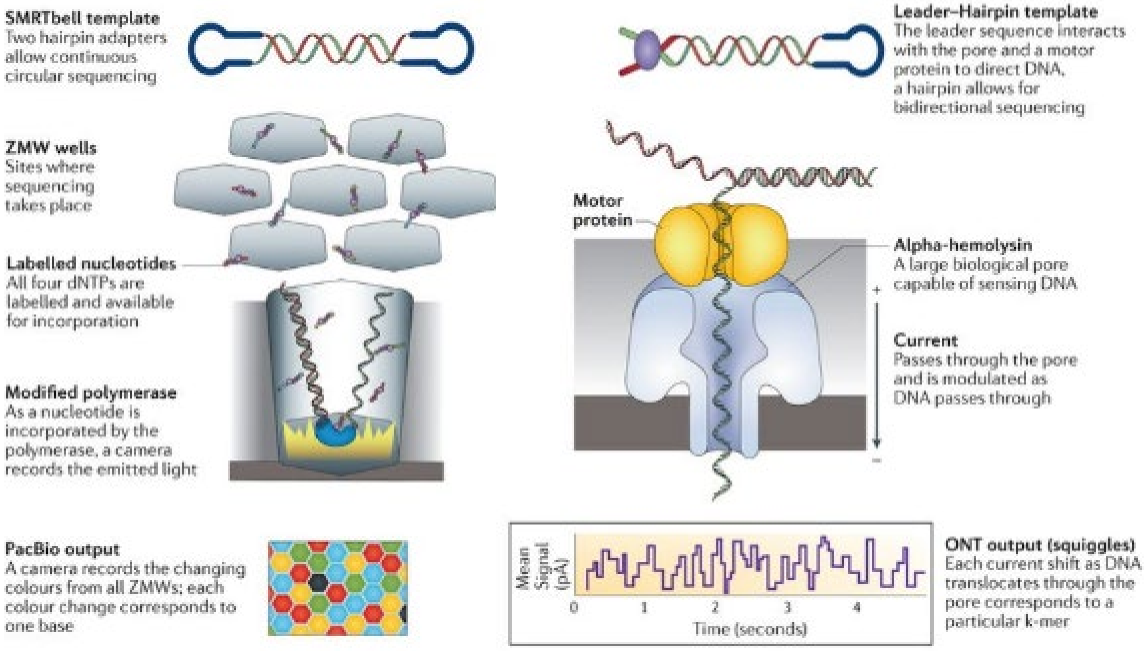
1. **Innovation: Resource, infrastructure, or tool development.**
   Activities: Data Visualization tools; other tools for clinical/phenotypic data
2. **Clinical/phenotypic data extraction, harmonization, & curation.**
   Activities: Collect, extract, organize, curate, harmonize, and submit deep clinical 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

2020 Kids First X01
Long Read Sequencing Pilot

BROAD INSTITUTE

HUDSONALPHA INSTITUTE FOR BIOTECHNOLOGY

St. Jude Children's Research Hospital


SMRTbell template
Two hairpin adapters allow continuous circular sequencing
ZMW wells
Sites where sequencing takes place
Labelled nucleotides
All four dNTPs are labelled and available for incorporation
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base
Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing
Motor protein
Alpha-hemolysin
A large biological pore capable of sensing DNA
Current
Passes through the pore and is modulated as DNA passes through
+
−
Mean Signal (pA)
0
1
2
3
4
Time (seconds)
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

PACIFIC BIOSCIENCES

OXFORD NANOPORE

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

7 Consensus Recommendation Themes Emerged

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

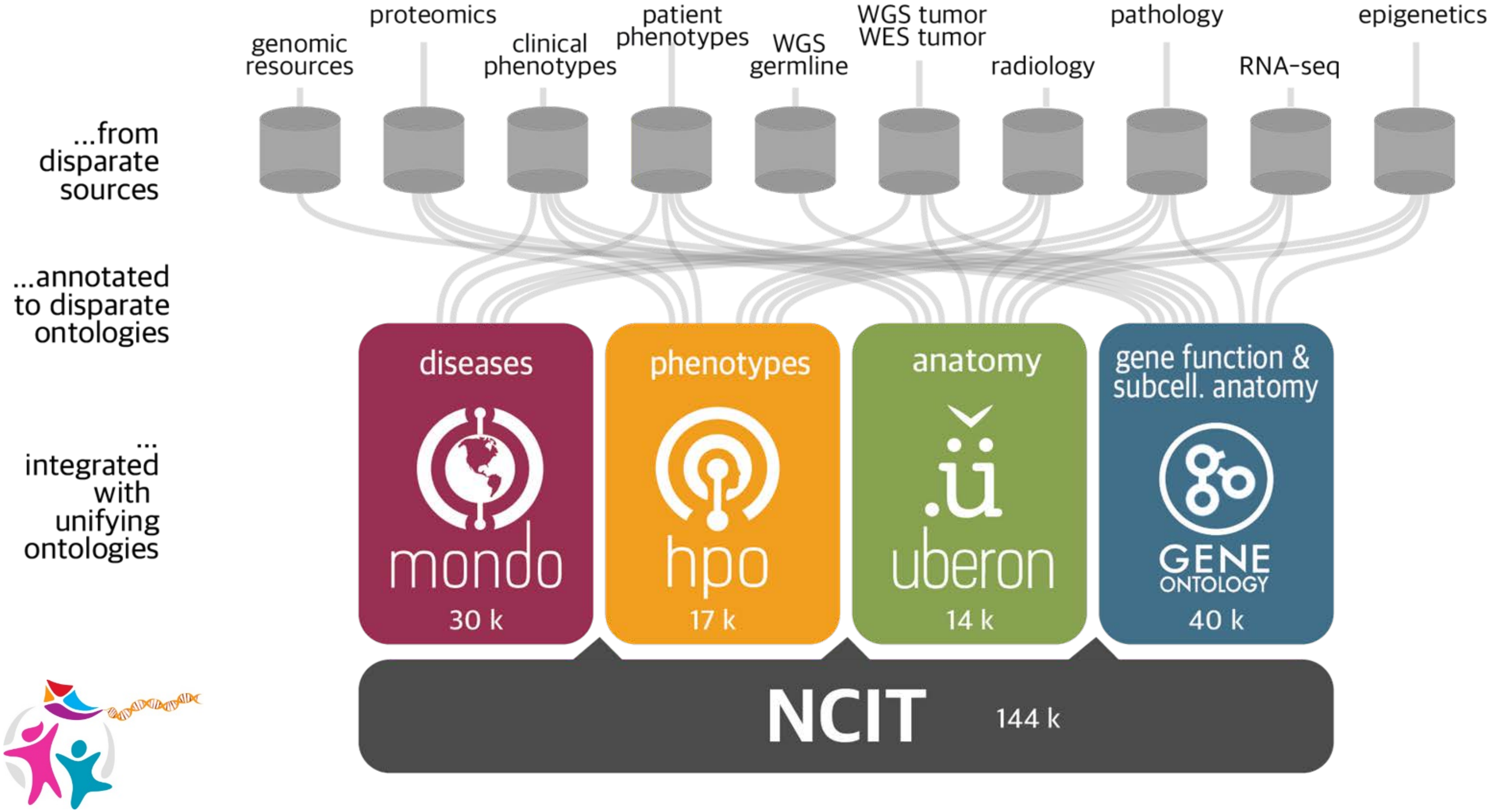
6. **Validation with model organisms.**
   Activities: validating KF findings/variants, deep phenotyping of animal models

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

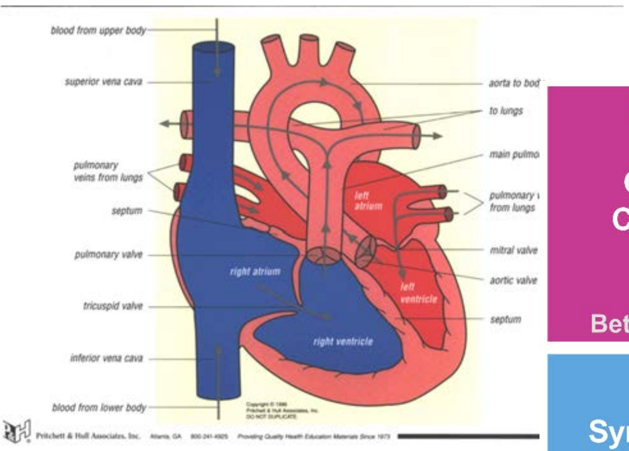
Innovation across the Phenotypic Translational Divide Webinar

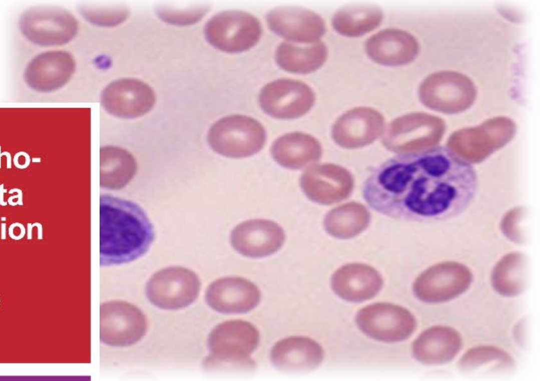
Information: https://monarch-initiative.github.io/phenomics/pages/clin-phen-webinar.html

Curation with ontologies that support heterogenous data types in Kids First



Innovation across the Phenotypic Translational Divide Webinar





Pediatric Cardiac Genomics Consortium (PCGC)
-
Betsy Goldmuntz

Down Syndrome -Congenital Heart Disease
-
Joaquin Espinosa

Neuropsychological Data Harmonization
-
Stephanie Sherman

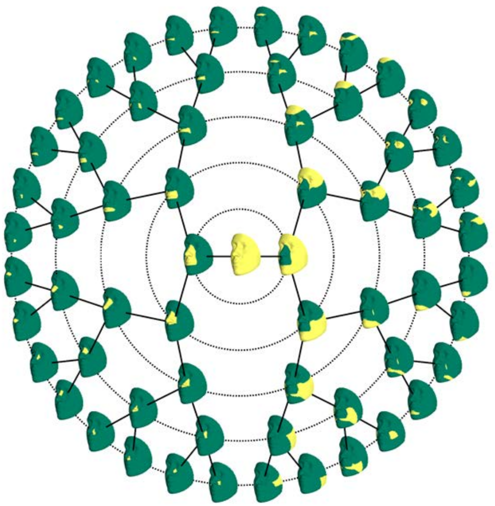
Down Syndrome-ALL & Rhabdomyosarcoma
-
Phillip Lupo

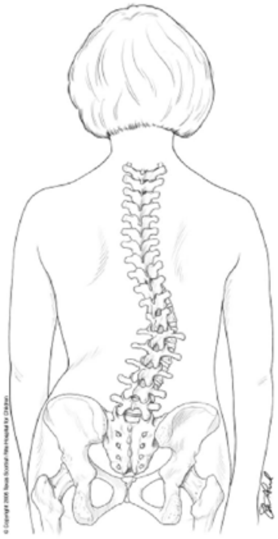
Enchondromatoses and Related Malignant Tumors
-
Nara Sobreira

Orofacial Clefts
-
Mary Marazita

Adolescent Idiopathic Scoliosis
-
Carole Wise

Cornelia de Lange Syndrome
-
Sarah Raible

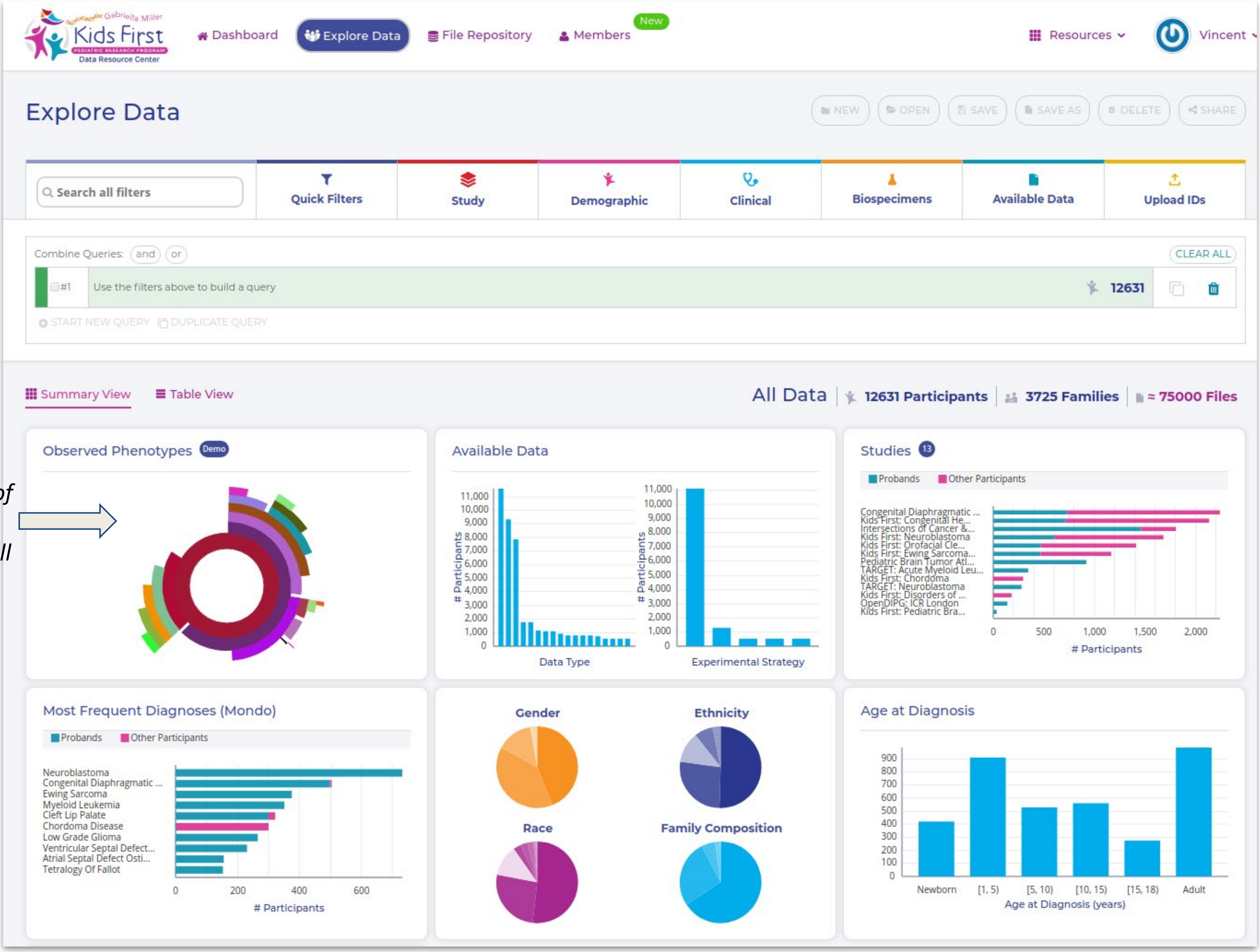




Webinar Information: https://monarch-initiative.github.io/phenomics/pages/clin-phen-webinar.html

New ontology search and visualisation tools

Sunburst: hierarchical view of participant counts per HPO term at all levels



7 Consensus Recommendation Themes Emerged

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*

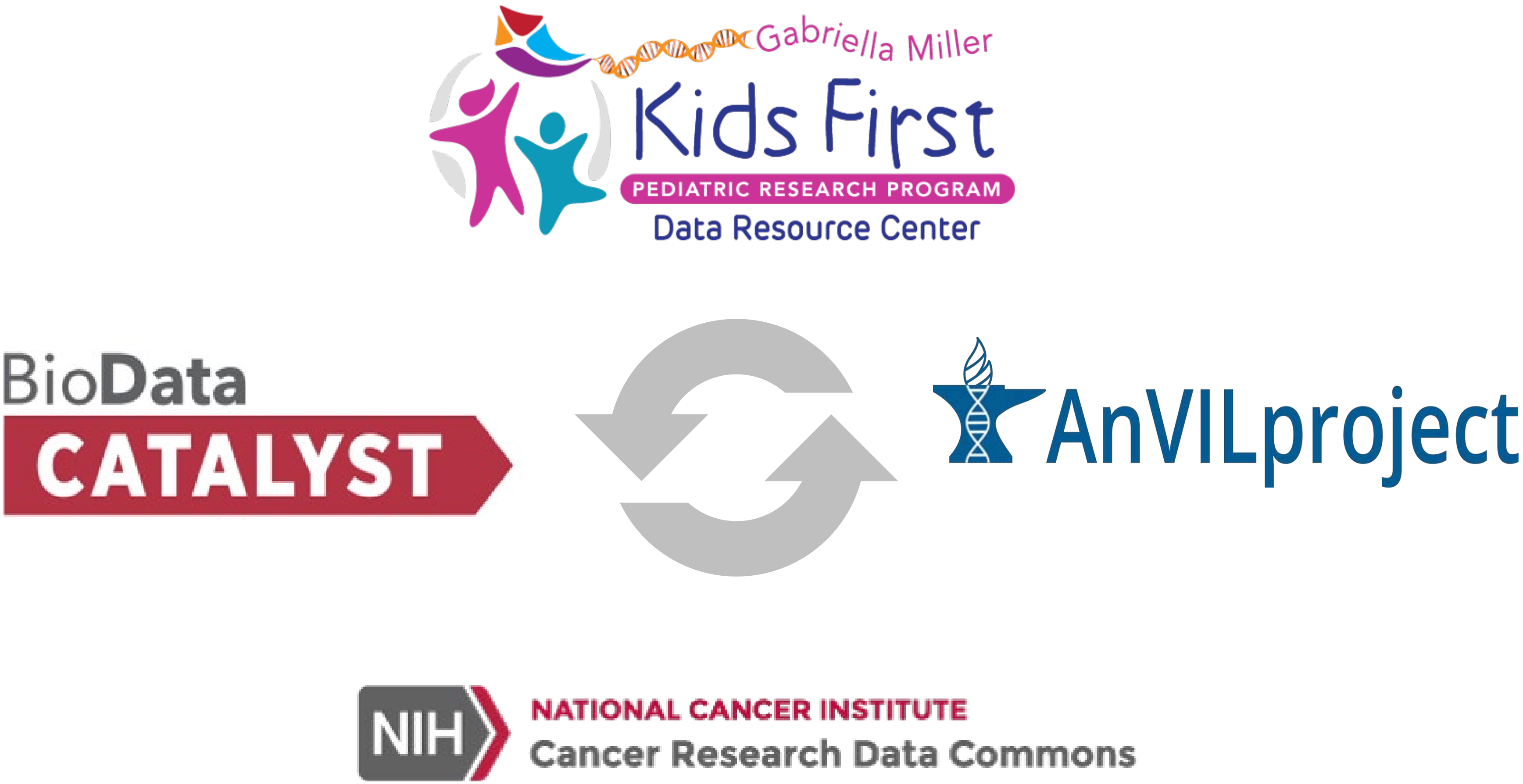
4. **Integration and interoperability of external pediatric datasets.**
   Activities: Using DRC workflow and best practices to harmonize external pediatric datasets; Building tools that can operate across multiple spaces

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 Cloud Based Platforms Interoperability (NCPI)


Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center
BioData
CATALYST
AnVILproject
NIH
NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

NIH Cloud Based Platforms Interoperability efforts

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

BioData CATALYST

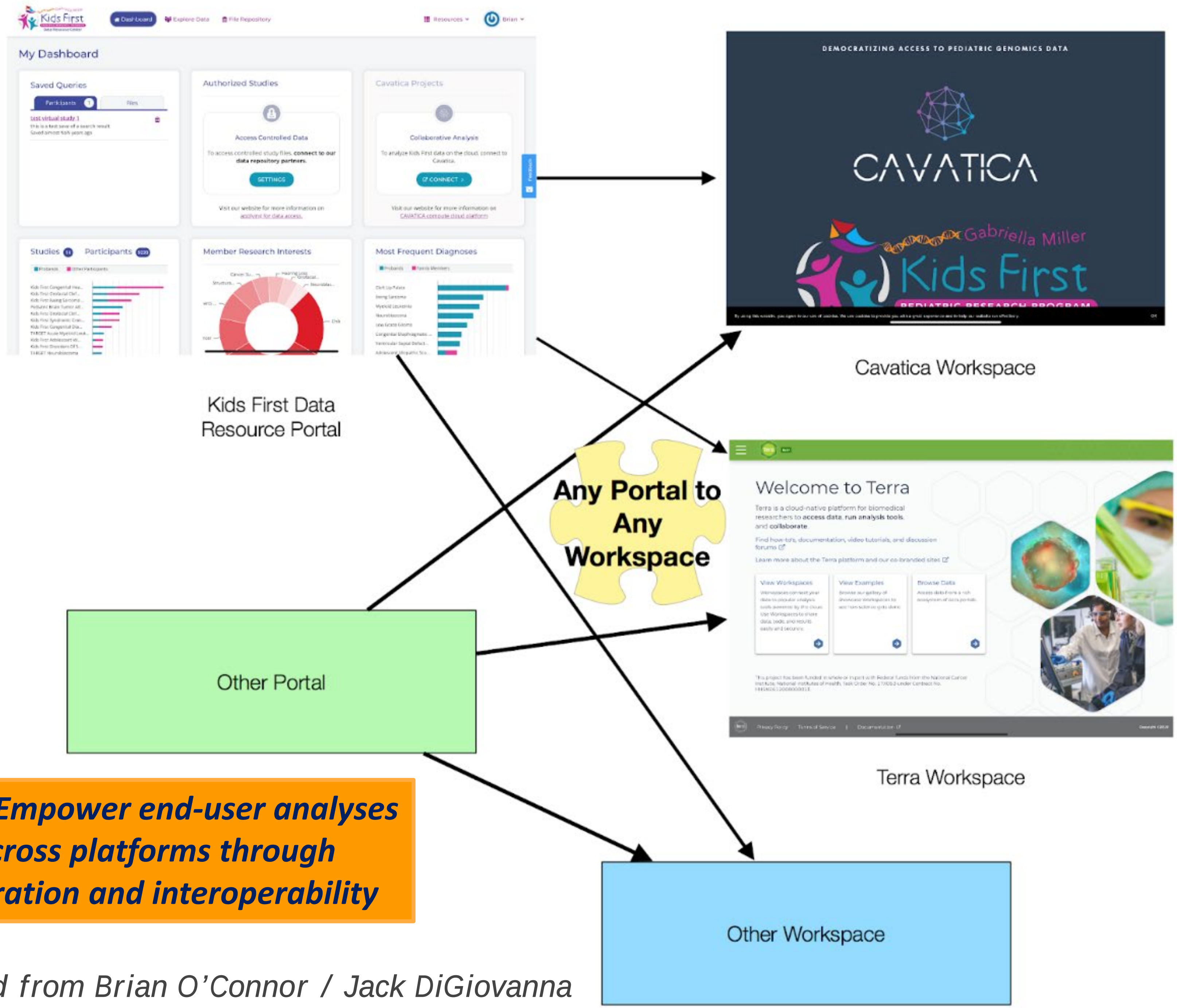
TOPMed/PCGC

AnVILproject

Undiagnosed Diseases Network, Centers for Mendelian Genomics, GTEx

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

TARGET



My Dashboard

Saved Queries

Authorized Studies

Access Controlled Data

To access controlled study files, connect to our data repository partners.

SETTINGS

Cavatica Projects

Collaborative Analysis

To analyze Kids First data on the cloud, connect to Cavatica.

CONNECT

Studies

Participants

Member Research Interests

Most Frequent Diagnoses

Kids First Data Resource Portal

DEMOCRATIZING ACCESS TO PEDIATRIC GENOMICS DATA

CAVATICA

Gabriella Miller Kids First

Cavatica Workspace

Welcome to Terra

Terra is a cloud-native platform for biomedical researchers to **access data**, **run analysis tools**, and **collaborate**.

View Workspaces

View Examples

Browse Data

Terra Workspace

Any Portal to Any Workspace

Other Portal

Other Workspace

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

Borrowed from Brian O'Connor / Jack DiGiovanna

Layers of Interoperability

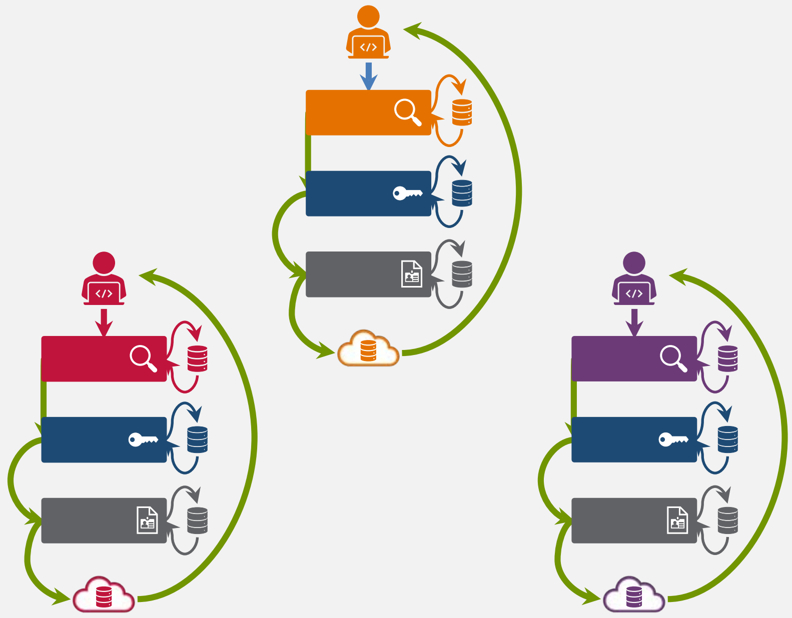
Challenge	NCPI Activities
Operational barriers to trans-platform data sharing	Establish principles for promoting interoperability across multiple platforms.
Inability to search & access data across platforms	Test & implement technical standards for data exchange (e.g. GA4GH APIs) based on key use cases
Teach researchers to use the cloud	Create public “knowledge base” with training materials and cloud cost guide.
Lack of standards for clinical data exchange	Pilot and assess FHIR resources to model and share complex clinical and phenotypic data

NIH Researcher Auth Services (RAS)

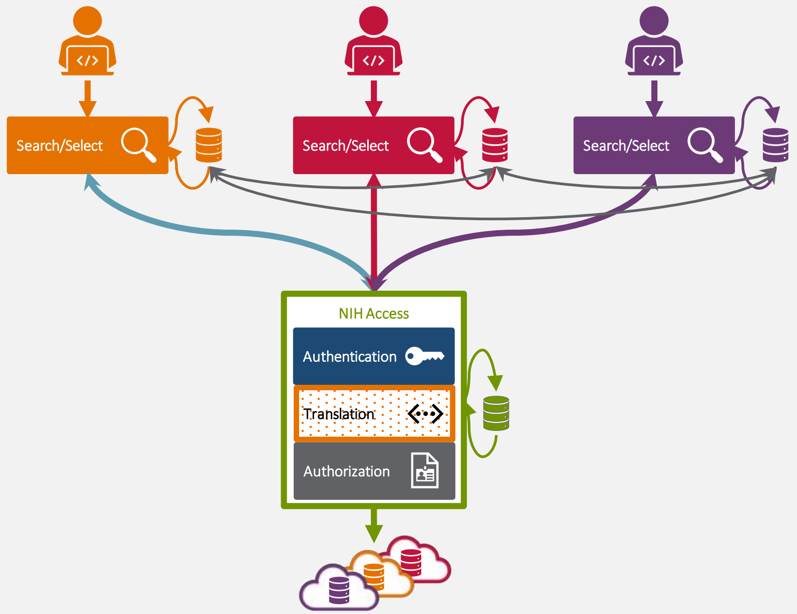
Simplify researcher access to NIH data through
federated **authentication** (linking user identity account; “passport”)
and **authorization** (claim to access specific studies/datasets; “visa”)

https://datascience.nih.gov/data-infrastructure/researcher-auth-service

Current

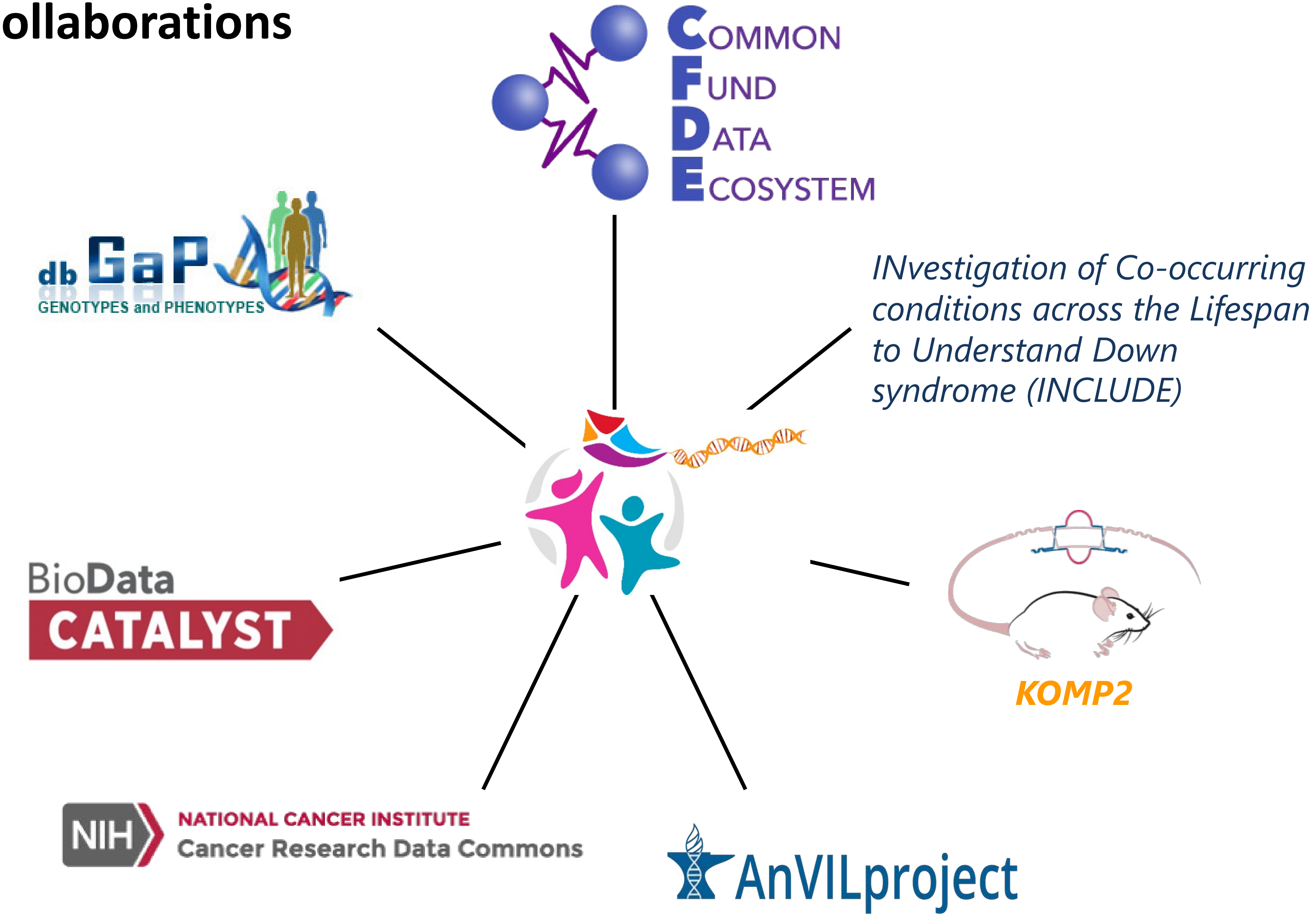


Summer 2020



Adapted from Susan Gregurick, ODSS

Collaborations


Common Fund Data Ecosystem
db GaP
GENOTYPES and PHENOTYPES
BioData
CATALYST
NIH
NATIONAL CANCER INSTITUTE
Cancer Research Data Commons
AnVILproject

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

KOMP2

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 can I get involved?

How do I access data?

What funding opportunities are available?

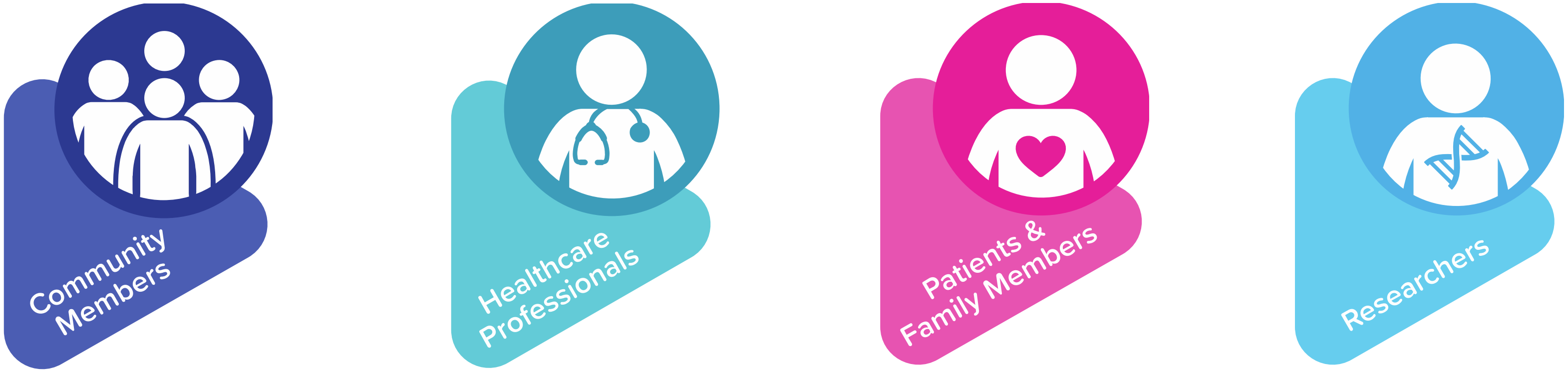
See: FAQs for Funding Opportunities Announcements (FOAs) to Support Data Analyses of Kids First Datasets (https://commonfund.nih.gov/kidsfirst/FAQ)

- **Kids First cohort sequencing opportunity (X01):**
  - 1 more reissue of PAR-19-390 **for 2021**
- **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

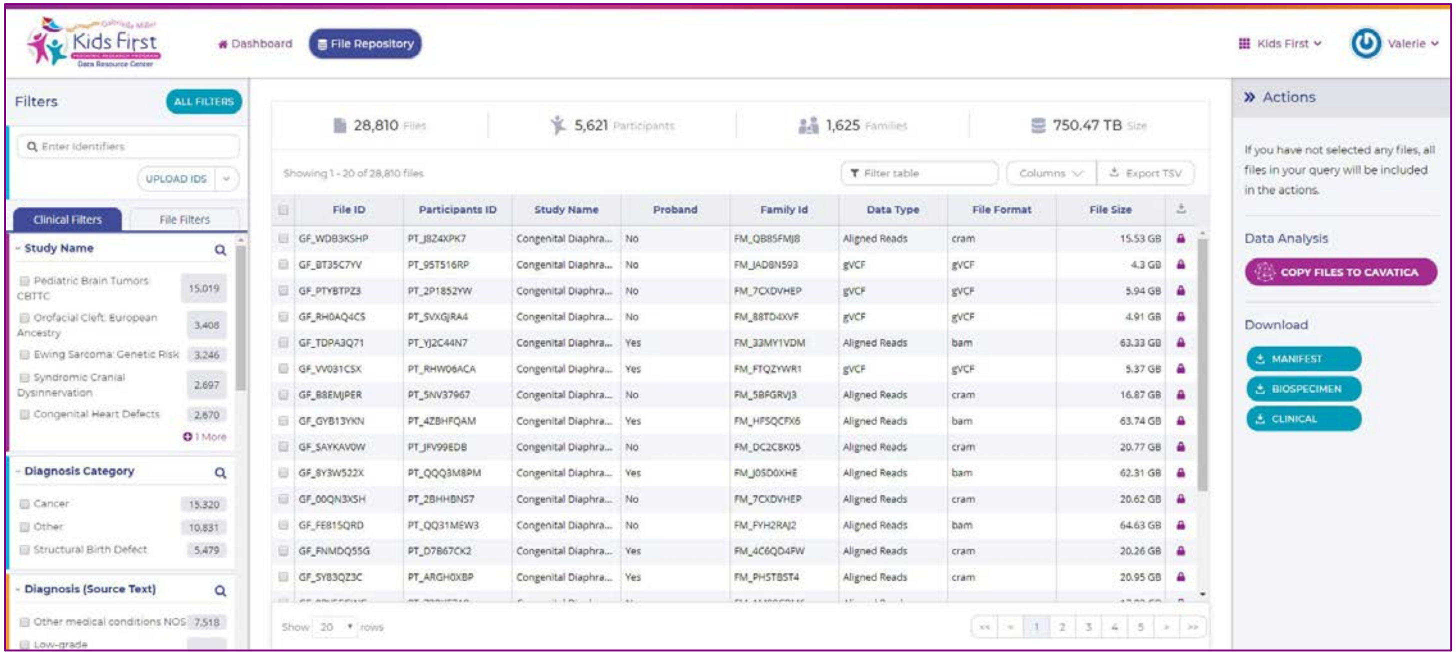
How can I get involved?

- **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/



How do I access data?

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



Kids First Data Resource Center | Dashboard | File Repository | Kids First | Valerie

Filters | ALL FILTERS

Enter Identifiers | UPLOAD IDS

Clinical Filters | File Filters

Study Name

- □ Pediatric Brain Tumors CBTTC 15,019
- □ Orofacial Cleft: European Ancestry 3,408
- □ Ewing Sarcoma: Genetic Risk 3,246
- □ Syndromic Cranial Dysinnervation 2,697
- □ Congenital Heart Defects 2,670
- 1 More

Diagnosis Category

- □ Cancer 15,320
- □ Other 10,831
- □ Structural Birth Defect 5,479

Diagnosis (Source Text)

- □ Other medical conditions NOS 7,518
- □ Low-grade

28,810 Files | 5,621 Participants | 1,625 Families | 750.47 TB Size

Showing 1 - 20 of 28,810 files | Filter table | Columns | Export TSV

□	File ID	Participants ID	Study Name	Proband	Family Id	Data Type	File Format	File Size	
□	GF_WDB3KSHP	PT_J8Z4XPK7	Congenital Diaphra...	No	FM_QB85FMJ8	Aligned Reads	cram	15.53 GB	🔒
□	GF_BT35C7YV	PT_95T516RP	Congenital Diaphra...	No	FM_JAD8N593	gVCF	gVCF	4.3 GB	🔒
□	GF_PTYBTPZ3	PT_2P1852YW	Congenital Diaphra...	No	FM_7CXDVHEP	gVCF	gVCF	5.94 GB	🔒
□	GF_RH0AQ4CS	PT_SVXGJRA4	Congenital Diaphra...	No	FM_88TD4XVF	gVCF	gVCF	4.91 GB	🔒
□	GF_TDPA3Q71	PT_YJ2C44N7	Congenital Diaphra...	Yes	FM_33MY1VDM	Aligned Reads	bam	63.33 GB	🔒
□	GF_VV031CSX	PT_RHW06ACA	Congenital Diaphra...	Yes	FM_FTQZYWR1	gVCF	gVCF	5.37 GB	🔒
□	GF_B8EMJPER	PT_5NV37967	Congenital Diaphra...	No	FM_5BFGRVJ3	Aligned Reads	cram	16.87 GB	🔒
□	GF_GYB13YKN	PT_4ZBHFQAM	Congenital Diaphra...	Yes	FM_HFSQCFX6	Aligned Reads	bam	63.74 GB	🔒
□	GF_SAYKAV0W	PT_JFV99EDB	Congenital Diaphra...	No	FM_DC2C8K05	Aligned Reads	cram	20.77 GB	🔒
□	GF_8Y3W522X	PT_QQQ3M8PM	Congenital Diaphra...	Yes	FM_J0SD0XHE	Aligned Reads	bam	62.31 GB	🔒
□	GF_00QN3XSH	PT_2BHHBNS7	Congenital Diaphra...	No	FM_7CXDVHEP	Aligned Reads	cram	20.62 GB	🔒
□	GF_FE815QRD	PT_QQ31MEW3	Congenital Diaphra...	No	FM_FYH2RAJ2	Aligned Reads	bam	64.63 GB	🔒
□	GF_FNMDQ55G	PT_D7B67CK2	Congenital Diaphra...	Yes	FM_4C6QD4FW	Aligned Reads	cram	20.26 GB	🔒
□	GF_SY83QZ3C	PT_ARGH0XBP	Congenital Diaphra...	Yes	FM_PHSTBST4	Aligned Reads	cram	20.95 GB	🔒

Show 20 rows | << < 1 2 3 4 5 > >>

Actions

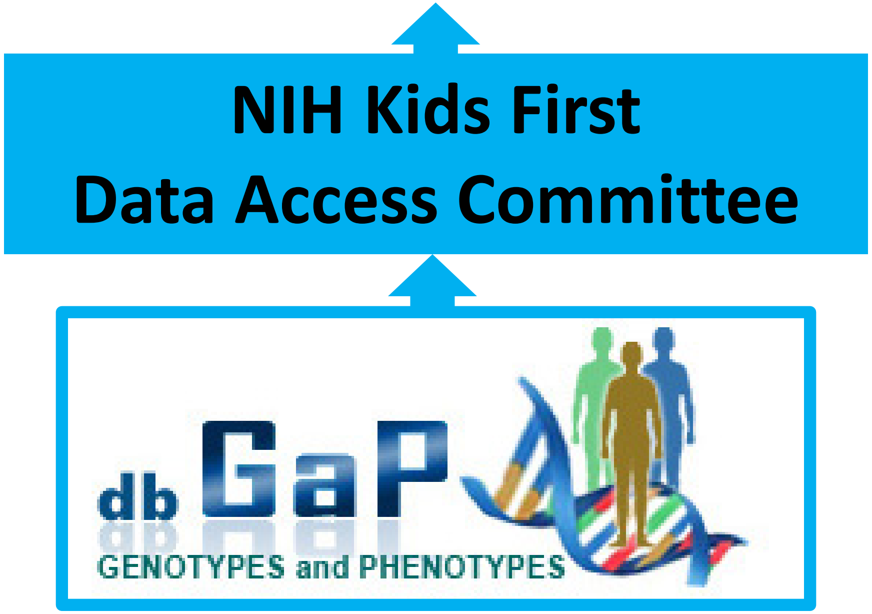
If you have not selected any files, all files in your query will be included in the actions.

Data Analysis

COPY FILES TO CAVATICA

Download

- MANIFEST
- BIOSPECIMEN
- CLINICAL



NIH Kids First Data Access Committee

Submit [dbGaP Data Access Requests (DARs)]() for **individual-level sequence data**

Push approved sequence data to Cavatica from the portal:
https://kidsfirstdrc.org/support/analyze-data/

Individual-level sequence data

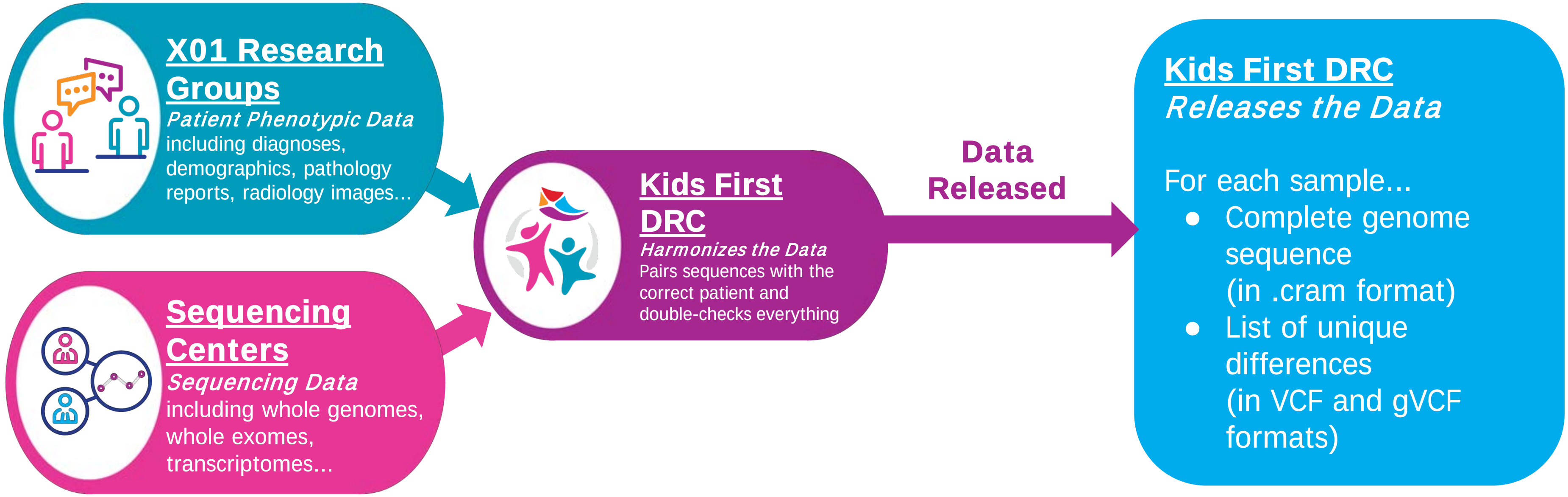
- To learn more about submitting dbGaP Data Access Requests (DARs) watch: https://www.youtube.com/watch?v=39cba0gF2tw&index=3&t=503s&list=PLoXwgZflAe4aMwWpVQU_WVeWHzyhI3BCu

dbGaP
GENOTYPES and PHENOTYPES

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

Submitting an Approvable dbGaP Data Access Request
Vivian Ota Wang, Ph.D
Office of Data Sharing
NCI

How are sequences released by the Kids First DRC?



X01 Research Groups

Patient Phenotypic Data

including diagnoses, demographics, pathology reports, radiology images...

Sequencing Centers

Sequencing Data

including whole genomes, whole exomes, transcriptomes...

Kids First DRC

Harmonizes the Data

Pairs sequences with the correct patient and double-checks everything

Data Released

Kids First DRC

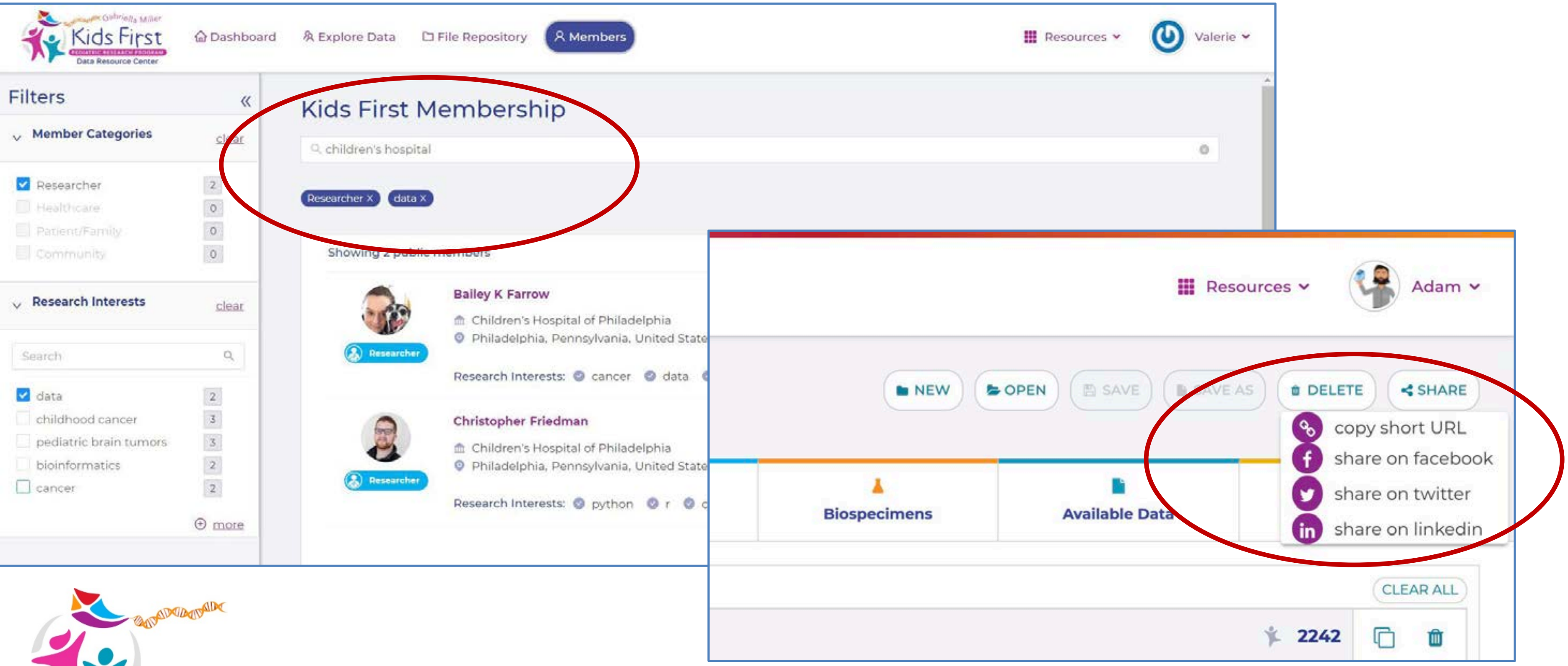
Releases the Data

For each sample...

- Complete genome sequence (in .cram format)
- List of unique differences (in VCF and gVCF formats)

How can I interact with other community members?

What community resources are available?



Kids First Membership

children's hospital

Researcher X data X

Showing 2 public members

☑ Researcher	2
☐ Healthcare	0
☐ Patient/Family	0
☐ Community	0

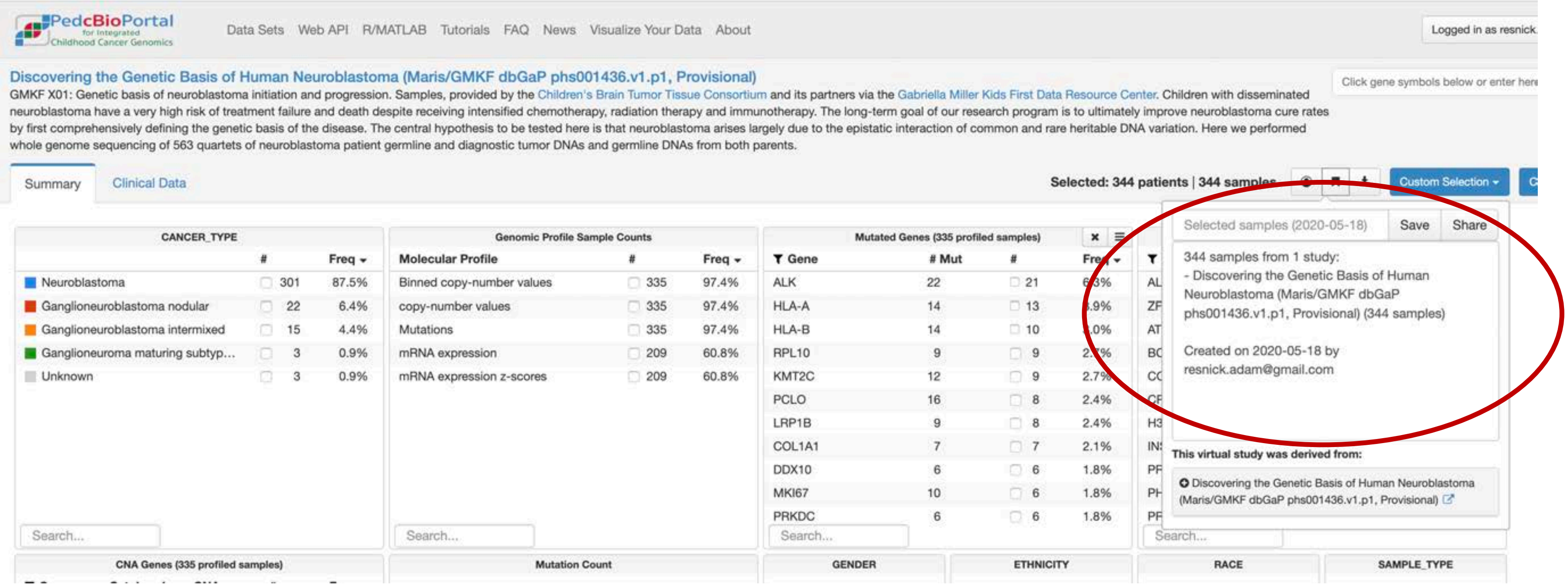
☑ data	2
☐ childhood cancer	3
☐ pediatric brain tumors	3
☐ bioinformatics	2
☐ cancer	2

Bailey K Farrow
Researcher
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, United State
Research Interests: cancer data

Christopher Friedman
Researcher
Children's Hospital of Philadelphia
Philadelphia, Pennsylvania, United State
Research Interests: python r

How can I interact with other community members?

What community resources are available?



Discovering the Genetic Basis of Human Neuroblastoma (Maris/GMKF dbGaP phs001436.v1.p1, Provisional)

GMKF X01: Genetic basis of neuroblastoma initiation and progression. Samples, provided by the Children's Brain Tumor Tissue Consortium and its partners via the Gabriella Miller Kids First Data Resource Center. Children with disseminated neuroblastoma have a very high risk of treatment failure and death despite receiving intensified chemotherapy, radiation therapy and immunotherapy. The long-term goal of our research program is to ultimately improve neuroblastoma cure rates by first comprehensively defining the genetic basis of the disease. The central hypothesis to be tested here is that neuroblastoma arises largely due to the epistatic interaction of common and rare heritable DNA variation. Here we performed whole genome sequencing of 563 quartets of neuroblastoma patient germline and diagnostic tumor DNAs and germline DNAs from both parents.

Selected: 344 patients | 344 samples

CANCER_TYPE

	#	Freq ▾
🟦 Neuroblastoma	□ 301	87.5%
🟥 Ganglioneuroblastoma nodular	□ 22	6.4%
🟧 Ganglioneuroblastoma intermixed	□ 15	4.4%
🟩 Ganglioneuroma maturing subtyp...	□ 3	0.9%
⬜ Unknown	□ 3	0.9%

Genomic Profile Sample Counts

Molecular Profile	#	Freq ▾
Binned copy-number values	□ 335	97.4%
copy-number values	□ 335	97.4%
Mutations	□ 335	97.4%
mRNA expression	□ 209	60.8%
mRNA expression z-scores	□ 209	60.8%

Mutated Genes (335 profiled samples)

Gene	# Mut	#	Freq ▾
ALK	22	□ 21	6.3%
HLA-A	14	□ 13	3.9%
HLA-B	14	□ 10	3.0%
RPL10	9	□ 9	2.7%
KMT2C	12	□ 9	2.7%
PCLO	16	□ 8	2.4%
LRP1B	9	□ 8	2.4%
COL1A1	7	□ 7	2.1%
DDX10	6	□ 6	1.8%
MKI67	10	□ 6	1.8%
PRKDC	6	□ 6	1.8%

Selected samples (2020-05-18)

344 samples from 1 study:
- Discovering the Genetic Basis of Human Neuroblastoma (Maris/GMKF dbGaP phs001436.v1.p1, Provisional) (344 samples)

Created on 2020-05-18 by resnick.adam@gmail.com

This virtual study was derived from:

Discovering the Genetic Basis of Human Neuroblastoma (Maris/GMKF dbGaP phs001436.v1.p1, Provisional)

CNA Genes (335 profiled samples)

Mutation Count

GENDER

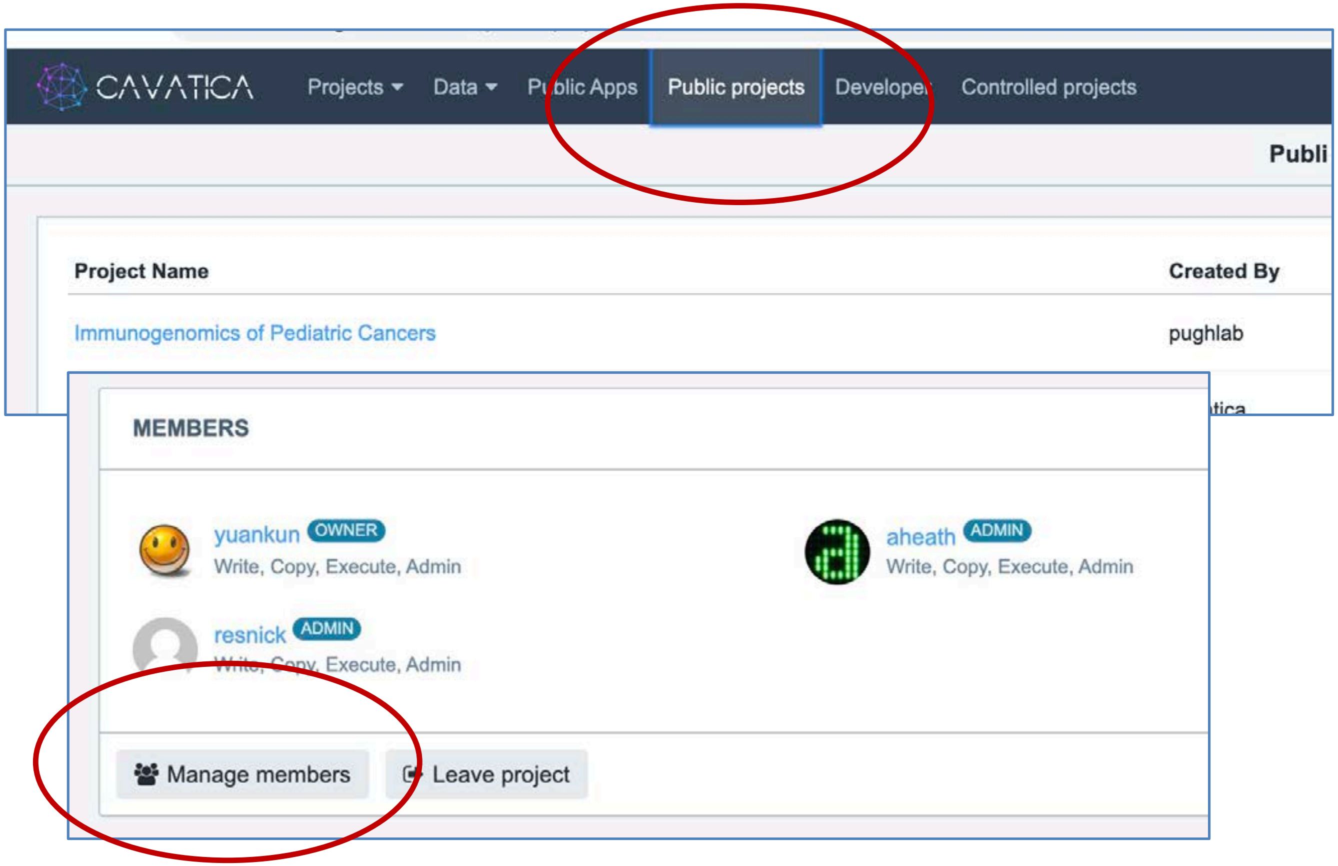
ETHNICITY

RACE

SAMPLE_TYPE

How can I interact with other community members?

What community resources are available?



Project Name	Created By
Immunogenomics of Pediatric Cancers	pughlab

MEMBERS

yuankun OWNER
Write, Copy, Execute, Admin

aheath ADMIN
Write, Copy, Execute, Admin

resnick ADMIN
Write, Copy, Execute, Admin

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 can I get involved?

How do I access data?

Thank You!

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

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM