Interoperability between Kids First & Undiagnosed Diseases Network (UDN) Data via dbGaP/SRA

Valerie Cotton & Allison Heath

Overall Goals

Use Case: *Enable researchers to easily co-analyze data from Kids First & the Undiagnosed Disease Network in the cloud to leverage large-scale pediatric cohorts from Kids First to resolve variants of unknown significance in UDN cases.*

Kids First: The goal of Kids First is to help researchers uncover new insights into the biology of childhood cancer and structural birth defects.

UDN: The Undiagnosed Diseases Network (UDN) is an initiative to facilitate the diagnosis of conditions that have eluded diagnosis through the coordinated action of leading clinical and research centers.

UDN & Pediatric Genomics

GENOME SEQUENCING

1,142 participants (716 children and 426 adults) have undergone genome sequencing. Many of these participants had non-diagnostic exome sequencing prior to enrollment in the UDN. The most common symptom category for participants undergoing genome sequencing is neurology (51%), followed by multiple congenital anomalies (9%).

- **Data access provided by:** dbGaP Authorized Access
- **Release Date:** September 27, 2021
- **Embargo Release Date:** September 27, 2021
- Data Use Certification Requirements (DUC)
- **Public Posting of Genomic Summary Results:** Allowed
- **Use Restrictions**

Consent group	Is IRB required?	Data Access Committee	Number of participants
General Research Use	No	National Human Genome Research Institute (nhgridac@mail.nih.gov)	4239

Scientific Narrative (specific use case)

...To address the challenge of VUS's, we have developed a pipeline to assess variants found on clinical sequencing using biobank cohorts with linked phenotyped data.

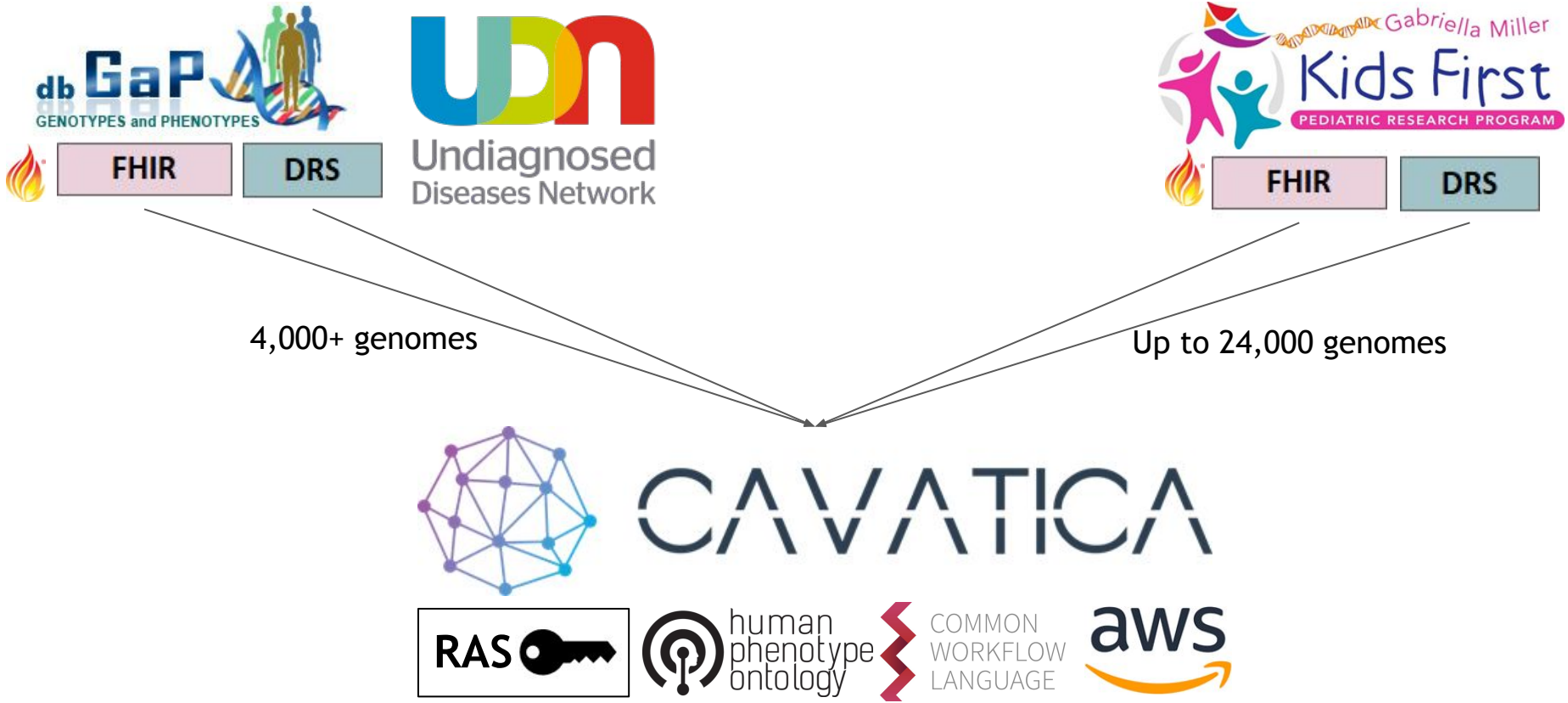
Our pipeline creates a **phenotype risk score (PheRS)** of the proband based on their clinical presentation described in human phenotype ontology terms (HPO). We then apply the PheRS to the biobank cohort, such that individuals with many overlapping features have a high PheRS, and those with no or few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

We have been using this pipeline to analyze **Undiagnosed Disease Network (UDN)** patients, using a biobank cohort called BioVU… We believe that expanding our search for variant matched individuals to a large cohort like **Kids First** would enable us better interpret candidate variants for unsolved UDN cases.....

Lisa Bastarache

Undiagnosed Diseases Network

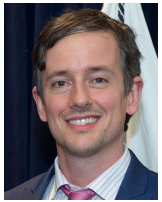
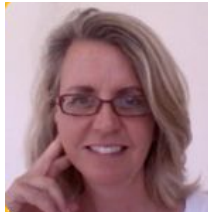
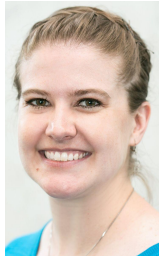
Overview of Standards Used



Solution Matrix

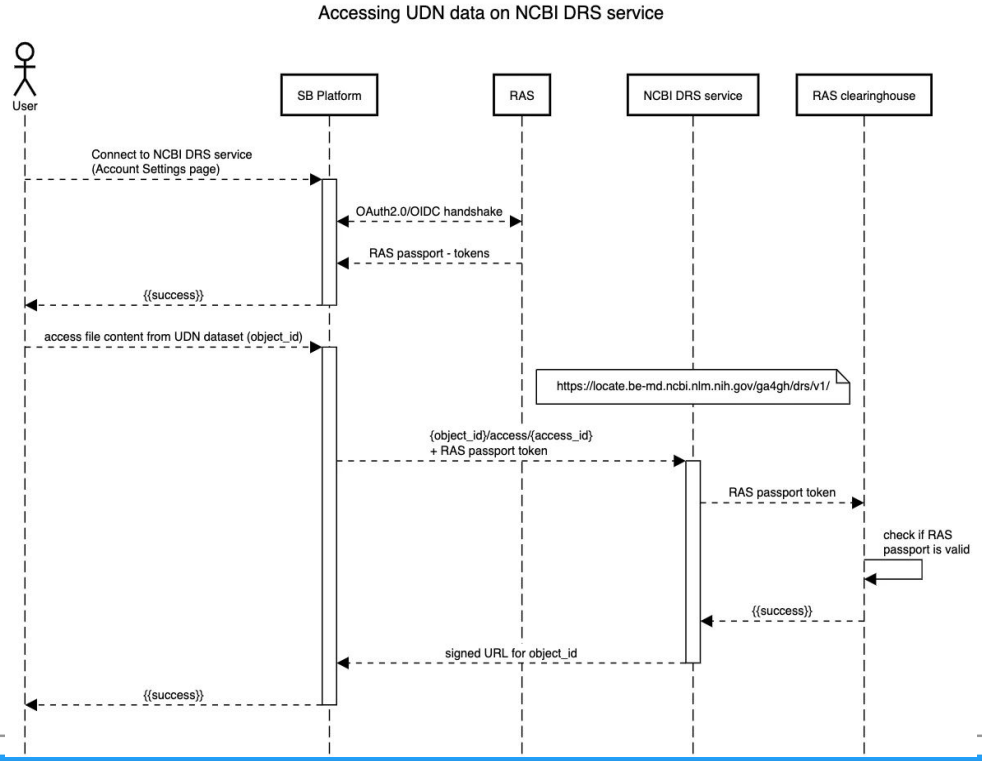
	Kids First Data Resource	NLM/NCBI	Analysis Tools
Genomic data	CAVATICA already integrated with the Kids First/Gen3 **DRS** server. **RAS** Milestone 3 is underway.	Connect CAVATICA to dbGaP **DRS** server, using RAS v1.1 Passports - Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identity variants of unknown significance (VUSs) underlying undiagnosed conditions and “matched” cases in Kids First
Phenotypic data	CAVATICA is building a FHIR client to ingest from the Kids First FHIR-based data service	dbGaP on FHIR is in development. FHIR & RAS integration will be needed for controlled-access phenotypes	PheRS to compare phenotypes of individuals with the same/similar VUSs

Collaboration Matrix

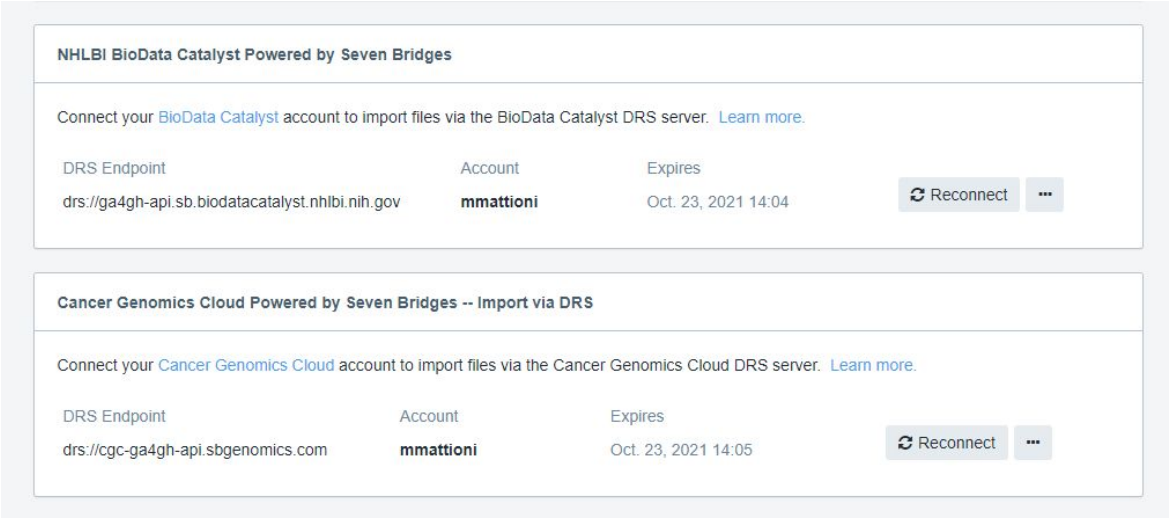
	Kids First Data Resource	NLM/NCBI	Tester/User
Genomic data	Michele Mattioni & Jack DiGiovanna & Adam Resnick   	Kurt Rodarmer & Yuriy Skripchenko  	Yuankun Zhu & Anne Deslattes Mays  
Phenotypic data	Allison Heath & Robert Carroll  	Liz Amos & Mike Feolo  	Lisa Bastarache 

Genomic Data Interoperability

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



CAVATICA: RAS Connection



Connect with the NCBI DRS Server

DRS EndPoint
https://locate.be-md.ncbi.nlm.nih.gov/ga4gh/drs/v1/ *Connect*

- Seven Bridges identified solution to add a **new "card"** in the Account DataSets configuration tab

DRS links

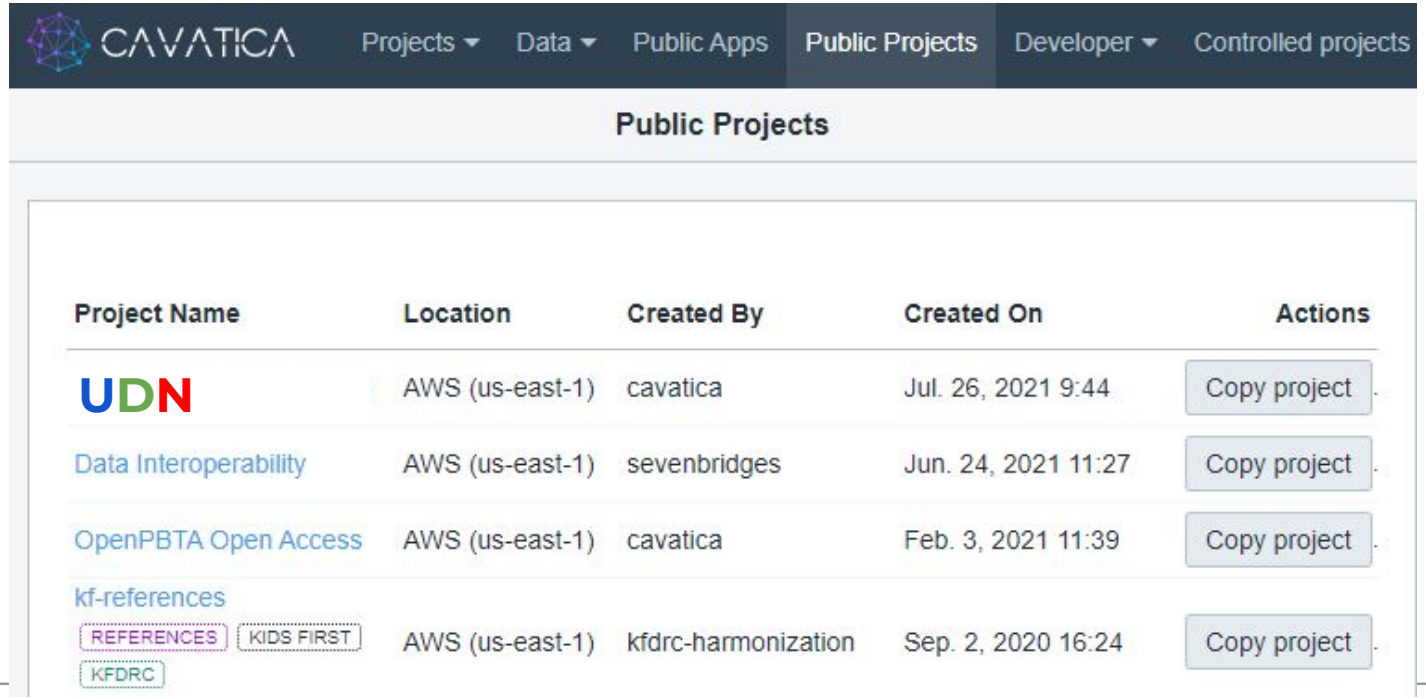
1. Use NCBI Run Selector to obtain a manifest which contains SRA Runs
2. Use the IDX service to obtain the DRS links connected with the SRA Runs
   - *Note: The DRS Links are offered in bundles, which Seven Bridges needs to build support for*
   - At the moment Seven Bridges extract the bundles, and then obtains the DRS pointer to the file
3. Import the DRS File into Cavatica

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

	Run	BioSample	alignment_software	analyte_type	Assay Type	biospecimen_repository_sample_id	body_site	Bytes	Center Name
1	SRR5031422	SAMN05980034	BWA-mem v0.7.12	DNA	WGS	8657f8fb-432b-4473-a31b-060384c4b79f	Blood	55.83 Gb	NHGRI-PHS001232
2	SRR5031424	SAMN05980042	BWA-mem v0.7.12	DNA	WGS	57b49db5-2778-4557-9fbb-9ff454cf4212	Blood	61.43 Gb	NHGRI-PHS001232
3	SRR5031427	SAMN05980030	BWA-mem v0.7.12	DNA	WGS	a2529ebc-4e29-4d60-93a8-fa07ed9f84a4	Blood	78.20 Gb	NHGRI-PHS001232
4	SRR5031429	SAMN05980037	BWA-mem v0.7.12	DNA	WGS	33ad6df6-f122-4e14-b75f-82733c39a220	Blood	82.83 Gb	NHGRI-PHS001232
5	SRR5031431	SAMN05980040	BWA-mem v0.7.12	DNA	WGS	6f94ba61-73d7-4551-80b3-6591001c437a	Blood	74.80 Gb	NHGRI-PHS001232
6	SRR5031434	SAMN05980032	BWA-mem v0.7.12	DNA	WGS	e8bb68df-e276-4604-94cf-05b57902f337	Blood	72.77 Gb	NHGRI-PHS001232
7	SRR8257099	SAMN10087985	BWA-mem v0.7.12	DNA	WGS	c6c974cc-86e8-42d8-92ba-ab10f1b37557	Blood	17.33 Gb	HMS-CC
8	SRR8060841	SAMN10087770	BWA-mem v0.7.12	DNA	WGS	31e6d861-ccb8-41c2-9ebc-c4e05251e690	Blood	51.81 Gb	HMS-CC
9	SRR8060848	SAMN10087458	BWA-mem v0.7.12	DNA	WGS	1735b288-f786-4148-86f2-0efc61d7cc2f	Blood	17.78 Gb	HMS-CC

Draft Approach for UDN Data Findability

The dataset will be findable/searchable as a CAVATICA Public Project (dbGaP approval still required). The DRS file would be built into the Project.



Variant Identification

- For functional equivalence, call UDN variants using Kids First workflows
- Use Kids First Portal variant search to identify datasets of interest → Apply for those datasets in dbGaP
- Use Kids First VCFs to identify variant matched individuals
- Run PheRS

Variant	Type	dbSnp	Consequences	CLINVAR	Studies	Participants
chrX:g.48792004del	deletion	--	frameshift_variant GATA1 G126X	--	1	1 / 4843
chrX:g.48794116del	deletion	--	frameshift_variant GATA1 G397X	--	1	1 / 4843
chrX:g.48791978C>A	SNV	--	missense_variant GATA1 Q119K	--	1	1 / 4843
chrX:g.48792194C>T	SNV	rs140561920	missense_variant GATA1 R191C	Benign	1	4 / 4843

Solution Matrix

	Kids First Data Resource	NLM/NCBI	Analysis Tools
Genomic data	CAVATICA already integrated with the Kids First/Gen3 **DRS** server. **RAS** Milestone 3 is underway.	Connect CAVATICA to dbGaP **DRS** server, using RAS v1.1 Passports - Requires BAMs in S3 storage (US East1 to avoid egress)	Variant calling and searching across UDN & Kids First to identity variants of unknown significance (VUSs) underlying undiagnosed conditions and “matched” cases in Kids First
Phenotypic data	CAVATICA is building a FHIR client to ingest from the Kids First FHIR-based data service	dbGaP on FHIR is in development. FHIR & RAS integration will be needed for controlled-access phenotypes	PheRS to compare phenotypes of individuals with the same/similar VUSs

PheRS pipeline

- R-based tool creates a phenotype risk score (PheRS) of the proband based on their clinical presentation described in human phenotype ontology terms (HPO).
  - ✔**Kids First already maps phenotypes to HPO**
- Apply PheRS to the cohort, such that individuals with many overlapping features have a high PheRS, and those with no or few overlapping features have a low score.
- Identify variant matched individuals and test if they have unexpectedly elevated phenotype risk scores
- Make available to the community and path for utilization/comparision with other work like LIRICAL

Undiagnosed Diseases Network

Proband phenotype

Clinical symptoms and physical findings

GROWTH PARAMETERS
- Failure to thrive

CARDIOVASCULAR
- Patent ductus arteriosus

GASTROINTESTINAL
- Elevated hepatic transaminase
- Gastroesophageal reflux

GENITOURINARY
- Hydrocele testis

BEHAVIOR, COGNITION AND DEVELOPMENT
- Global developmental delay
- Delayed speech and language development

DIGESTIVE SYSTEM
- Hepatomegaly

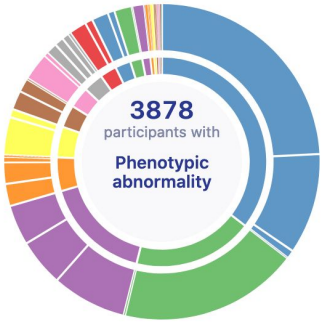
METABOLISM/HOMEOSTASIS
- Recurrent hypoglycemia
- Neonatal hypoglycemia

Candidate variants

Heterozygous Variants

Gene	Chr Position rs#	Change	Effect	Proband	Mother (Unaff)	Father (Unaff)
COL9A1 NM_001851.4	chr6	A → T	splice donor 10.9>2.7	●○	○○	●○
	70991091	c.876+2T>A				
	rs149830493					
ELN NM_000501	chr7	G → A	missense	●○	○○	●○
	73470684	c.1234G>A				
	rs375116795	p.Gly412Arg				
PIGN NM_012327	chr18	T → C	missense	●○	○○	●○
	59757754	c.2238A>G				
	rs200658159	p.Ile746Met				
POLG NM_002693.2	chr15	G → C	missense	●○	○○	●○
	89872002	c.1084C>G				
	rs763248358	p.Leu362Val				
RFT1 NM_052859.3	chr3	C → T	missense	●○	●○	○○
	53140879	c.782G>A				
	rs374781452	p.Arg261Gln				

Observed Phenotypes



Phenotypic abnormality (HP:0000118)	27	3878
Abnormality of head or neck (HP:0000152)	0	1480
Abnormality of the musculoskeletal system (HP:0033127)	0	1328
Abnormality of the cardiovascular system (HP:0001626)	41	957
Abnormality of the nervous system (HP:0000707)	15	431
Abnormality of the eye (HP:0000478)	7	355
Abnormality of the genitourinary system (HP:0000119)	25	341
Abnormality of the digestive system (HP:0025031)	80	327
Abnormality of the respiratory system (HP:0002086)	0	303
Neoplasm (HP:0002664)	0	278
Abnormality of the ear (HP:0000598)	8	266
Abnormality of the integument (HP:0001574)	1	196
Abnormality of limbs (HP:0040064)	53	190
Growth abnormality (HP:0001507)	0	90
Abnormality of the immune system (HP:0002715)	0	59
Abnormality of the endocrine system (HP:0000818)	0	41
Abnormality of prenatal development or birth (HP:0001197)	0	26
Abnormality of blood and blood-forming tissues (HP:0001871)	0	24
Abnormality of the breast (HP:0000769)	0	18
Abnormal cellular phenotype (HP:0025354)	0	10
Abnormality of metabolism/homeostasis (HP:0001939)	0	9

chr18:g.62090521T>C Germline

Summary | Frequencies | Clinical Associations

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

4 Studies

18 Participants

3.72e-3 Frequency

Type	Ref Genome	ClinVar	dbSNP
SNV	GRCh38	539565	rs200658159

Gene Consequences

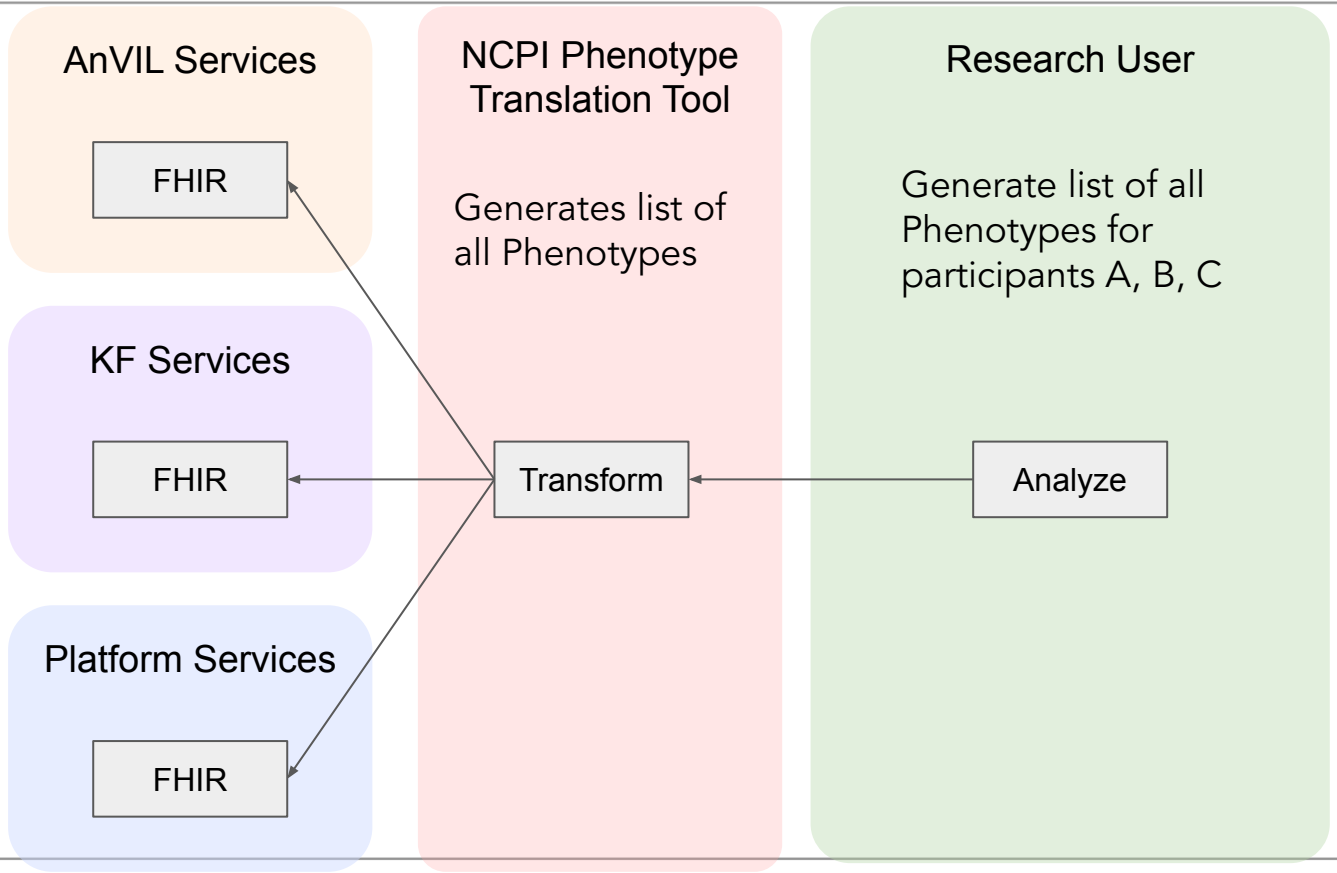
Gene PIGN

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

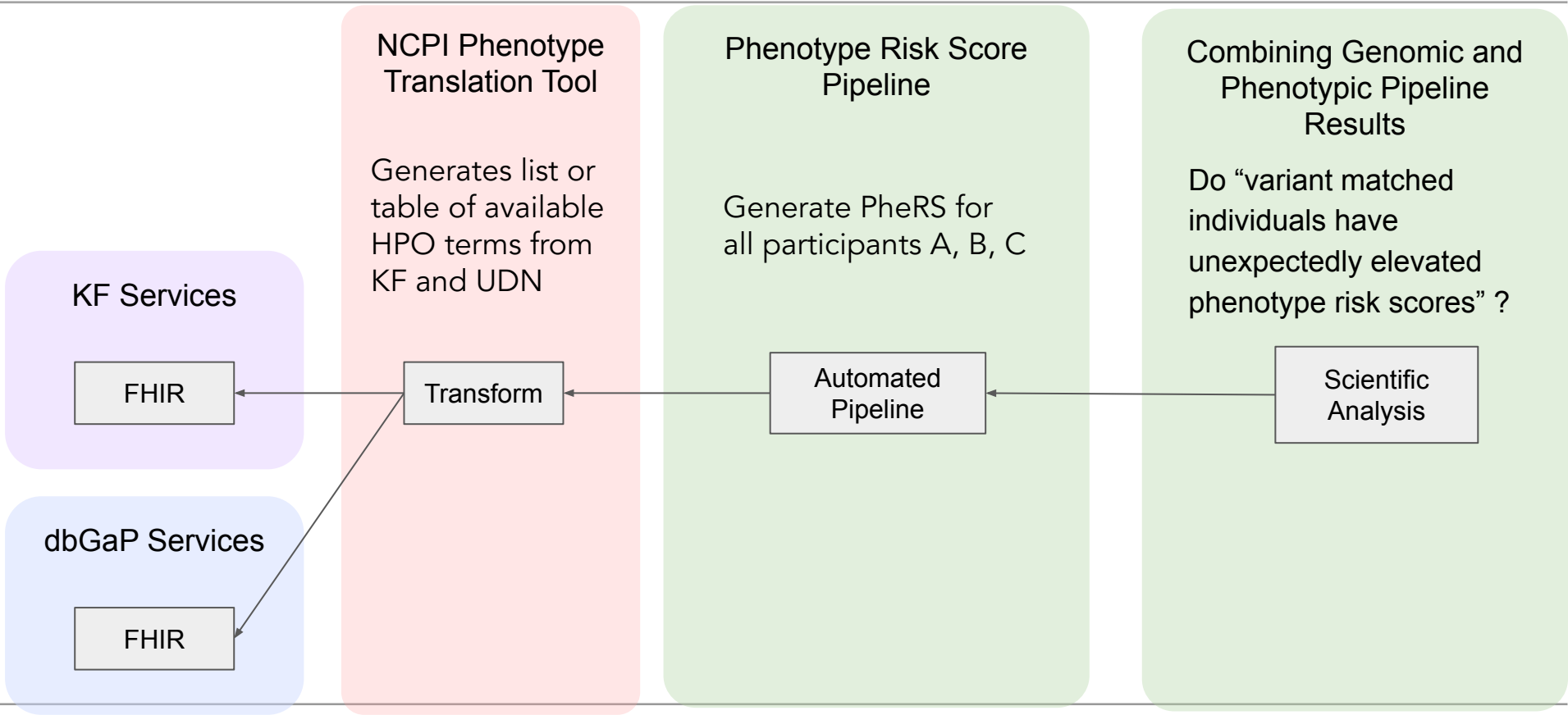
Show Transcripts (28)

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

Driving Tool / Service Layers: General



Driving Tool / Service Layers: Use Case

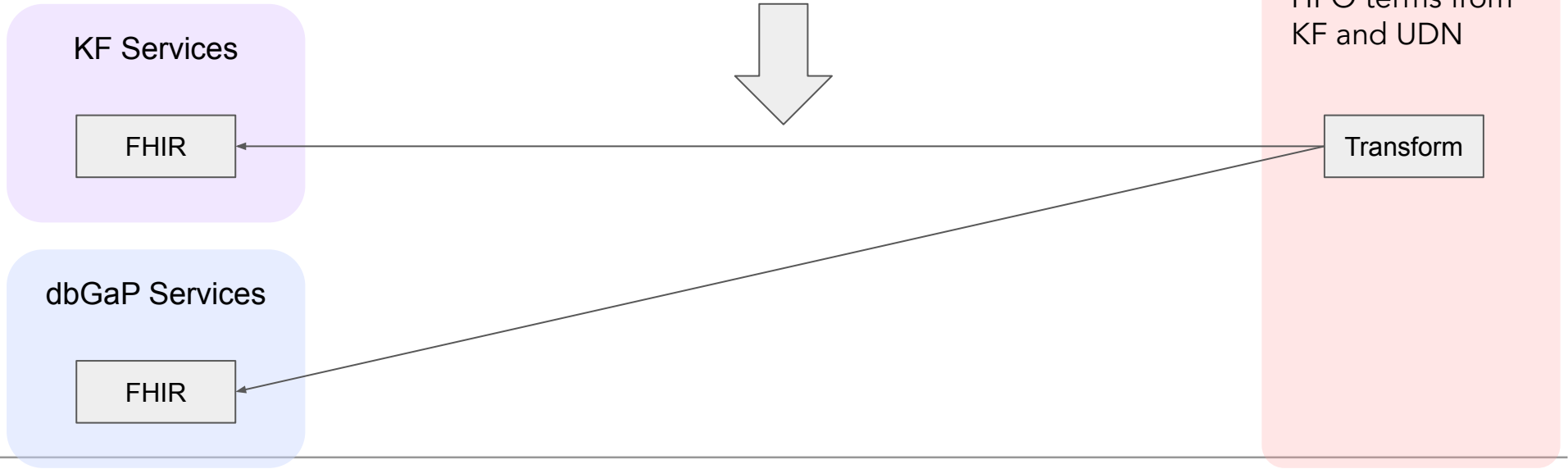


Concrete Progress on Each Step

What protocol is the most practical/useful?

- PFB?
- ndjson?
- FHAvros?

RAS-based access to FHIR data



Concrete Progress on Each Step

NCPI Phenotype Translation Tool

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

Transform

Phenotypic pipeline/analysis often a different “modality” than genomic pipeline/analysis - statistical analysis from a database. What current cloud workspace tooling fits best here? Do we need to be able to support additional capabilities?

Phenotype-based Pipeline

Generate PheRS for all participants A, B, C

Automated Pipeline

Concrete Progress on Each Step

Phenotype-based Pipeline

Generate PheRS for all participants A, B, C

Automated Pipeline

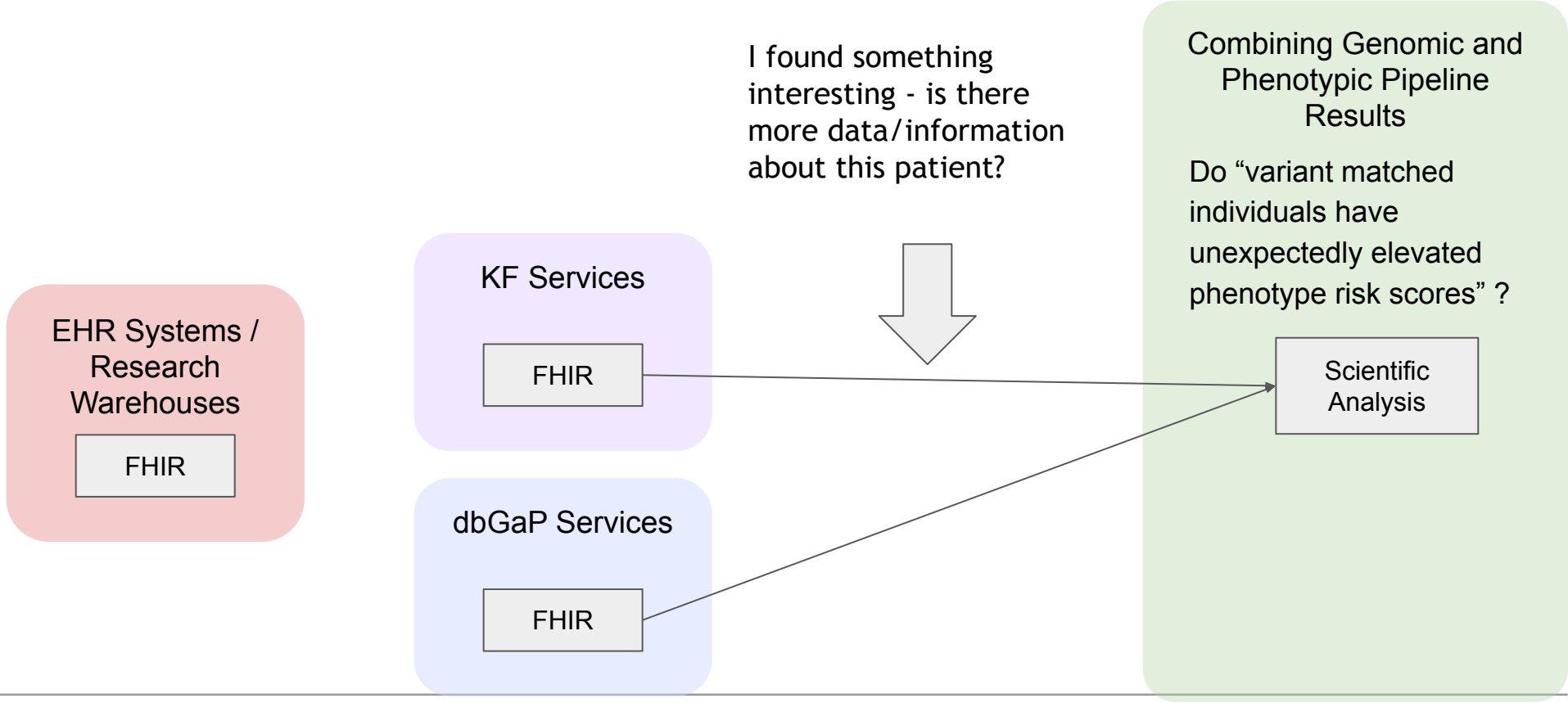
May be the most well-defined? Happens in a R Studio or Jupyter notebook environment?

Combining Genomic and Phenotypic Pipeline Results

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

Scientific Analysis

Doors to New Capabilities

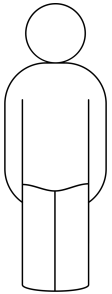


High-throughput phenotyping for Marfan Syndrome

MARFAN SYNDROME	HPO	Phecodes
HEAD & NECK		
Eyes		
- Retinal detachment	HP:0000541	[361] Retinal detachment & defects
- Iris hypoplasia	HP:0001083	[753.1] Congenital cataract & lens anomalies
CARDIOVASCULAR		
Heart		
- Aortic regurgitation	HP:0001653	[394.2] Mitral valve disease
Vascular		
- Aortic root dilatation	HP:0002616	[442.1] Aortic aneurysm
- Aortic dissection	HP:0002647	
SKELETAL		
Limbs		
- Joint hypermobility	HP:0001382	[728.2] Laxity of ligament or hypermobility
CHEST		
Ribs Sternum Clavicles & Scapulae		
- Pectus excavatum	HP:0000767	[756.21] Pectus excavatum
RESPIRATORY		
Lung		
- Pneumothorax	HP:0002107	[506] Empyema and pneumothorax

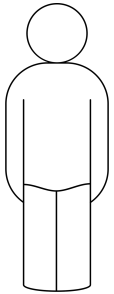
Phenotype Risk Score (PheRS) for Marfan Syndrome

Looks a lot like they have Marfan syndrome



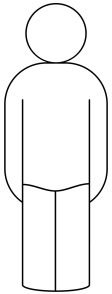
- Aortic aneurysm
- Tall stature (> 2 S.D.)
- Pectus excavatum
- Ectopia lentis

Doesn't look anything like they have Marfan syndrome

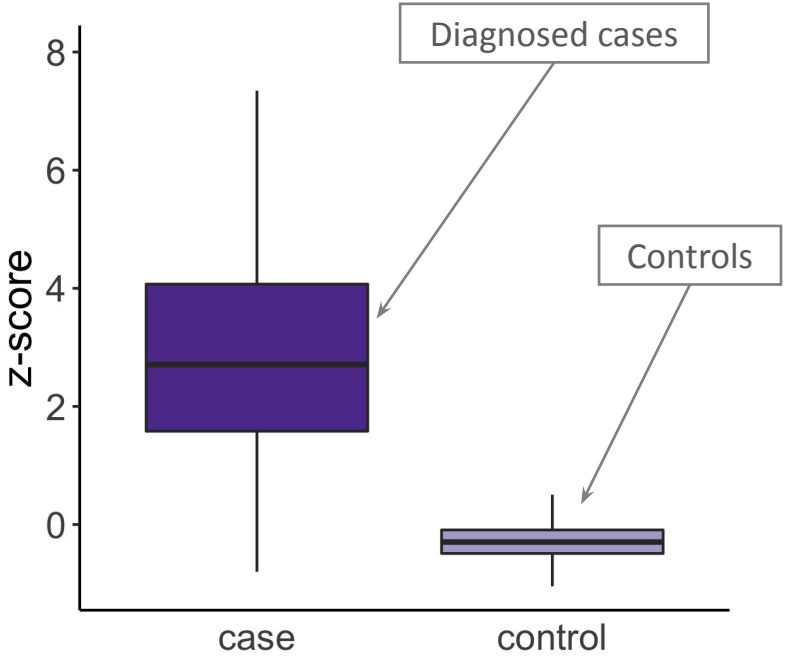


No Marfan symptoms

A little bit look like they have Marfan syndrome



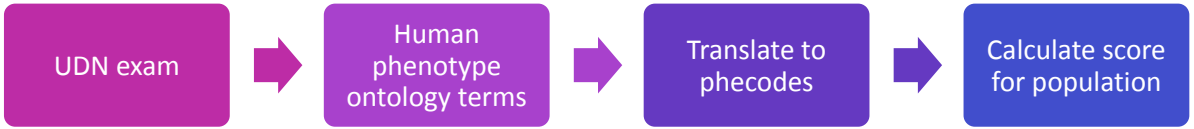
Aortic aneurysm

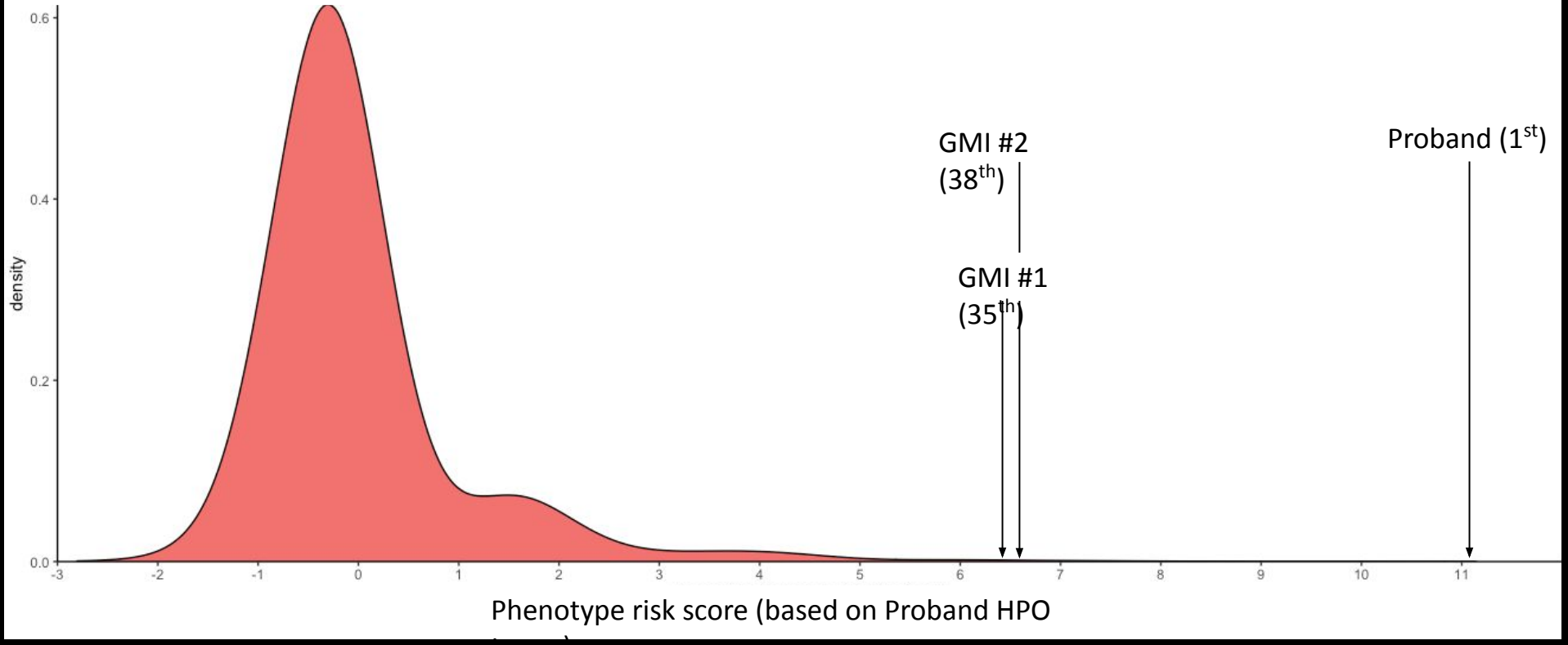


*You can differentiate a cohort diagnosed with Marfan syndrome using **only the features** of the disease*

Bastarache, et al. Science 2018

PHENOTYPE RISK SCORE FOR *MSL2*





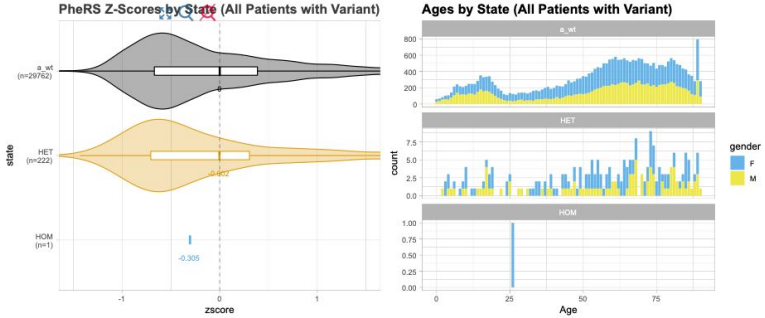
Variant Summary

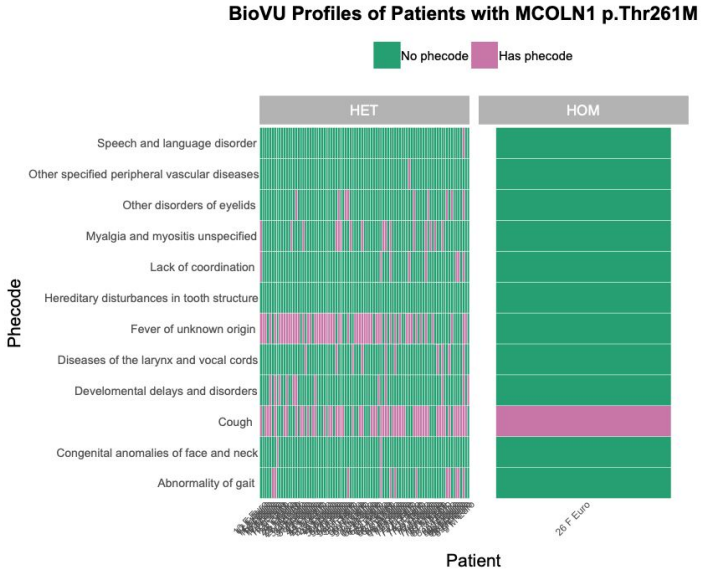
	ALL				ADULTS				PEDS			
	HOM		HET		HOM		HET		HOM		HET	
SNP	nHOM	zHOM	nHET	zHET	nHOM_A	zHOM_A	nHET_A	zHET_A	nHOM_P	zHOM_P	nHET_P	zHET_P
PLOD1 p.Arg512Cys (cHET) exm15796	2	1.52	287	0.08	1	1.58	236	0.11	1	0.90	51	0.05
SAMD9 p.Tyr896His (HET) exm634380	1	0.44	68	0.11	1	0.43	60	0.15	0	NA	8	-0.07
MCOLN1 p.Thr261Met (HET) exm1416067	1	-0.31	222	0.00	0	NA	181	0.04	1	0.06	41	-0.06
VPS13B p.Lys1129Arg (cHET) exm712265	1	-0.58	319	0.05	0	NA	273	0.05	1	-0.26	46	-0.05
EDNRB p.Val260Phe (HET) exm1073577	0	NA	100	-0.09	0	NA	89	-0.13	0	NA	11	0.40
FLNB p.Ala1341Gly (HET) exm326265	0	NA	22	0.47	0	NA	18	0.35	0	NA	4	0.41
CARD11 p.Ile544Leu (HET) exm600786	0	NA	176	0.05	0	NA	153	-0.03	0	NA	23	0.36
EHMT1 p.Ala369Thr (cHET) exm804322	0	NA	3	-0.23	0	NA	3	-0.27	0	NA	0	NA
EHMT1 p.Ala369Thr (HET) exm804322	0	NA	3	-0.23	0	NA	3	-0.27	0	NA	0	NA

Interactive visualization by Leigh Anne Tang

MCOLN1 P.THR261MET IS NOT LIKELY TO CAUSE PROBANDS PHENOTYPE

MCOLN1 p.Thr261Met (HET)





Interactive visualization by Leigh Anne Tang

Back up notes about DRS