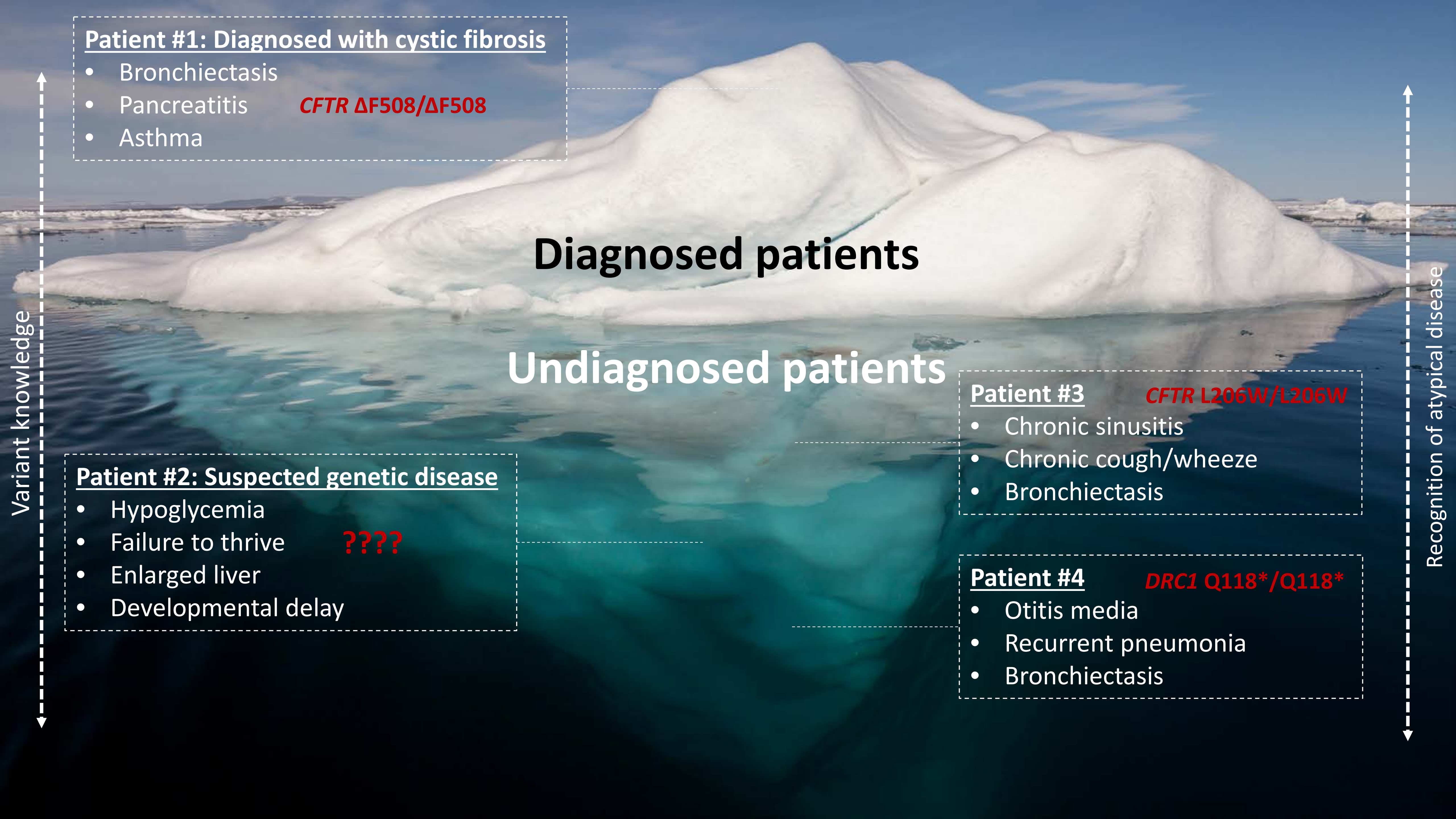
Phenotype mining in electronic health records

Genomic Medicine XI - September 5th, 2018

VANDERBILT UNIVERSITY

MEDICAL CENTER



Variant knowledge

Patient #1: Diagnosed with cystic fibrosis

- Bronchiectasis
- Pancreatitis
- Asthma

CFTR ΔF508/ΔF508

Diagnosed patients

Undiagnosed patients

Patient #2: Suspected genetic disease

- Hypoglycemia
- Failure to thrive
- Enlarged liver
- Developmental delay

????

Patient #3 *CFTR* L206W/L206W

- Chronic sinusitis
- Chronic cough/wheeze
- Bronchiectasis

Patient #4 *DRC1* Q118*/Q118*

- Otitis media
- Recurrent pneumonia
- Bronchiectasis

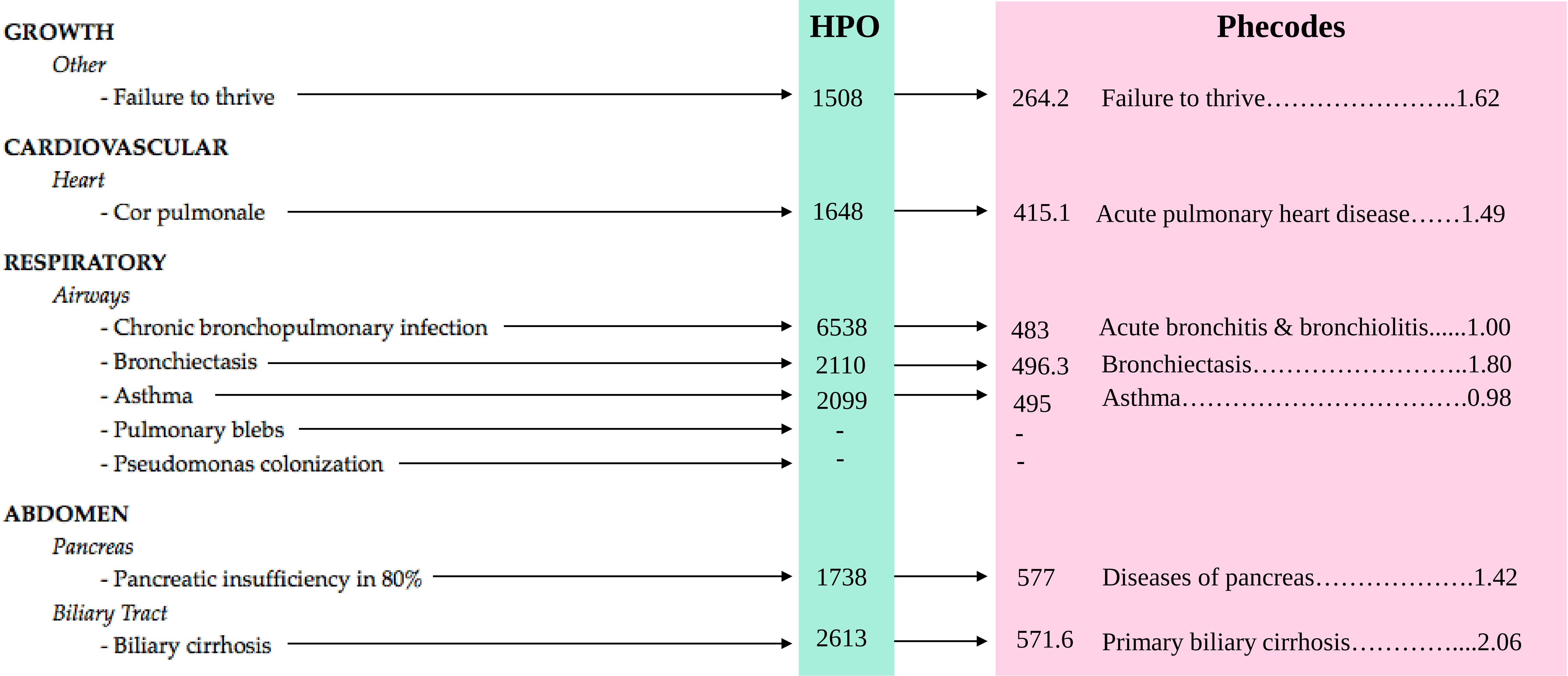
Recognition of atypical disease

50 YEARS OMIM
Human Genetics Knowledge for the World

CYSTIC FIBROSIS; CF

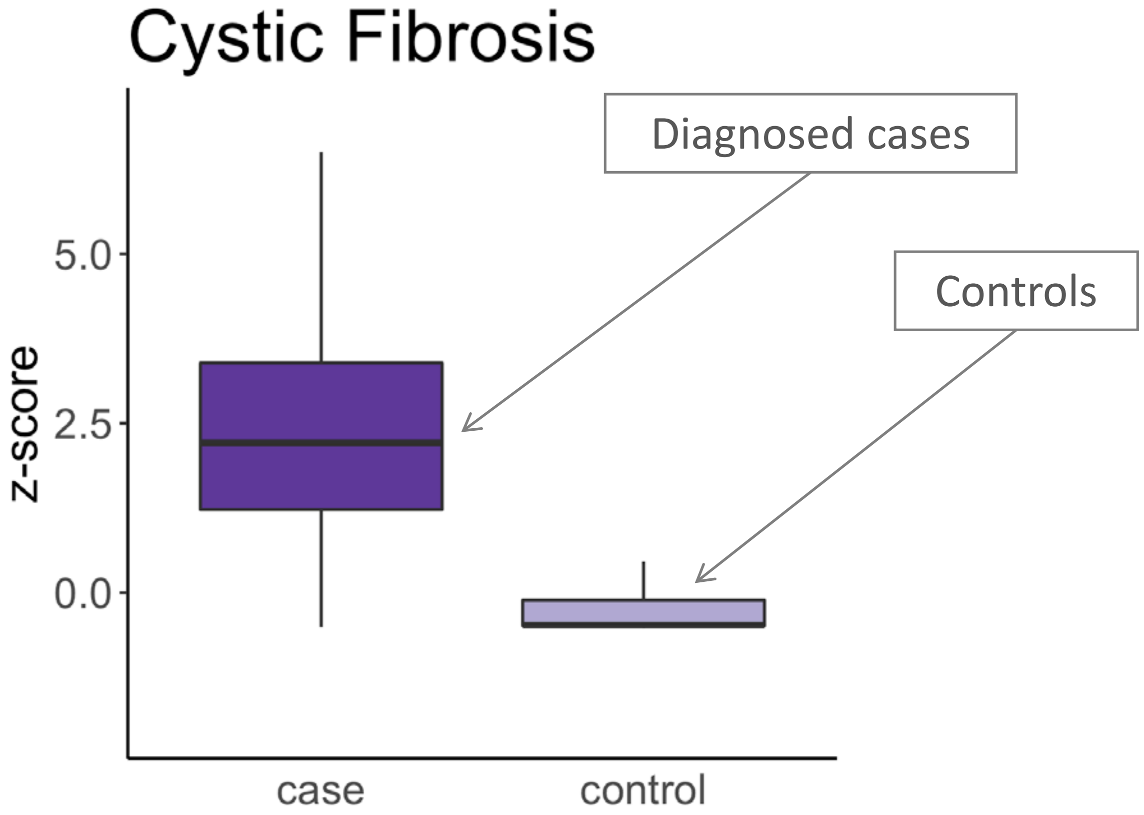
INHERITANCE

- Autosomal recessive



	HPO	Phecodes	
GROWTH			
Other			
- Failure to thrive	1508	264.2	Failure to thrive…………………..1.62
CARDIOVASCULAR			
Heart			
- Cor pulmonale	1648	415.1	Acute pulmonary heart disease……1.49
RESPIRATORY			
Airways			
- Chronic bronchopulmonary infection	6538	483	Acute bronchitis & bronchiolitis......1.00
- Bronchiectasis	2110	496.3	Bronchiectasis……………………..1.80
- Asthma	2099	495	Asthma…………………………….0.98
- Pulmonary blebs	-	-	
- Pseudomonas colonization	-	-	
ABDOMEN			
Pancreas			
- Pancreatic insufficiency in 80%	1738	577	Diseases of pancreas……………….1.42
Biliary Tract			
- Biliary cirrhosis	2613	571.6	Primary biliary cirrhosis…………....2.06

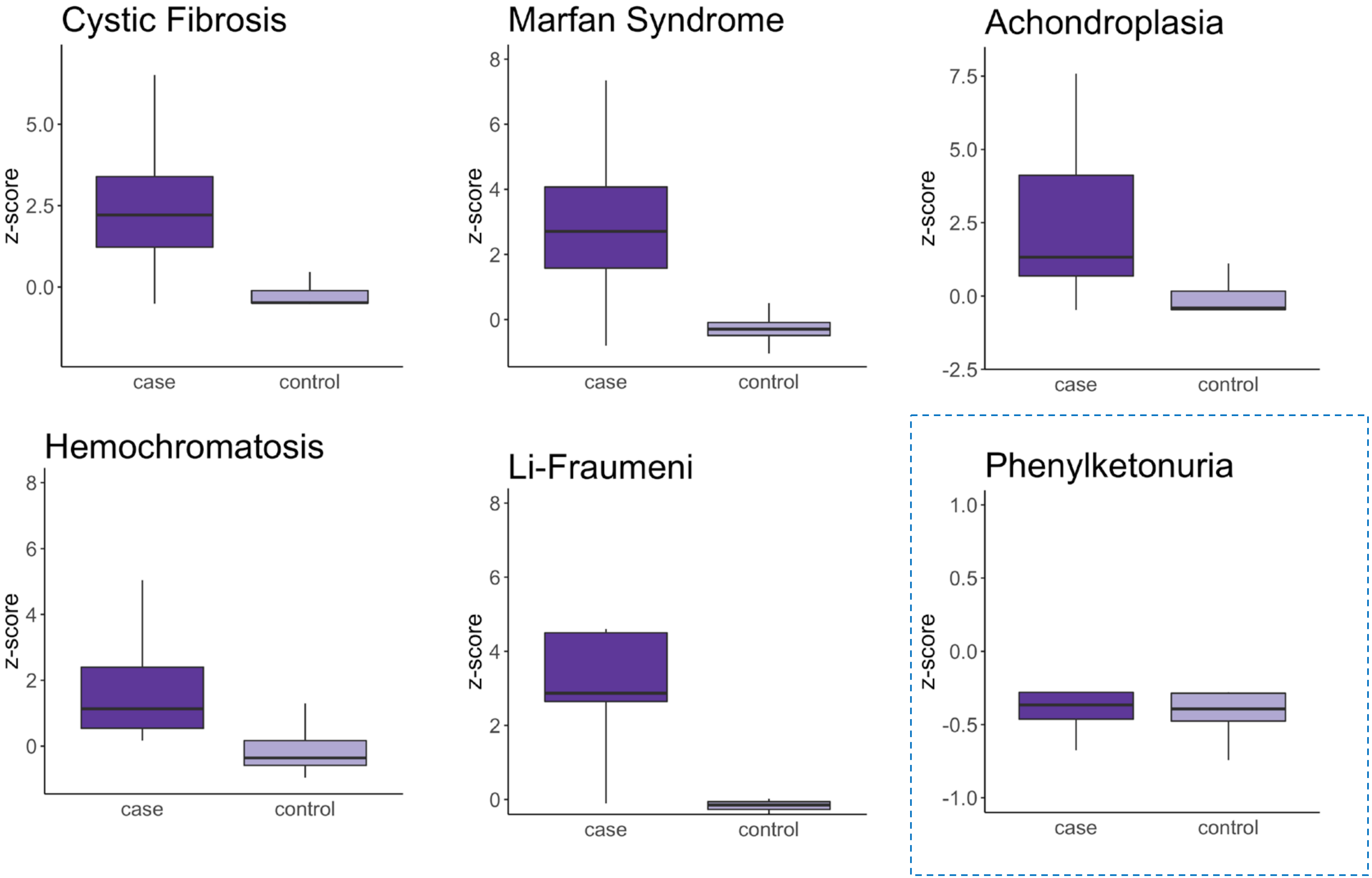
Do diagnosed patients have higher phenotype risk scores?



Diagnosed cases

Controls

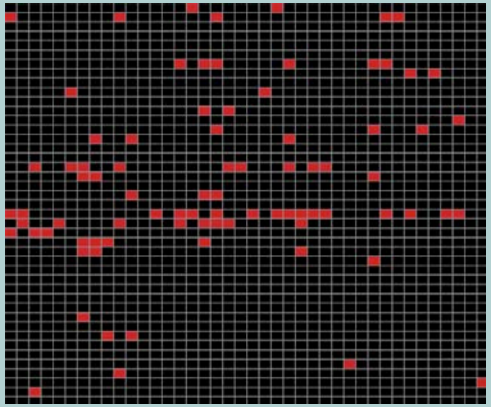
*You can differentiate a group individuals diagnosed with a disease using **only the features*** of the disease


Cystic Fibrosis
z-score
5.0
2.5
0.0
case
control
Marfan Syndrome
z-score
8
6
4
2
0
case
control
Achondroplasia
z-score
7.5
5.0
2.5
0.0
-2.5
case
control
Hemochromatosis
z-score
8
6
4
2
0
case
control
Li-Fraumeni
z-score
8
6
4
2
0
case
control
Phenylketonuria
z-score
1.0
0.5
0.0
-0.5
-1.0
case
control

Hypothesis free

~60,000 rare
Exonic
markers

X

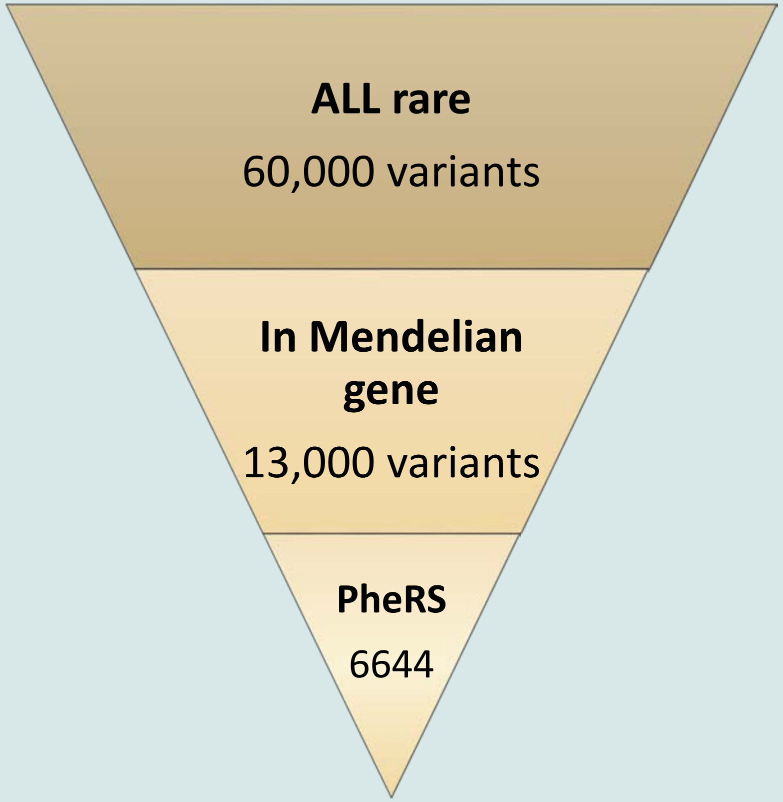


“Phenotype array” of
~1500 phenotypes

= ~90
million tests

Hypothesis driven

If a variant in gene X is linked to a phenotypic pattern, other variants in gene X will produce a similar pattern.



ALL rare

60,000 variants

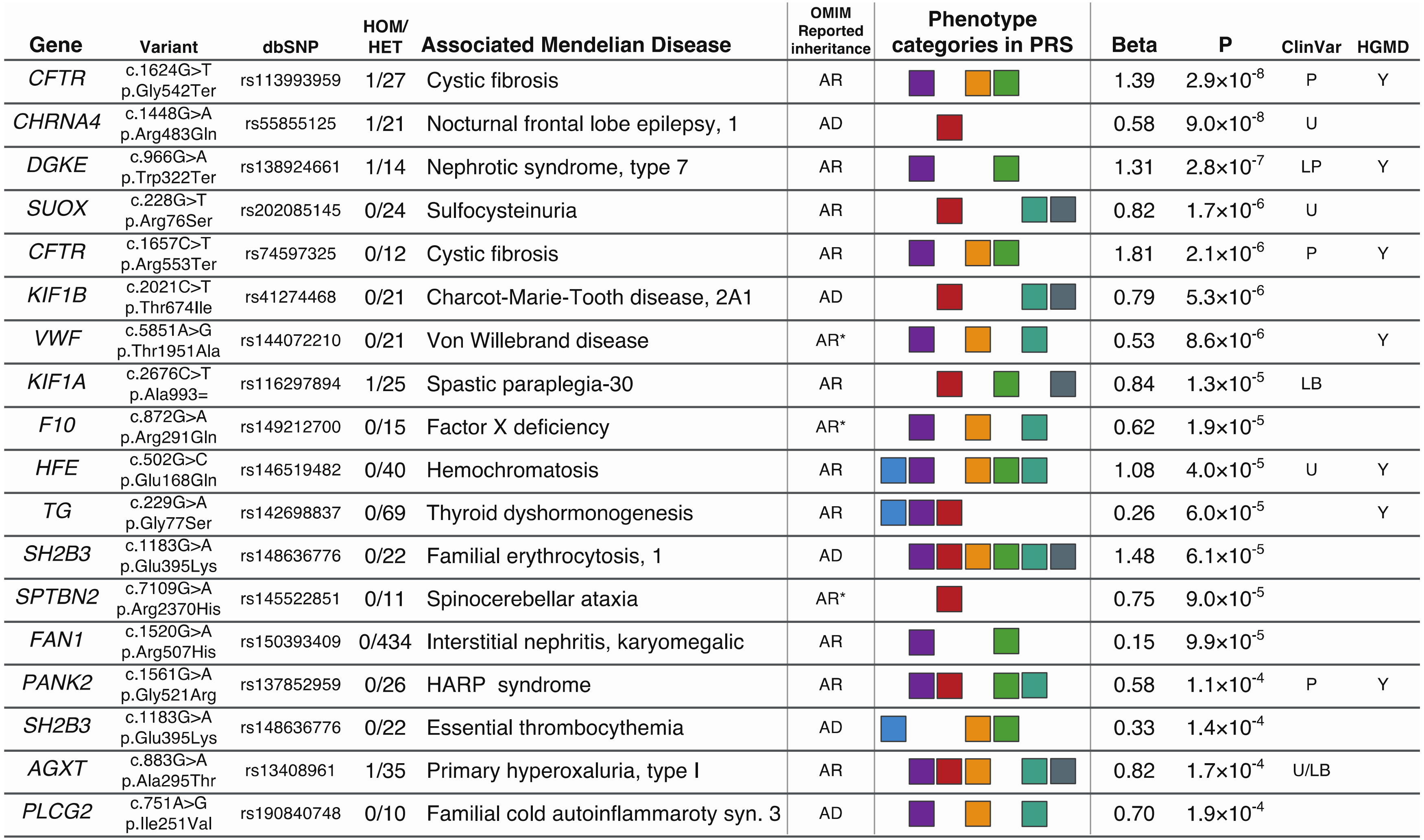
In Mendelian gene

13,000 variants

PheRS

6644

Application #1: Variant interpretation



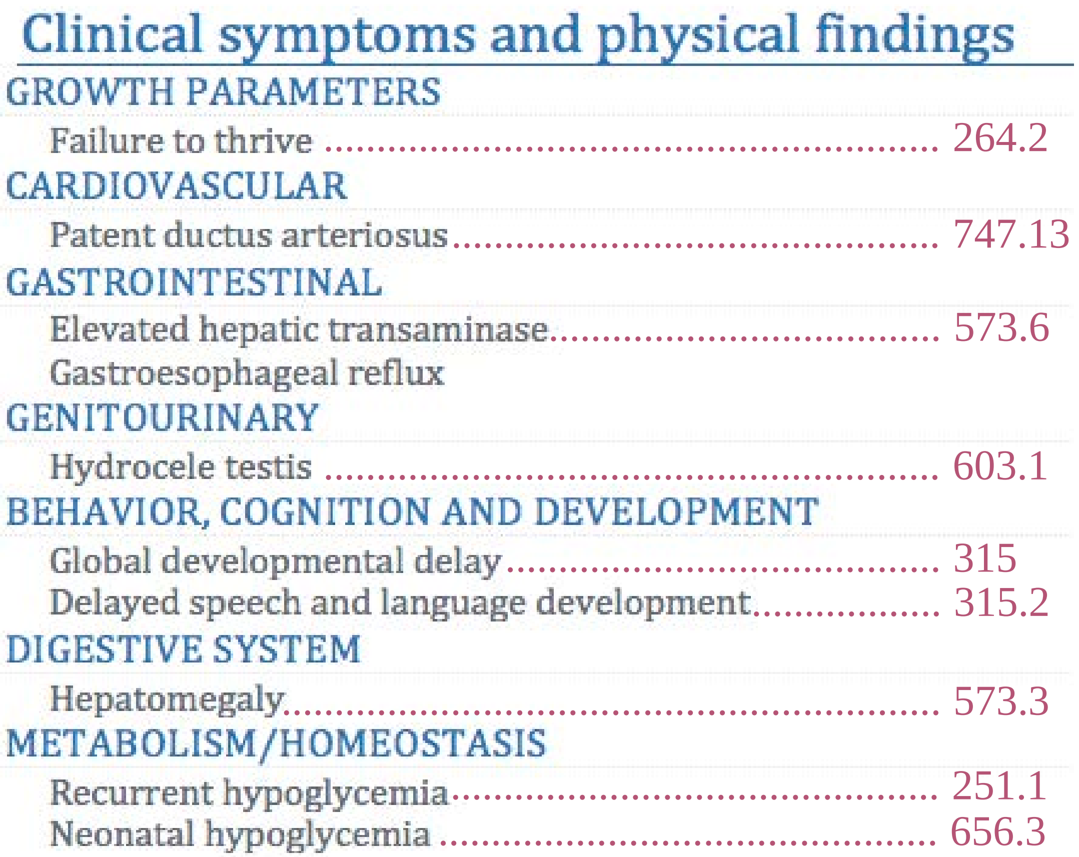
Gene	Variant	dbSNP	HOM/ HET	Associated Mendelian Disease	OMIM Reported inheritance	Phenotype categories in PRS	Beta	P	ClinVar	HGMD
CFTR	c.1624G>T p.Gly542Ter	rs113993959	1/27	Cystic fibrosis	AR	🟪 🟧 🟩	1.39	2.9×10^{-8}	P	Y
CHRNA4	c.1448G>A p.Arg483Gln	rs55855125	1/21	Nocturnal frontal lobe epilepsy, 1	AD	🟥	0.58	9.0×10^{-8}	U	
DGKE	c.966G>A p.Trp322Ter	rs138924661	1/14	Nephrotic syndrome, type 7	AR	🟪 🟩	1.31	2.8×10^{-7}	LP	Y
SUOX	c.228G>T p.Arg76Ser	rs202085145	0/24	Sulfocysteinuria	AR	🟥 ▣ ⬛	0.82	1.7×10^{-6}	U	
CFTR	c.1657C>T p.Arg553Ter	rs74597325	0/12	Cystic fibrosis	AR	🟪 🟧 🟩	1.81	2.1×10^{-6}	P	Y
KIF1B	c.2021C>T p.Thr674Ile	rs41274468	0/21	Charcot-Marie-Tooth disease, 2A1	AD	🟥 ▣ ⬛	0.79	5.3×10^{-6}		
VWF	c.5851A>G p.Thr1951Ala	rs144072210	0/21	Von Willebrand disease	AR*	🟪 🟧 ▣	0.53	8.6×10^{-6}		Y
KIF1A	c.2676C>T p.Ala993=	rs116297894	1/25	Spastic paraplegia-30	AR	🟥 🟩 ⬛	0.84	1.3×10^{-5}	LB	
F10	c.872G>A p.Arg291Gln	rs149212700	0/15	Factor X deficiency	AR*	🟪 🟧 ▣	0.62	1.9×10^{-5}		
HFE	c.502G>C p.Glu168Gln	rs146519482	0/40	Hemochromatosis	AR	🟦 🟪 🟧 🟩 ▣	1.08	4.0×10^{-5}	U	Y
TG	c.229G>A p.Gly77Ser	rs142698837	0/69	Thyroid dyshormonogenesis	AR	🟦 🟪 🟥	0.26	6.0×10^{-5}		Y
SH2B3	c.1183G>A p.Glu395Lys	rs148636776	0/22	Familial erythrocytosis, 1	AD	🟪 🟥 🟧 🟩 ▣ ⬛	1.48	6.1×10^{-5}		
SPTBN2	c.7109G>A p.Arg2370His	rs145522851	0/11	Spinocerebellar ataxia	AR*	🟥	0.75	9.0×10^{-5}		
FAN1	c.1520G>A p.Arg507His	rs150393409	0/434	Interstitial nephritis, karyomegalic	AR	🟪 🟩	0.15	9.9×10^{-5}		
PANK2	c.1561G>A p.Gly521Arg	rs137852959	0/26	HARP syndrome	AR	🟪 🟥 🟩 ▣	0.58	1.1×10^{-4}	P	Y
SH2B3	c.1183G>A p.Glu395Lys	rs148636776	0/22	Essential thrombocythemia	AD	🟦 🟧 🟩	0.33	1.4×10^{-4}		
AGXT	c.883G>A p.Ala295Thr	rs13408961	1/35	Primary hyperoxaluria, type I	AR	🟪 🟥 🟧 ▣ ⬛	0.82	1.7×10^{-4}	U/LB	
PLCG2	c.751A>G p.Ile251Val	rs190840748	0/10	Familial cold autoinflammaroty syn. 3	AD	🟪 🟧 ▣	0.70	1.9×10^{-4}		

- 🟦 Neoplastic
- 🟪 Endocrine, metabolic/Blood
- 🟥 Nervous/Psychiatric/Sensory
- 🟧 Circulatory/Respiratory
- 🟩 Digestive/Genitourinary
- ▣ Musculoskeletal/Dermatologic
- ⬛ Other symptoms/Injuries

Application #2: WES interpretation

Undiagnosed Diseases Network

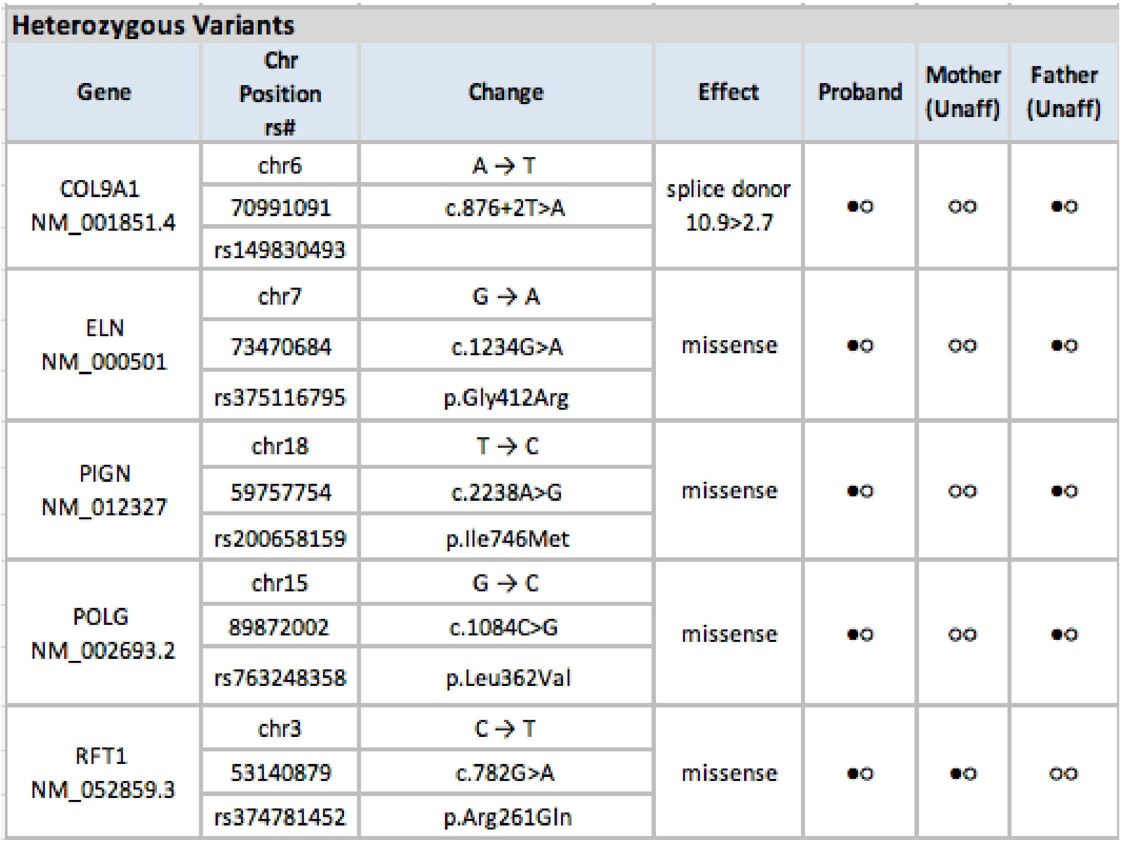
Proband phenotype



Clinical symptoms and physical findings

GROWTH PARAMETERS	
Failure to thrive	264.2
CARDIOVASCULAR	
Patent ductus arteriosus	747.13
GASTROINTESTINAL	
Elevated hepatic transaminase	573.6
Gastroesophageal reflux	
GENITOURINARY	
Hydrocele testis	603.1
BEHAVIOR, COGNITION AND DEVELOPMENT	
Global developmental delay	315
Delayed speech and language development	315.2
DIGESTIVE SYSTEM	
Hepatomegaly	573.3
METABOLISM/HOMEOSTASIS	
Recurrent hypoglycemia	251.1
Neonatal hypoglycemia	656.3

Candidate variants

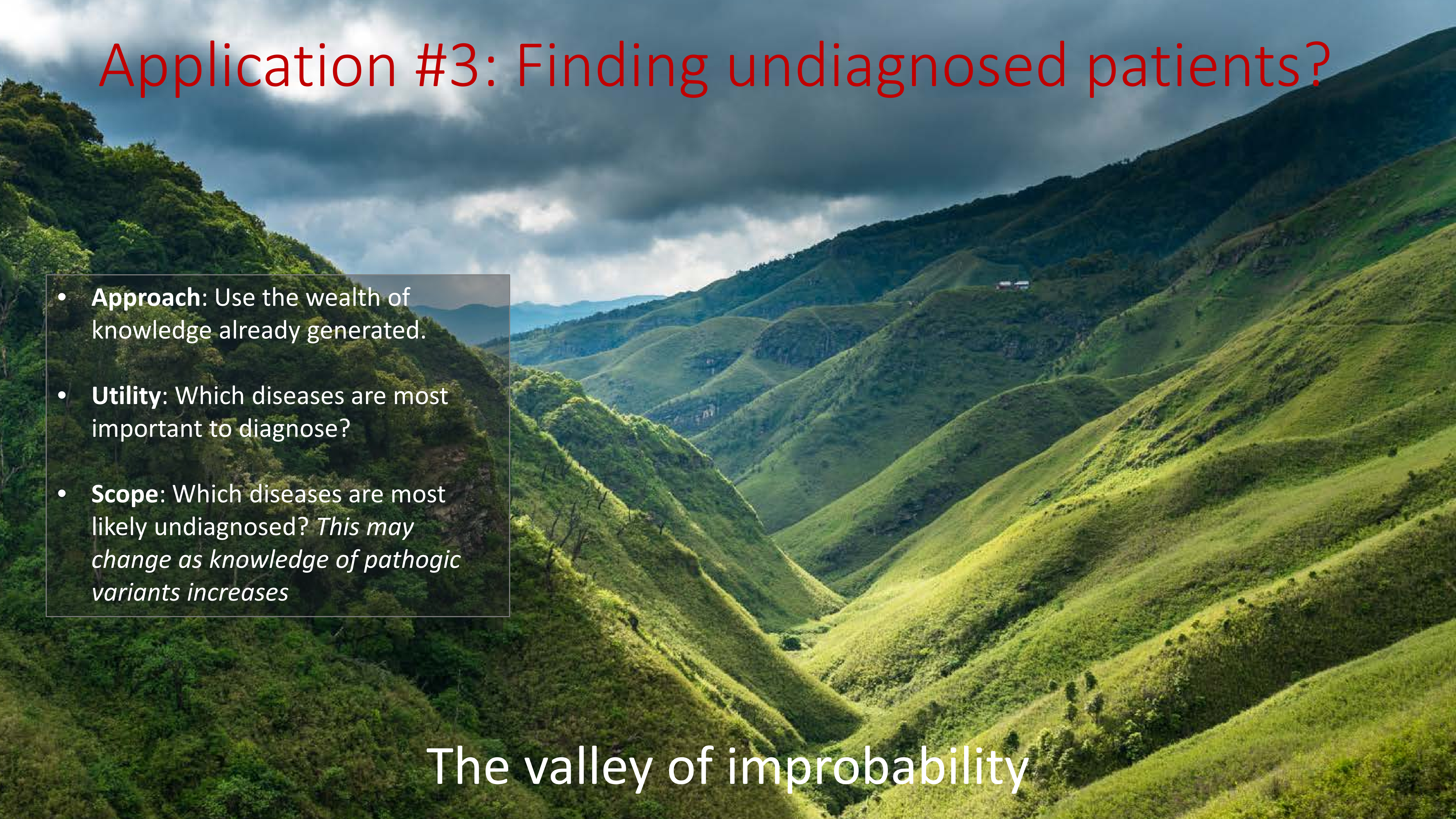


Heterozygous Variants

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

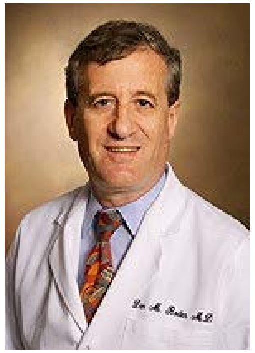
Application #3: Finding undiagnosed patients?

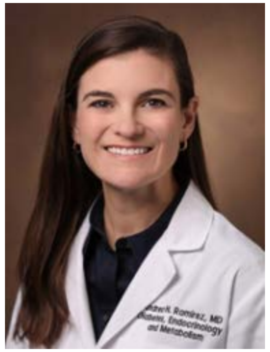
- **Approach**: Use the wealth of knowledge already generated.
- **Utility**: Which diseases are most important to diagnose?
- **Scope**: Which diseases are most likely undiagnosed? *This may change as knowledge of pathogic variants increases*



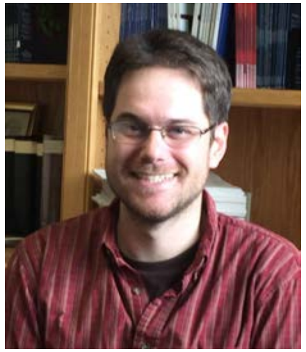
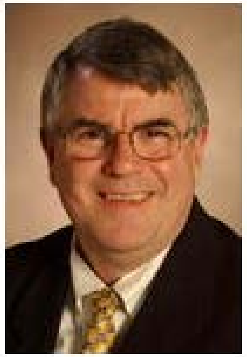
The valley of improbability

Acknowledgements





VANDERBILT UNIVERSITY
MEDICAL CENTER

udn
Undiagnosed
Diseases Network

emerge
network