Exome sequencing for drug discovery

Jonathan Marchini

Head of Statistical Genetics and Machine Learning
The Regeneron Genetics Center

Genetics and Drug Discovery

90% OF EXPERIMENTAL MEDICINES FAIL, with costs exceeding $2B per approved drug

>$100B SPENT on worldwide R&D by biopharma industry → only 10–20 new drugs per year

HIGHER PROBABILITY OF SUCCESS for drugs with human genetics evidence

YOU CANNOT PURSUE MODERN DRUG DISCOVERY & DEVELOPMENT WITHOUT INCORPORATING HUMAN GENETICS

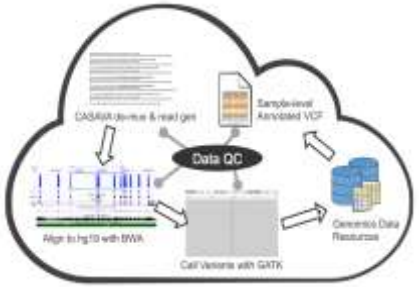
The Regeneron Genetics Center (RGC)

Automated Biobank (1.4M Samples)

Library Prep Automation (>500,000 Samples/Yr)

Illumina Fleet (>500,000 Exomes/Yr)

Cloud Based Informatics & Analysis



- **Key Technologies and Capabilities**
  - Automated biobank with 1.4M+ sample capacity
  - Custom fully-automated sequencing and genotyping sample preparation workflows
  - More than 600,000 exomes completed with a goal of >3,000,000 in the next few yea
  - First “genome center in the cloud” with fully automated analysis pipelines

REGENERON® GENETICS CENTER

June 2019

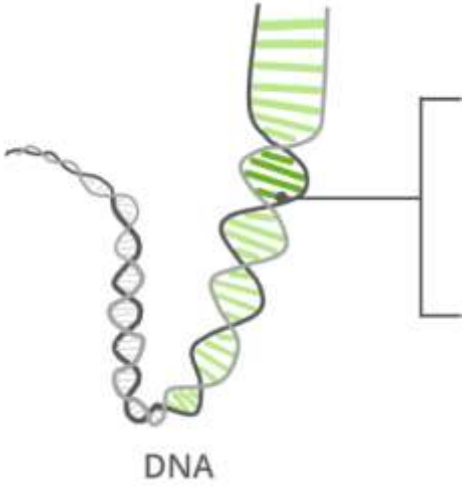
- >600,000 exome sequenced samples
- >670,000 arrays

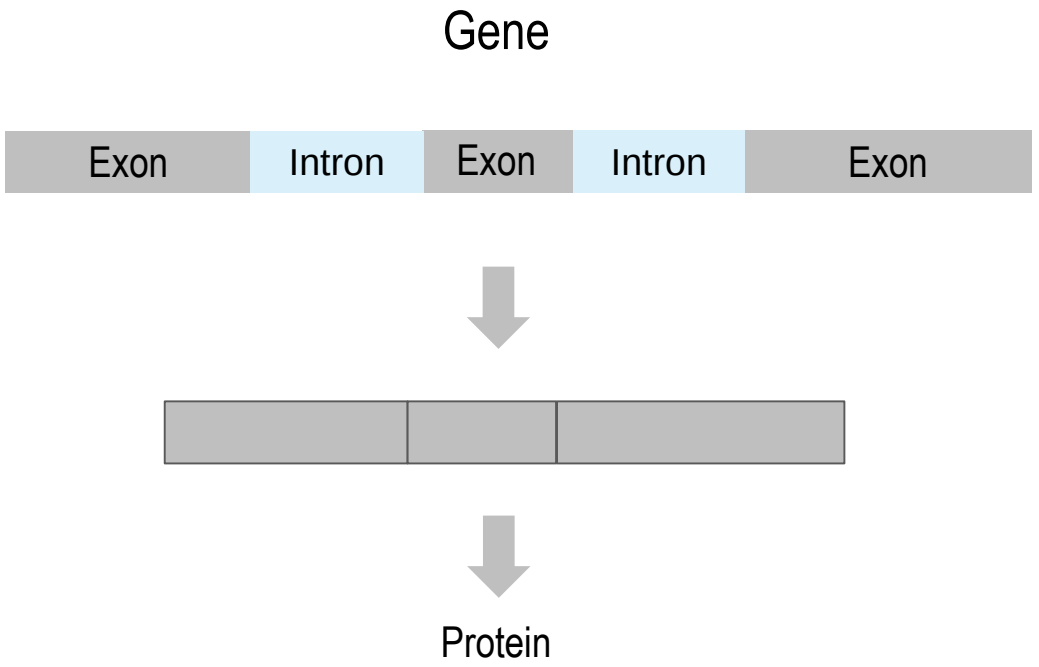
70+ collaborative projects

>100,000 datasets

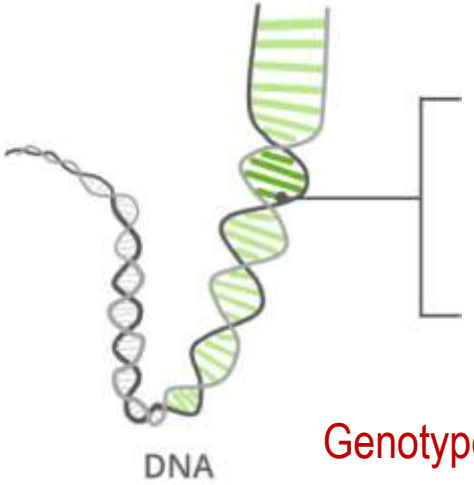
- **cohort/ancestry/exome/chip/imputation/analysis**

Exons are the 1-2% of the genome that encode proteins





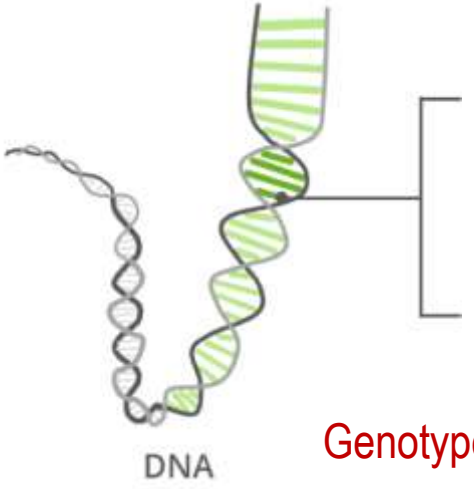
Exons are the 1-2% of the genome that encode proteins



Gene

Genotyped sites – Not many. Common variants. Hard to genotype rare variants.

Exons are the 1-2% of the genome that encode proteins

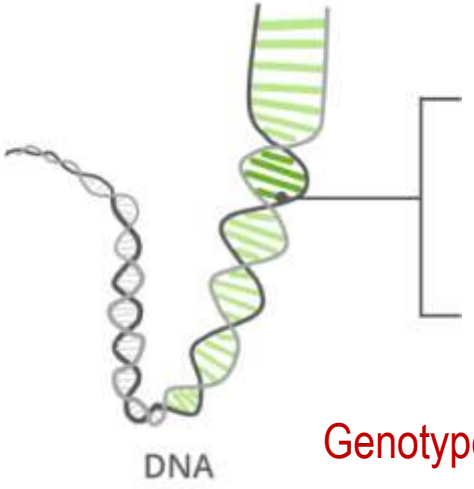


Gene

Genotyped sites – Not many. Common variants. Hard to genotype rare variants.

Imputed sites – 100 fold more sites. Dependent upon reference panel size/quality.

Exons are the 1-2% of the genome that encode proteins



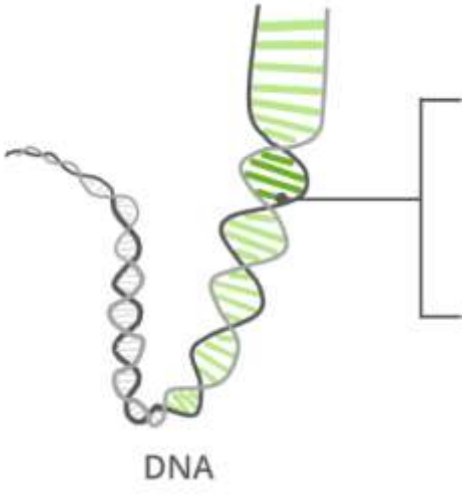
Gene

Genotyped sites – Not many. Common variants. Hard to genotype rare variants.

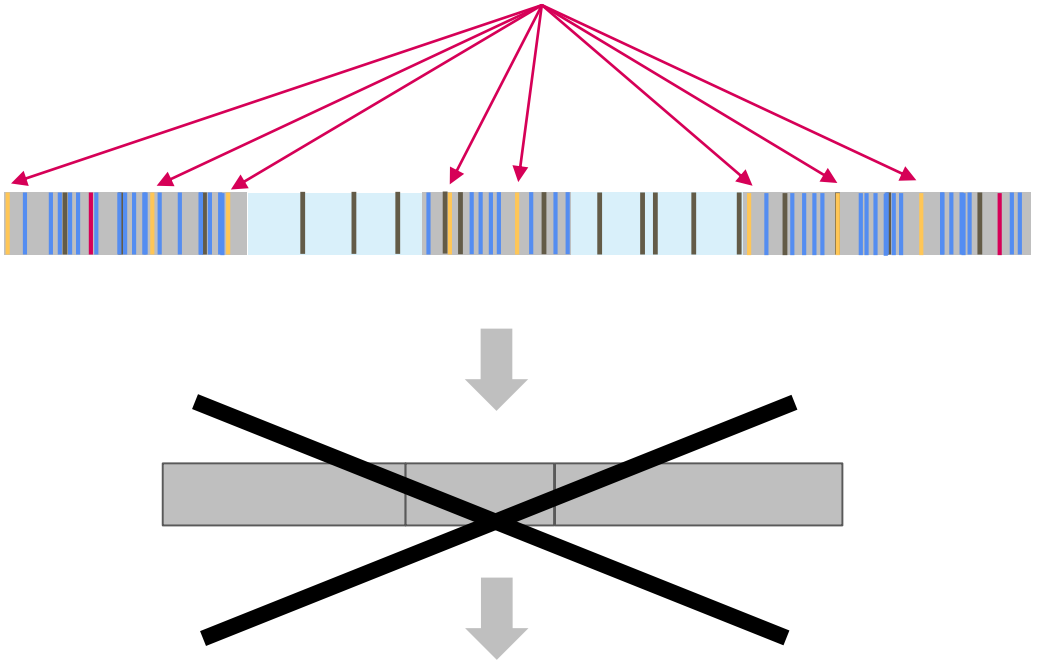
Imputed sites – 100 fold more sites. Dependent upon reference panel size/quality.

Exome Sequenced sites – ~10 fold more sites in Exons. Good quality calls at rare sites.

Exome sequencing

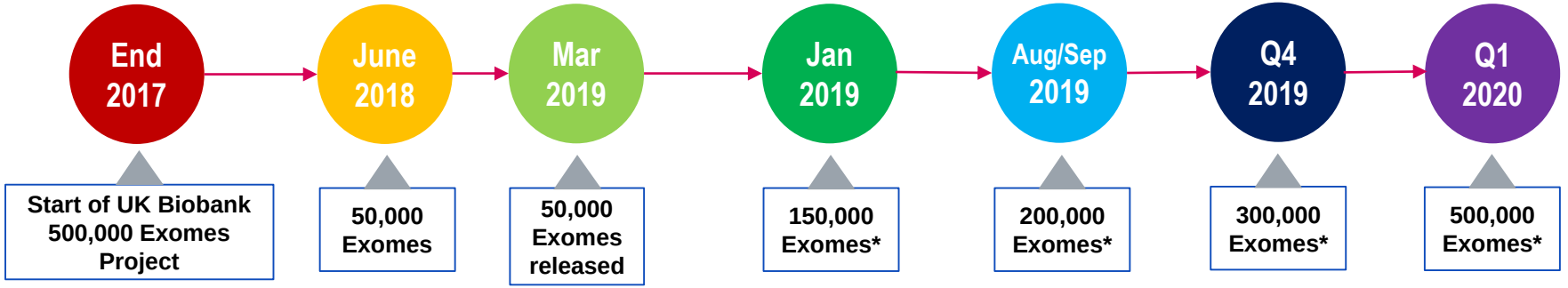


Identify many more Loss of Function (LOF) mutations



Sequencing is expensive so focus on the Exome first

UK Biobank 500,000 Exomes Project : Timelines



biobank uk
Improving the health of future generations

REGENERON GENETICS CENTER

biobank uk
Improving the health of future generations

REGENERON GENETICS CENTER

abbvie

Alnylam PHARMACEUTICALS

AstraZeneca

Biogen

Bristol-Myers Squibb

Pfizer

Takeda

gsk GlaxoSmithKline

*Release of data via UK Biobank will occur 12 months after these dates

50K dataset variant discovery

~4.7 million coding variants

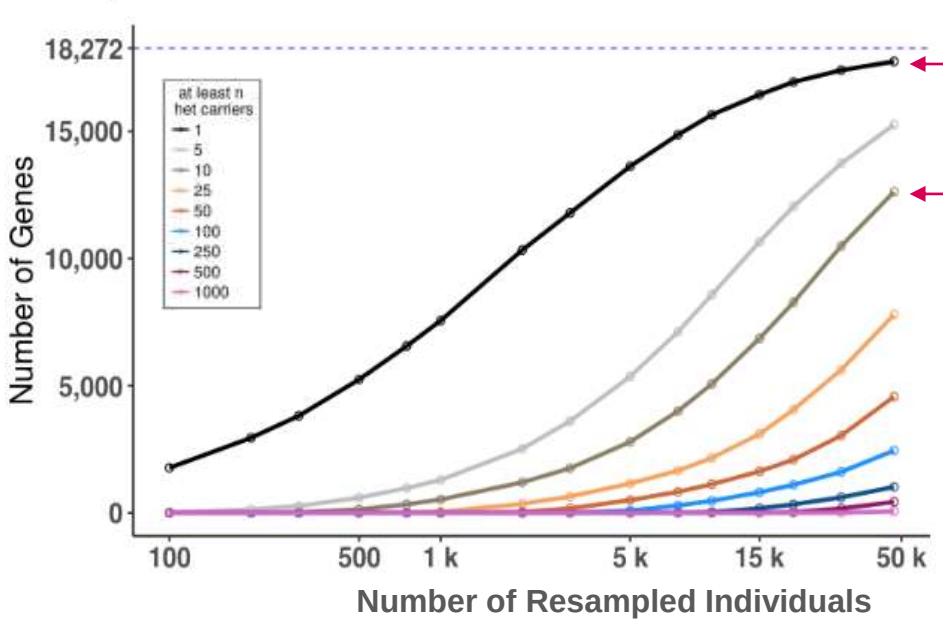
- ~98.4% have minor allele frequency (MAF) < 1%

Each individual

- Median #variants = 24,332
- Median #LOFs = 219
- Median #LOFs (MAF < 1%) = 24

	49,960 Participants	
	# Variants	MAF<1%
Total	9,693,526	9,547,730
Targeted Regions	4,735,722	4,665,684
SNVs	4,520,754	4,453,941
Indels	214,968	211,743
Multi-Allelic	591,340	580,728
Synonymous	1,229,303	1,203,043
Missense	2,498,947	2,472,384
LOF (any transcript)	231,631	230,790

LOF predictions



gnomAD

17,751 genes (97%)
>=1 LOF in 50k

17,946 genes
>=1 LOF in ~57k

12,629 genes (69%)
>=10 LOF in 50k

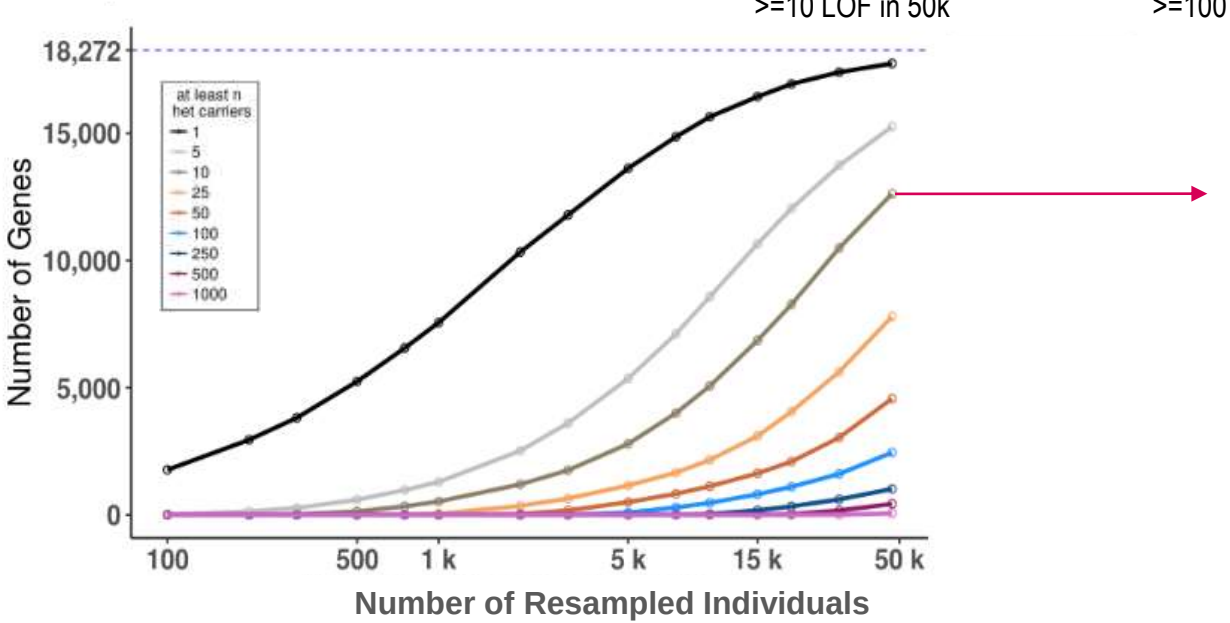
LOF definition

- SNPs + indels
- stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost and frameshift
- count losses in any transcript
- MAF <1%

LOF predictions

12,629 genes
>=10 LOF in 50k

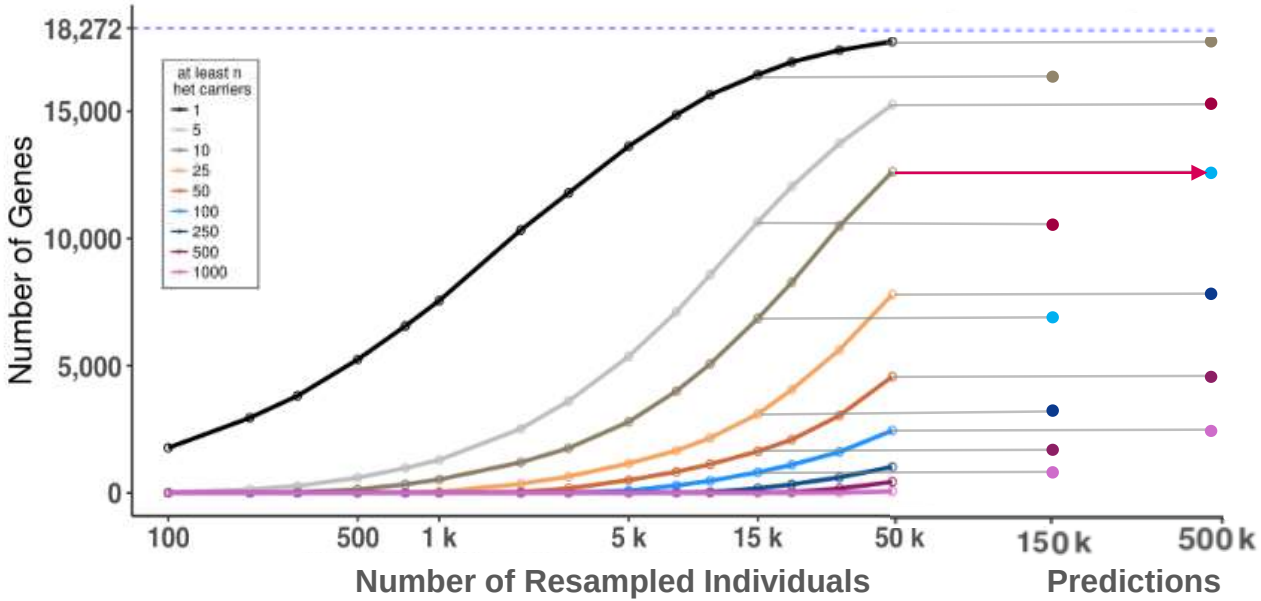
12,629 genes
>=100 LOF in 500k



LOF definition

- SNPs + indels
- stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost and frameshift
- count losses in any transcript
- MAF < 1%

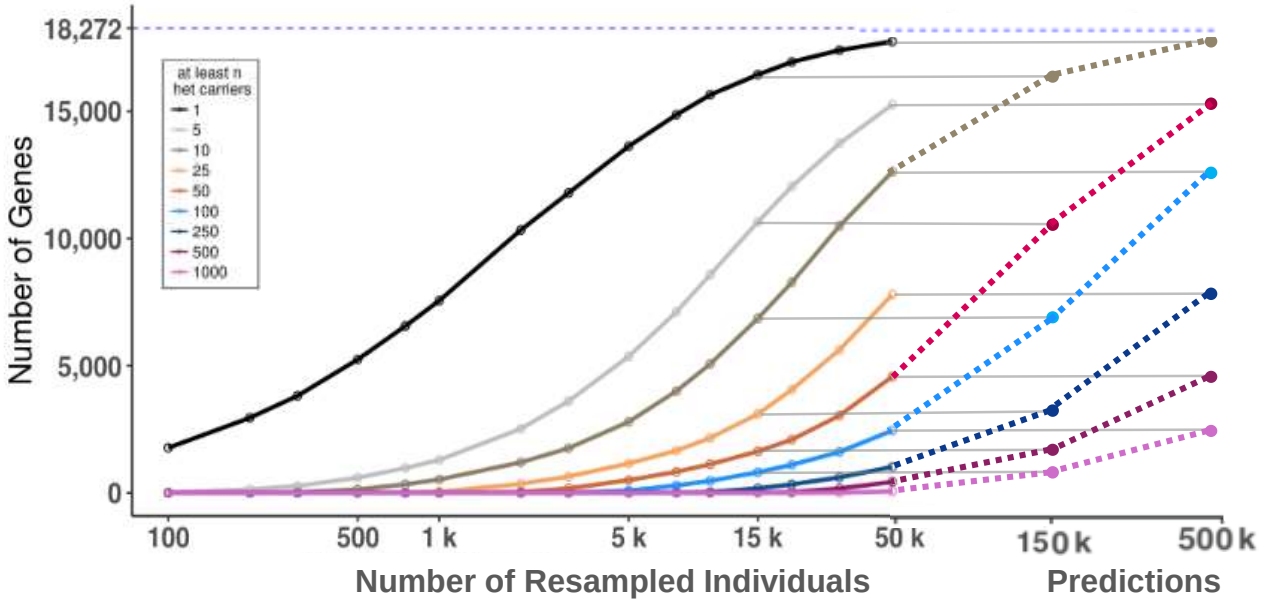
LOF predictions



LOF definition

- SNPs + indels
- stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost and frameshift
- count losses in any transcript
- MAF < 1%

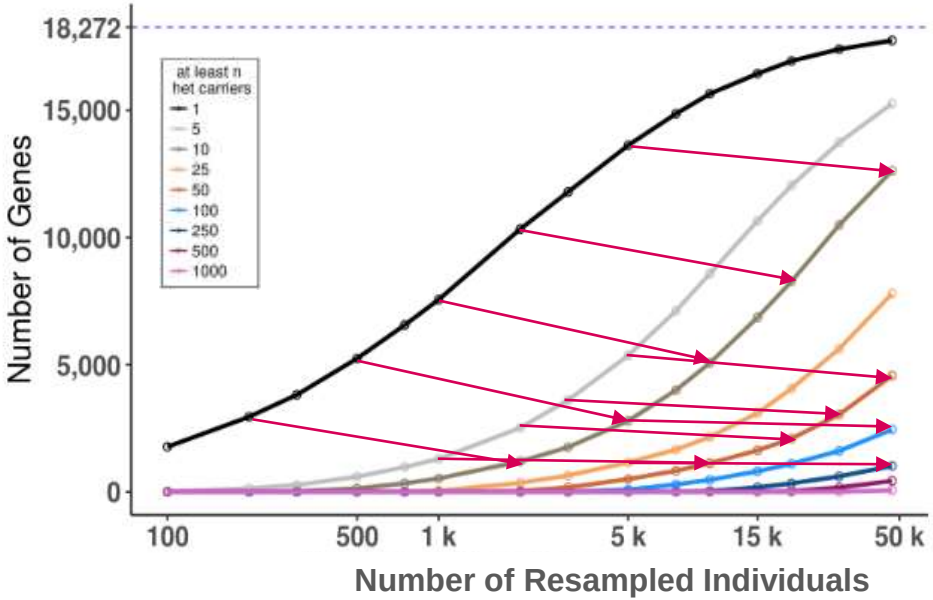
LOF predictions



LOF definition

- SNPs + indels
- stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost and frameshift
- count losses in any transcript
- MAF < 1%

LOF predictions - CAVEAT



Interpolation suggests that extrapolation likely over-estimates

Exome vs Imputed data

Only 13.7% of Exome variants exist in the UKB imputed data

LOFs

~12 fold increase Exome vs Imputed

	AAF	WES	Imputed 50k	Both
LOF	All	235,915	19,451	12,488
	<1%	234,716	18,162	11,639
	≥1%	1,199	1,289	849

Missense variants

~6 fold increase Exome vs Imputed

	AAF	WES	Imputed 50k	Both
Missense	All	2,518,075	420,194	317,623
	<1%	2,485,710	389,343	288,599
	≥1%	32,365	30,851	29,024

Medically actionable results

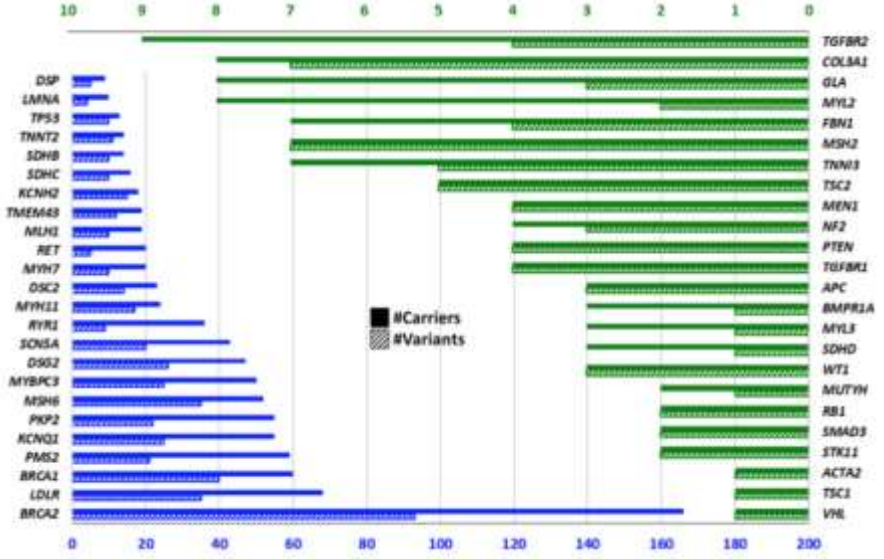
2% of UKB 50k carry medically actionable variants
in American College of Medical Genetics ACMG59 genes

Variants were observed in 48 genes

Most common ACMG59 pathogenic or likely pathogenic variant carriers:

- *BRCA2,* 93 variants 166 carriers
- *LDLR,* 35 variants 65 carriers
- *BRCA1,* 40 variants 60 carriers
- *PMS2*, 21 variants 59 carriers
- *KCNQ1*, 21 variants 55 carriers

Category	#Variants	% of Total Known ACMG59 Variants	#Carriers	% of individuals with reportable variants
Pathogenic (P)	316	4.23	694	1.39
Likely Pathogenic (LP)	239	-	315	0.63
P + LP	555	-	1,009	2.0[1]



[1]Percent of individuals with P or LP variants is not additive, as the 2.0% represents non-redundant carriers; 9 individuals were found to have 2 medically actionable variants.

Positive control LOF burden results in 50k/150k dataset

Gene	Binary Phenotype	UKB 50k exome		UKB 150k exome	
		OR (95% CI)	Burden P	OR (95% CI)	Burden P
MLH1	Personal history of malignant neoplasm of digestive organs	84 (31,230)	$3.5x10^{-11}$	28 (12,70)	$1.7x10^{-12}$
PKD1	Chronic kidney disease (CKD)	91 (36,229)	$2.9x10^{-9}$	32 (14,71)	$2.5x10^{-16}$
CALR	Other neoplasms of uncertain behavior of lymphoid, hematopoietic and related tissue	866 (194,3857)	$4.1x10^{-8}$	280 (57,1381)	$1.6x10^{-11}$
TTN	Cardiomyopathy	12 (6.5,22)	$1.4x10^{-8}$	6.7 (4.0,11)	$5.5x10^{-12}$

Gene	Quantitative Phenotype	UKB 50k exome		UKB 150k exome	
		Beta (95% CI)	Burden P	Beta (95% CI)	Burden P
*TUBB1**	Platelet distribution width	1.84 (1.48,2.20)	$2.5x10^{-23}$	1.99 (1.81,2.16)	$2.3x10^{-111}$
KALRN	Mean platelet thrombocyte volume	-0.62 (-0.74,-0.49)	$2.7x10^{-23}$	-0.52 (-0.59,-0.45)	$1.1x10^{-47}$
*KLF1**	Red blood cell erythrocyte distribution width	1.45 (1.06,1.83)	$1.5x10^{-13}$	1.39 (1.18,1.60)	$3.2x10^{-38}$
*IL33**	Eosinophill percentage	-0.31 (-0.40,-0.22)	$5.4x10^{-12}$	-0.27 (-0.32,-0.23)	$1.1x10^{-30}$
*ASXL1**	Red blood cell erythrocyte distribution width	0.62 (0.43,0.80)	$2.4x10^{-11}$	0.47 (0.36,0.57)	$2.2x10^{-19}$
*HBB**	Red blood cell erythrocyte count	2.96 (2.00,3.93)	$1.7x10^{-9}$	2.30 (1.87,2.73)	$1.10x10^{-25}$
GP1BA	Mean platelet thrombocyte volume	0.51 (0.32,0.69)	$6.4x10^{-8}$	0.55 (0.44,0.66)	$8.8x10^{-24}$
CHEK2	Platelet crit	0.30 (0.19,0.41)	$7.9x10^{-8}$	0.24 (0.17,0.30)	$1.0x10^{-13}$

*Multiple phenotype associations at p<1e-7
Betas measured in standard deviations

50k BT analysis included covariates: age, sex, PC1-4
50k QT analysis included covariates: age, sex, PC1-4, sample collection site

Novel LOF burden results in 50k/150k dataset

Gene	Binary Phenotype	UKB 50k exome Counts	UKB 50k exome OR (95% CI)	UKB 50k exome Burden P	UKB 150k exome Counts	UKB 150k exome OR (95% CI)	UKB 150k exome Burden P
PIEZO1	Asymptomatic varicose veins of lower extremities	Ctrls:43285|142|0 Cases:1267|20|0	4.9 (3.1,7.8)	$2.7x10^{-8}$	Ctrls:131514|443|0 Cases:3559|36|0	3.0 (2.1,4.4)	$1.8x10^{-8}$

Gene	Quantitative Phenotype	UKB 50k exome Counts	UKB 50k exome Beta (95% CI)	UKB 50k exome Burden P	UKB 150k exome Counts	UKB 150k exome Beta (95% CI)	UKB 150k exome Burden P
MEPE	BMD t-score	42836|149|0	-0.47(-0.63,-0.31)	$5.4x10^{-9}$	61510|232|0	-0.43 (-0.55,-0.31)	$8.9x10^{-12}$
*COL6A1**	Corneal resistance factor mean	35621|17|0	-1.50 (-1.96,-1.03)	$3.6x10^{-10}$	50007|30|0	-1.50 (-1.85,-1.15)	$3.8x10^{-17}$
IQGAP2	Mean platelet thrombocyte volume	45443|165|0	0.68 (0.53,0.82)	$1.1x10^{-19}$	138007|466|0	0.69 (0.61,0.78)	$3.6x10^{-60}$
GMPR	Mean corpuscular haemoglobin	45195|182|0	0.41 (0.27,0.55)	$1.1x10^{-8}$	137928|545|1	0.41 (0.33,0.48)	$3.6x10^{-24}$

*Multiple phenotype associations at $p \leq 1x10^{-7}$
Betas measured in standard deviations
50k binary trait analysis included covariates: age, sex, PC1-4
50k quantitative trait analysis included covariates: age, sex, PC1-4, sample collection site

Acknowledgements

Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank https://www.biorxiv.org/content/10.1101/572347v1

UK Biobank participants, organizers, researchers

UKB 50k exomes phase 1 and manuscript team

UKB-Pharma Consortium members