

**Machine Learning for Healthcare**

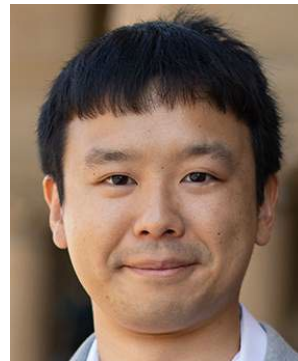
6.7930 [6.871], HST.956

**Lecture 19: Genetics**  
**Risk Prediction with Polygenic Risk Scores**

Prof. Manolis Kellis

Slides credit:

**Yosuke Tanigawa**



# Potential of PRS in clinical practice

---

## AHA SCIENTIFIC STATEMENT

---

### Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association

Jack W. O'Sullivan, MBBS, DPhil, Chair; Sridharan Raghavan, MD, PhD; Carla Marquez-Luna, PhD; Jasmine A. Luzum, PharmD, PhD; Scott M. Damrauer, MD, FAHA; Euan A. Ashley, MBChB, DPhil, FAHA; Christopher J. O'Donnell, MD, MPH; Cristen J. Willer, DPhil; Pradeep Natarajan, MD, MMSc, Vice Chair; on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease

*“These observations point to the **possibility of using genetic profiling to inform clinical practice** in significantly larger groups of individuals than for whom monogenic cardiovascular variants are considered. As a result of exponential increases in the proportion of individuals with broad genetic profiling, **cardiovascular PRSs are beginning to enter clinical practice**. Such PRSs may be appropriately considered in select scenarios, given the current evidence base.”*

# Potential relevance of PRS in clinical practice

Example: coronary artery disease

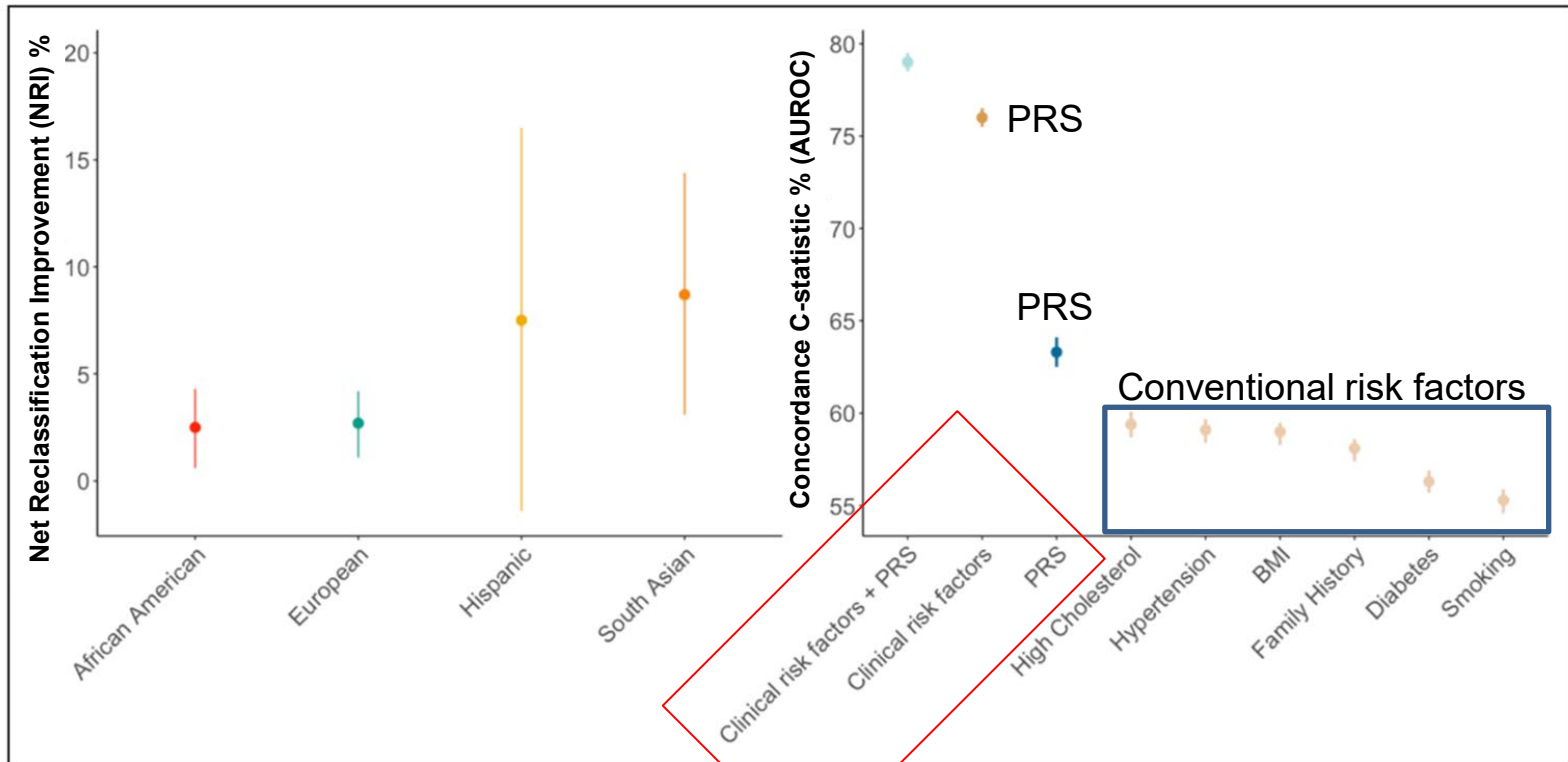


Figure 3. Predictive ability of polygenic risk scores for coronary artery disease.

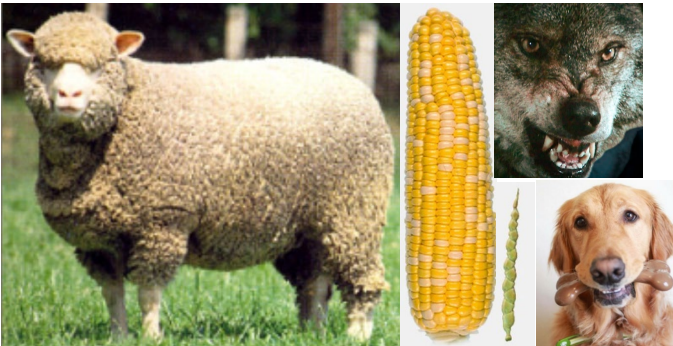
- PRS has higher risk stratification ability than conventional risk factors
- PRS & conventional risk factors leads to improvement

# Overview: Genetic prediction of complex traits

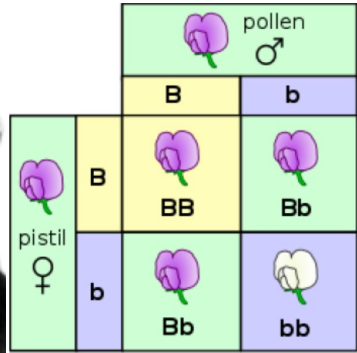
---

1. Foundations of Human Genetic Variation
2. Polygenic score (PGS) introduction
3. PGS Evaluation
4. Methods to fit PGS model
5. Challenges and opportunities in PGS research

# Genetics: Ancient Foreshadowings → Mendelian traits → Polygenicity



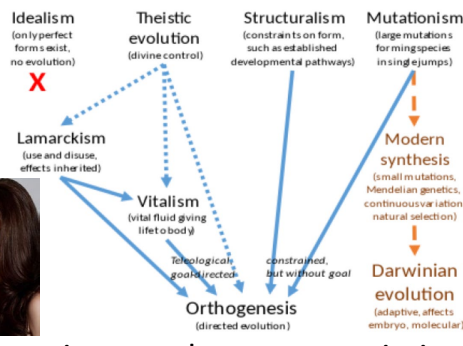
**9000BC:** Selective breeding of animals/plants  
**Inheritance:** Eye/hair color long understood



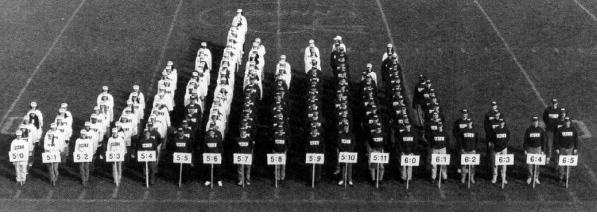
**1866: Mendel:** Discrete inheritance  
 No blending. Dominant/recessive alleles  
 Independent assortment



**Biometrics:** continuous phenotype variation.  
**Others:** Saltationism, orthogenesis, vitalism, neo-Lamarckism, theistic evolution...



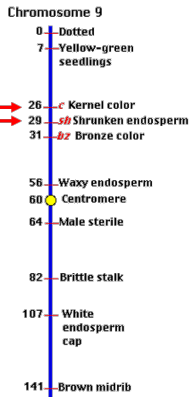
**Fisher**



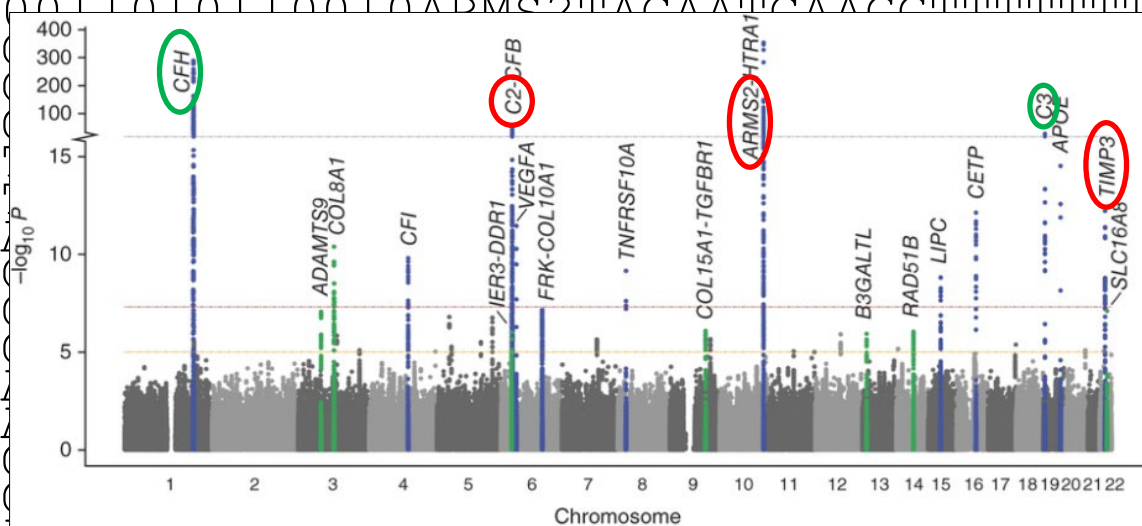
**1918:** Continuous phenotype variation explained by multiple Mendelian loci



**1913:** Linkage/mapping, Morgan, Sturtevant  
**1980s:** Mendelian Trait genes mapped



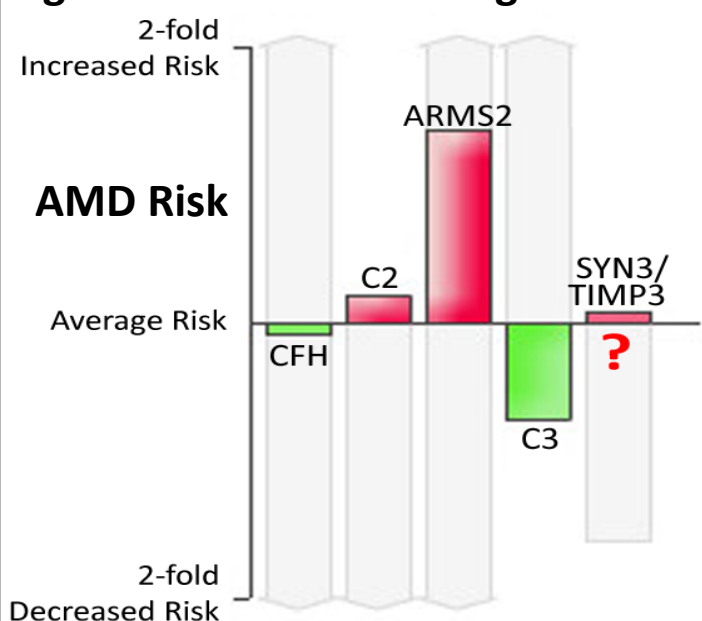
**2000s:** Human genome. Variation maps. Haplotypes. GWAS. Common/rare variants.



AAAAGGAAACAAGAAGACGCAGIAGGTCIGAGAAAG  
 GCTTATAGAAGGCCATCTGAGTGGCCCTCAAGCCGGTGAATTGGCTTTAGGGTTTACTG  
 AAGGAGGTGGAAACCTCAGCCTGCTTCTCGTCCGGGTGTTAGAGGAGTCATTTAGAAAN  
 NTIMP3AACATATATATTTTTTCA

TACTAGGGGACCTCTGTGTGCCTCCT  
 GACCACCCAACAATTCAGGGGTGGA  
 TGTGACGGGAAAAGACAATGCTCC  
 CAGCACCTTTGTCCACCACATATG  
 CGATGGTAACTGAGGCGGAGGGGA  
 TGGTTCCTGTGTCTTTCATTTCCA  
 GACAAGGAGCCAGTGACAAGCAGA  
 CTCTATAAATCCCACTGAGCTCT  
 GGAGCAAGCAGCCTCAGCACCC  
 TACCCAGACCTATTGAATCAGAA  
 AGCCTTCAGGTGCTTCTGATGCAT  
 AGGCAATTCAGCCTTCTCTGGTT  
 CCAATGCACCTGCTACATGCCAGA  
 TGATGGGGTGAGCAGAAACCCAAA  
 GTGAATTGGCTTTAGGGTTTACTG  
 TTTAGAGGAGTCATTTAGAAAN  
 CAGTGGCAGGAAGTCTTGCCCGAGGTGGGAATGTTACTG  
 AATATTTTTTCCCTTTGTTAGCTGGCTCTGGGCAGCCT  
 TGCTGCTTGGGACCTAATGACCTGCTTTCAATCCCTT  
 AATTTGGAAAACAACCAAGGCTTGGATGG  
 TCTGTACCCAGTTTTTCAAAGAGATTTTTTTTTTTTCA  
 AGTCCTGGACCTTTGGCAGCAAAGGGTGGGACTTCTG  
 AGCTCAGCGGGGCCCTCCCGCTTGGATGTTCCGGGAA  
 GGCGAGCCGCAGGTGCCAGAACACAGATTGTATAAAA  
 GGGAAAGGGAAATGTGACCAGGTCTAGGTCTGGAGTTTC  
 ACAAGCAAAGCAAGCCAGGACACACCATCCTGCCCCA  
 CAACGCCATGGGGAGCAATCTCAGCCCCCAACTCTGC  
 CTTGTCTGGAGGTAAGCGAGGGGTAACCTTCCCTTCTC  
 GCCTTTTGGGCCAGGCTTCATCAGCCTTTTCTCTTCA  
 TTTGGCCCGGCCAGGGATCCTGCTCTCTGGAGGGG  
 CCGACTTCTCAAAGAGGGCCAGGCACTGGAGTACGTG  
 CCCTGTGCAGACACGTACCTGCAGATCTACGGGGTCC  
 CAAAAGACTGTCAGGAAGGCAGAGTGCAGAGGTTTG  
 CCTAAGGCAGAAACAGGGCAGGCGGCAGCAAGGTCAG  
 TGACAAGGTGGGCTGACCGGGAGTAGGAGCAGTTTTA  
 AGAAAAAGCGGAGTTAACCTTACTAAGCATTTACCC  
 GTCAAGAGAACACTCAGAAATGGGGAGGGAGAAGCAG  
 GTAGGTAAAGATGCTGCTTCTGCGGGGACTG00110101

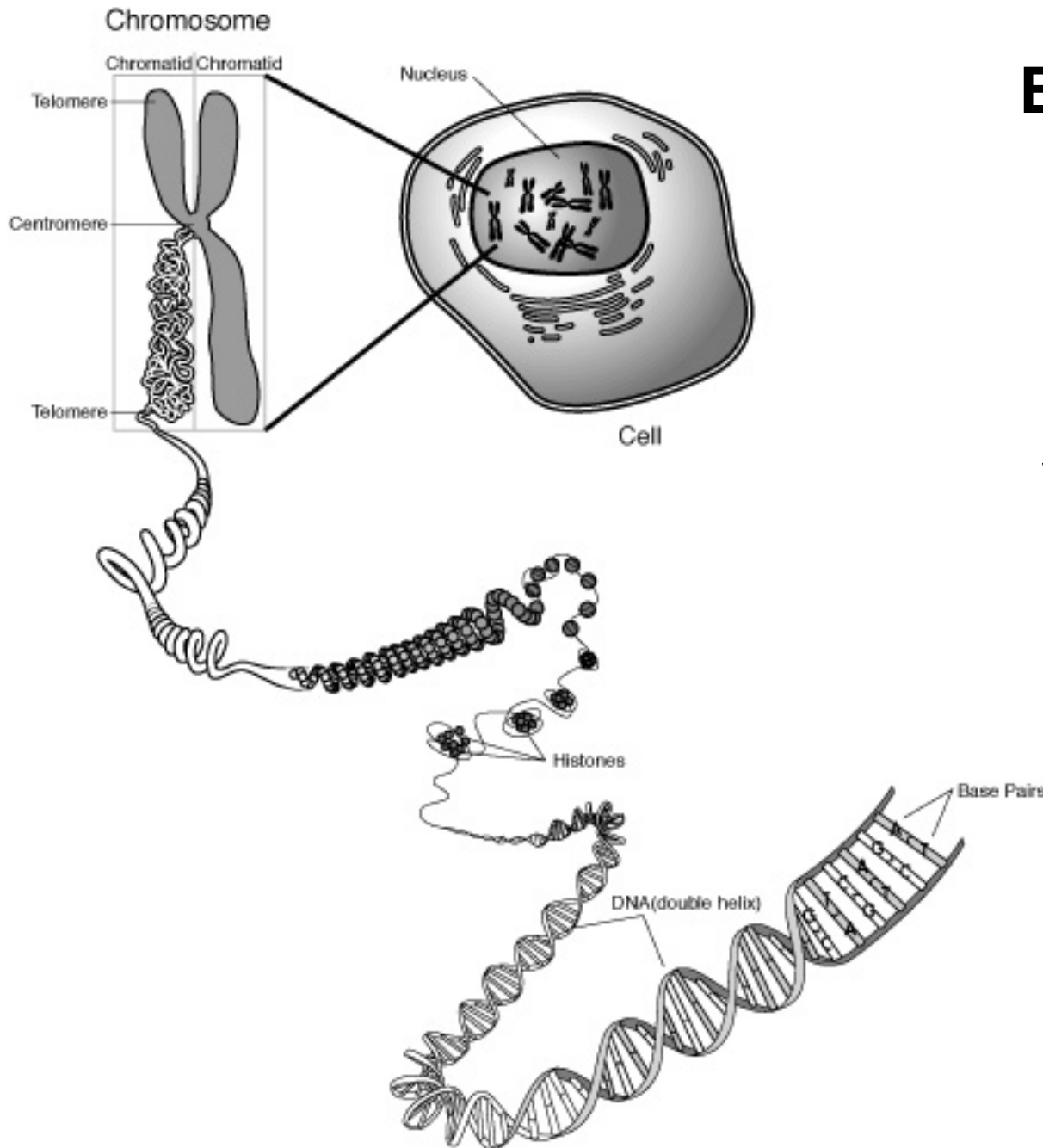
### Age-Related Macular Degeneration



### Three bad and two good alleles

TGGAAAATCCATAATGGGTTTGAAGGA

AATATTTTTTCCCTTTGTTAGCTGGCTCTGGGCAGCCT  
 TGCTGCTTGGGACCTAATGACCTGCTTTCAATCCCTT  
 AATTTGGAAAACAACCAAGGCTTGGATGG  
 TCTGTACCCAGTTTTTCAAAGAGATTTTTTTTTTTTCA  
 AGTCCTGGACCTTTGGCAGCAAAGGGTGGGACTTCTG  
 AGCTCAGCGGGGCCCTCCCGCTTGGATGTTCCGGGAA  
 GGCGAGCCGCAGGTGCCAGAACACAGATTGTATAAAA  
 GGGAAAGGGAAATGTGACCAGGTCTAGGTCTGGAGTTTC  
 ACAAGCAAAGCAAGCCAGGACACACCATCCTGCCCCA  
 CAACGCCATGGGGAGCAATCTCAGCCCCCAACTCTGC  
 CTTGTCTGGAGGTAAGCGAGGGGTAACCTTCCCTTCTC  
 GCCTTTTGGGCCAGGCTTCATCAGCCTTTTCTCTTCA  
 TTTGGCCCGGCCAGGGATCCTGCTCTCTGGAGGGG  
 CCGACTTCTCAAAGAGGGCCAGGCACTGGAGTACGTG  
 CCCTGTGCAGACACGTACCTGCAGATCTACGGGGTCC  
 CAAAAGACTGTCAGGAAGGCAGAGTGCAGAGGTTTG  
 CCTAAGGCAGAAACAGGGCAGGCGGCAGCAAGGTCAG  
 TGACAAGGTGGGCTGACCGGGAGTAGGAGCAGTTTTA  
 AGAAAAAGCGGAGTTAACCTTACTAAGCATTTACCC  
 GTCAAGAGAACACTCAGAAATGGGGAGGGAGAAGCAG  
 GTAGGTAAAGATGCTGCTTCTGCGGGGACTG00110101



# Building blocks of genetic variation

**Within each cell:**

2 copies of the genome

23 chromosomes

~20,000 genes

3.2B letters of DNA

Millions of polymorphic sites

# Types of genetic variation

- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

Name	Example	Frequency in one genome
Single nucleotide polymorphisms ( <b>SNPs</b> )	GAGGAGAACG[C/G]AACTCCGCCG	1 per 1,000 bp
Insertions/deletions ( <b>indels</b> )	CACTATTC[C/CTATGG]TGTCTAA	1 per 10,000 bp
Short tandem repeats ( <b>STRs</b> )	ACGGCAGTCGTCGTCGTCACCGTAT	1 per 10,000 bp
Structural variants ( <b>SVs</b> ) / Copy Number Variants ( <b>CNVs</b> )	Large (median 5,000 bp) deletions, duplications, inversions	1 per 1,000,000 bp



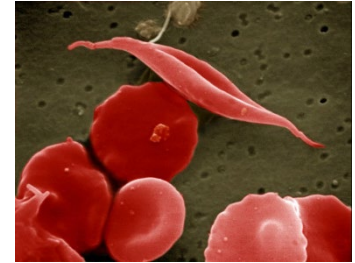
# Single-nucleotide polymorphisms (SNPs)

CATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTG  
 CATGGTGCATCTGACTCCTG**T**GGAGAAGTCTGCCGTTACTG

glutamic acid > valine

		Second letter				
		U	C	A	G	
First letter U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U	
	UUC } Leu	UCC } Ser	UAC } Tyr	UGC } Cys	C	
	UUA } Leu	UCA } Ser	UAA Stop	UGA Stop	A	
	UUG } Leu	UCG } Ser	UAG Stop	UGG Trp	G	
C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U	
	CUC } Leu	CCC } Pro	CAC } His	CGC } Arg	C	
	CUA } Leu	CCA } Pro	CAA } Gln	CGA } Arg	A	
	CUG } Leu	CCG } Pro	CAG } Gln	CGG } Arg	G	
A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U	
	AUC } Ile	ACC } Thr	AAC } Asn	AGC } Ser	C	
	AUA } Met	ACA } Thr	AAA } Lys	AGA } Arg	A	
	AUG } Met	ACG } Thr	AAG } Lys	AGG } Arg	G	
G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U	
	GUC } Val	GCC } Ala	GAC } Asp	GGC } Gly	C	
	GUA } Val	GCA } Ala	GAA } Asp	GGA } Gly	A	
	GUG } Val	GCG } Ala	GAG } Asp	GGG } Gly	G	

## Sickle Cell Anemia



rs189107123

GAGGAGAACG[**C/G**]AACTCCGCCG

- Many modern analyses (GWAS, eQTL) focus on SNPs/indels
- Often have only two **alleles** (states)
- Identified as reference SNP clusters (**rsid**)
- Submitted sequences containing a variant are clustered to build a database (**dbSNP**)
- To date, >100 M known variants in dbSNP

# Beyond SNPs: Tandem repeats and Indels

- Variable number tandem repeats

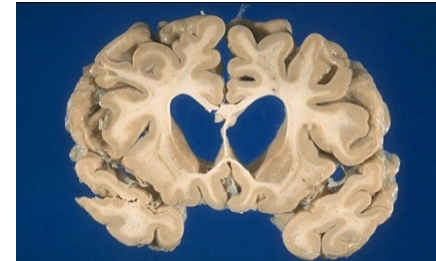
9 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTCATT

10 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTCATT

12 TCACAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGTTCATT

> 30 Huntington's Disease

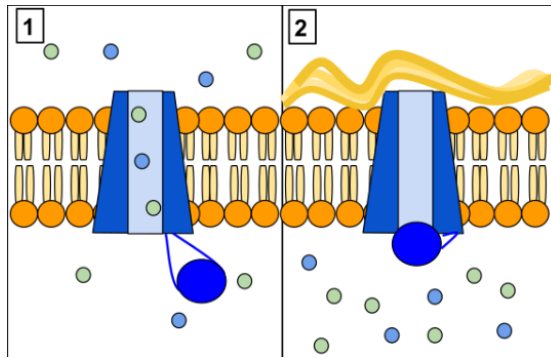
Abnormal protein, damages neurons, brain cell death, mood, coordination, speaking, dementia, etc



- Insertion/Deletions

Cystic fibrosis transmembrane conductance regulator (CFTR) -> Lung infections, cysts, fibrosis

CATTAAAGAAAATATCAT**CTT**TGGTGTTCCTATGATGAATA  
 CATTAAAGAAAATATCATTGGTGTTCCTATGATGAATA



CFTR Sequence:

Nucleotide	ATC	AT	<b>CTT</b>	T	GGT	GTT
Amino Acid	Ile	Ile	<b>Phe</b>		Gly	Val
	506		<b>508</b>			510

Deleted in  $\Delta F508$

$\Delta F508$  CFTR Sequence:

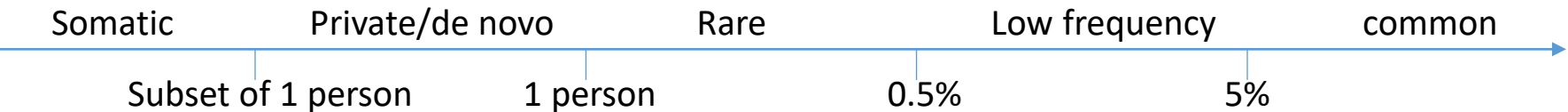
Nucleotide	ATC	ATT	GGT	GTT
Amino Acid	Ile	Ile	Gly	Val
	506			

# Variant alleles: ref/alt; maj/min; risk/prot; anc/der

Distinguishing the two alleles:

- Matching the human reference sequence (reference/alternate)
- Being more frequent in the population (major/minor)
- Matching the most recent common ancestor between human and chimpanzee (ancestral/derived)
- Based on their disease association (risk/non-risk)

Classifying variants by minor allele frequency:



Example: rs189107123

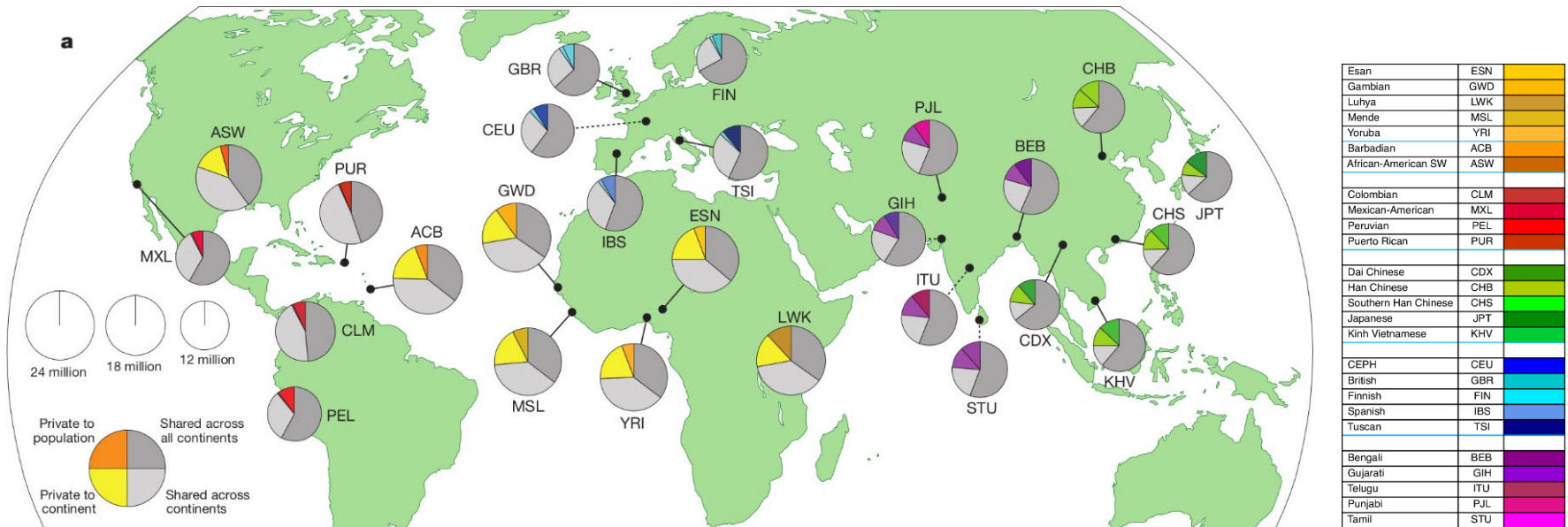
GAGGAGAACG[C/G]AACTCCGCCG

Reference allele: C

Minor allele: G (frequency 0.03 in Europeans)

Ancestral allele: unknown (**why?**)

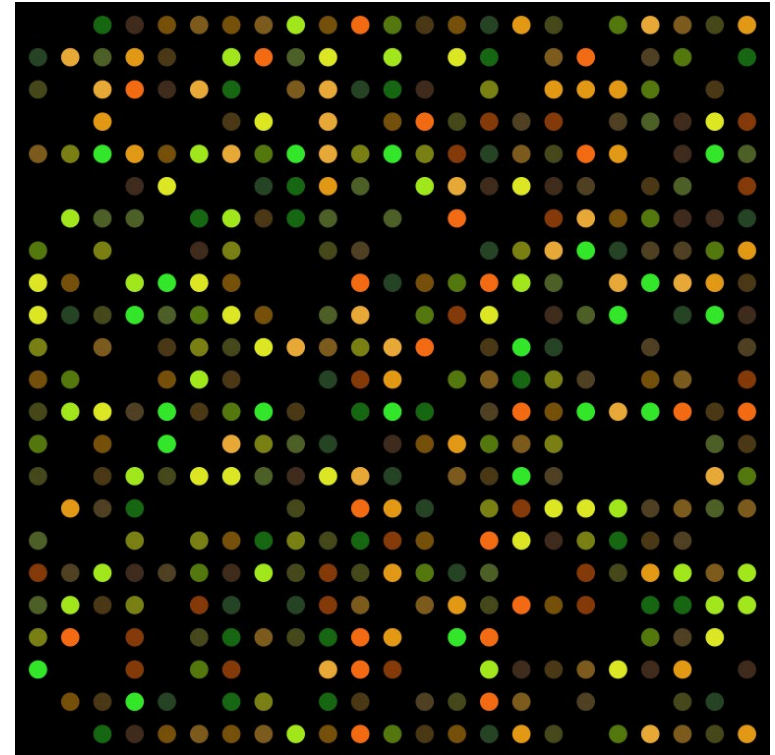
# Cataloguing genetic variants: Thousand Genomes Project



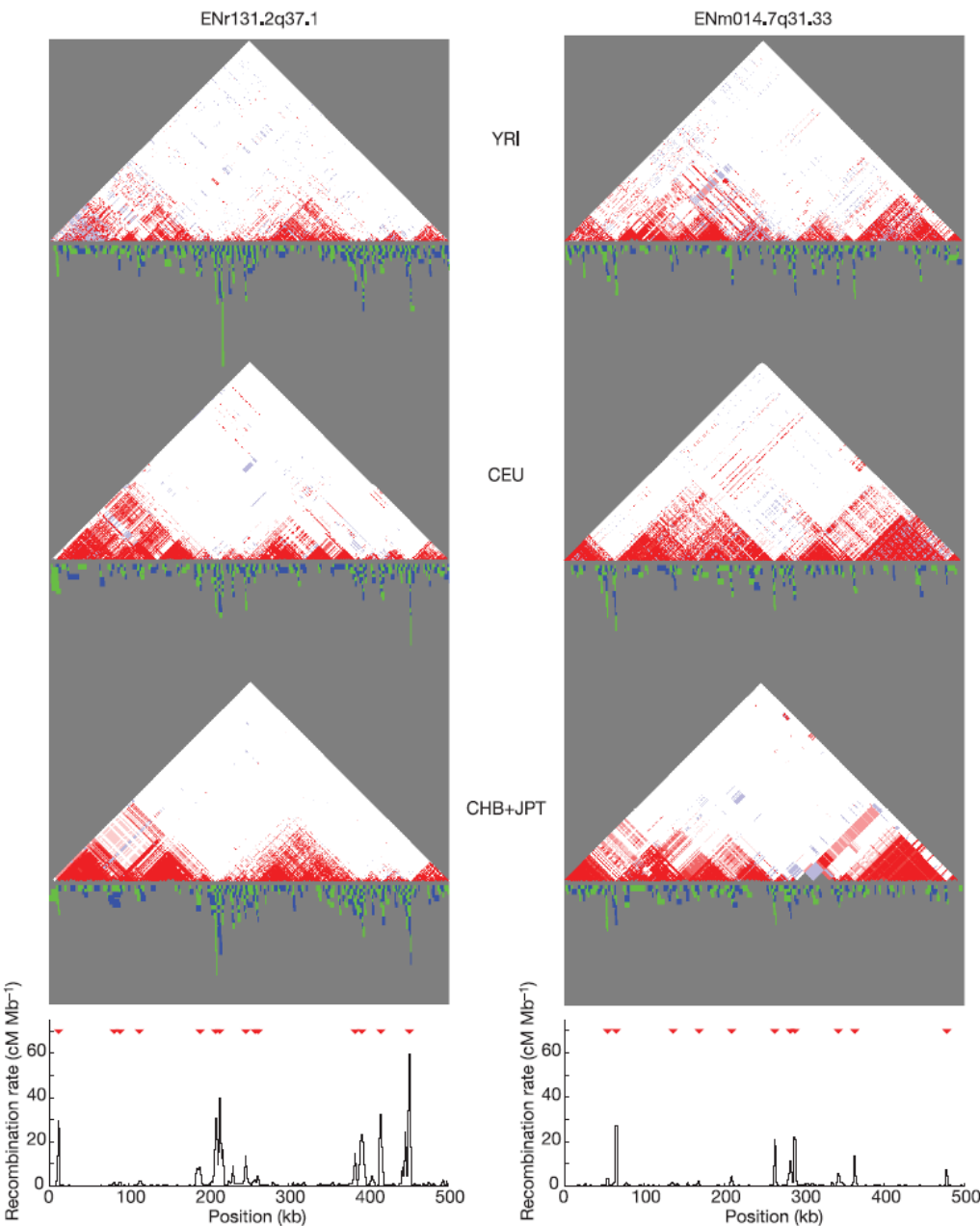
- 2,504 whole genome sequences at low depth (4x) across 26 subpopulations spanning the globe
- Develop sophisticated statistical tools (**phasing**, **imputation**) to account for noise, known patterns of variation (**linkage disequilibrium**; next section)

# Measuring known genetic variation: genotyping

- Key insight: Most **genetic variants** in an individual are recurrent in the population. Once they've been discovered/catalogued, build a **common array** for measuring them
- DNA microarrays were the key technological advance of the 1990s
- Idea: fragments of sample DNA containing SNPs will **hybridize** (reverse complement) to array **probes** (engineered DNA fragments)
- Tag fragments with fluorescent compound, use intensity to recover which probes were bound, which alleles were present in the sample
- Today: still the fundamental technology used in large-scale population genetic assays (GWAS, 23andMe)
- Next: study disease associations across populations, requiring new array designs due to differences in polymorphisms, LD across populations

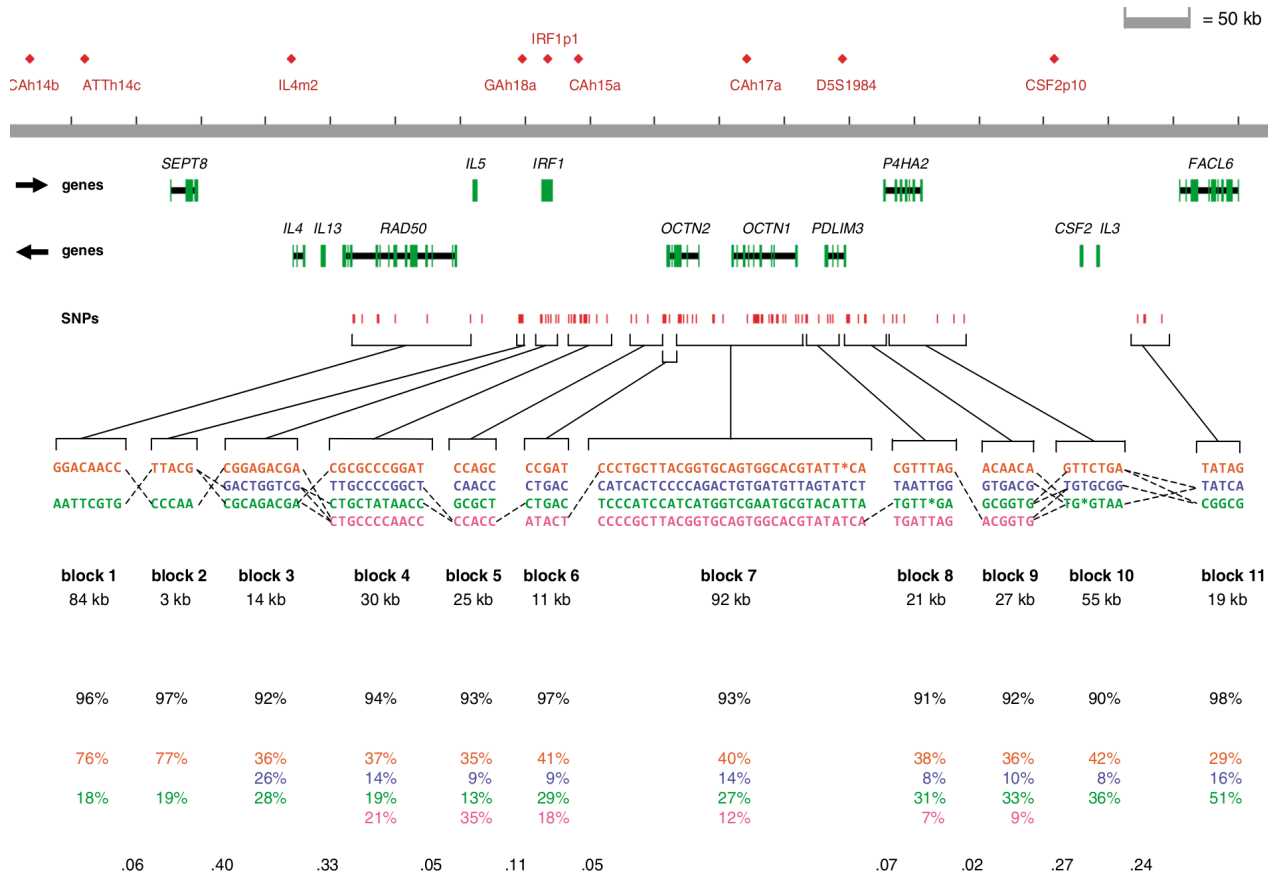


# $r^2$ and recombination events across regions/populations



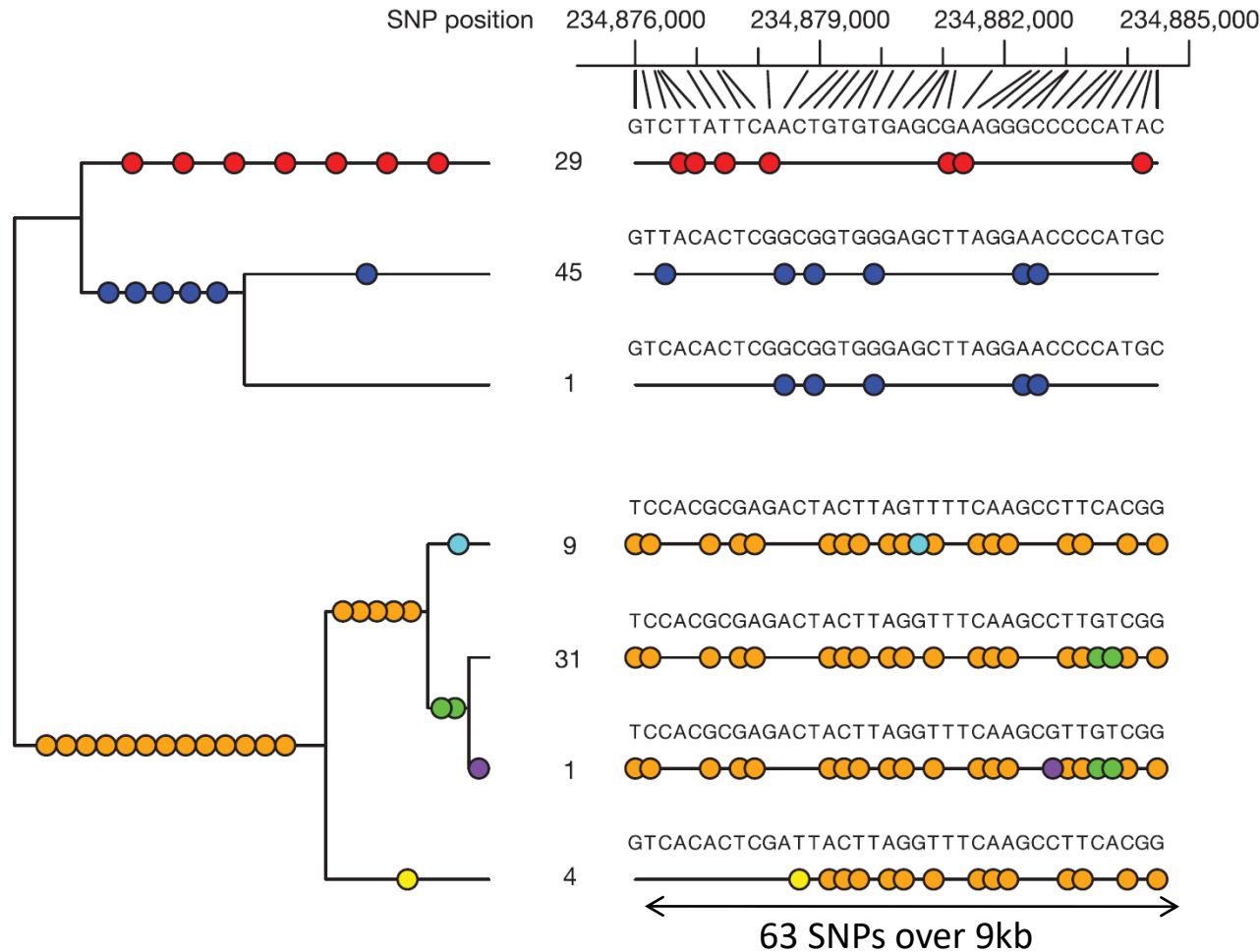
- Recurrent recombination events occur at hotspots
- $r^2$  correlations between SNPs depend on **historical order** in which they arose (not in their physical order on the chromosome)

# Long-range threading of haplotype blocks



- Relatively few haplotypes exist in the human population (consider 10M SNPs: we don't see  $2^{10M}$  haplotypes!)
- Implies high level of genotype sharing even for unrelated individuals

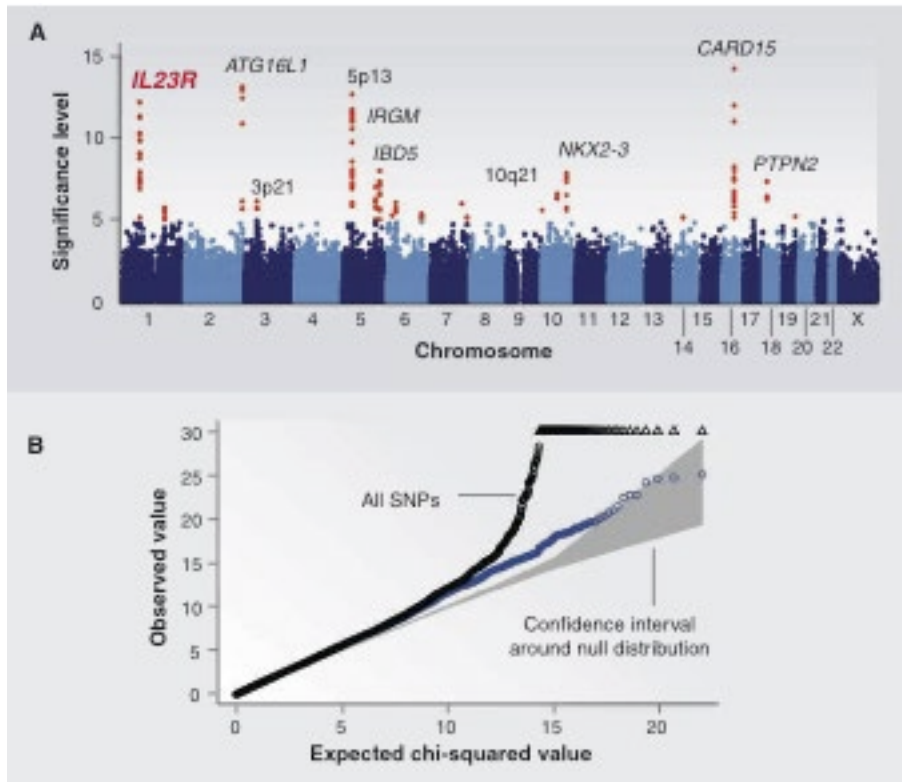
# Mutational history of multiple haplotypes



- Example region: 36 SNPs spanning 9kb
- In principle:  $2^{36}$  possible allele combinations (haplotypes)
- Sample 120 parental European chromosomes.
- In practice: only 5 recurrent haplotypes seen (and 2 singleton haplotypes)



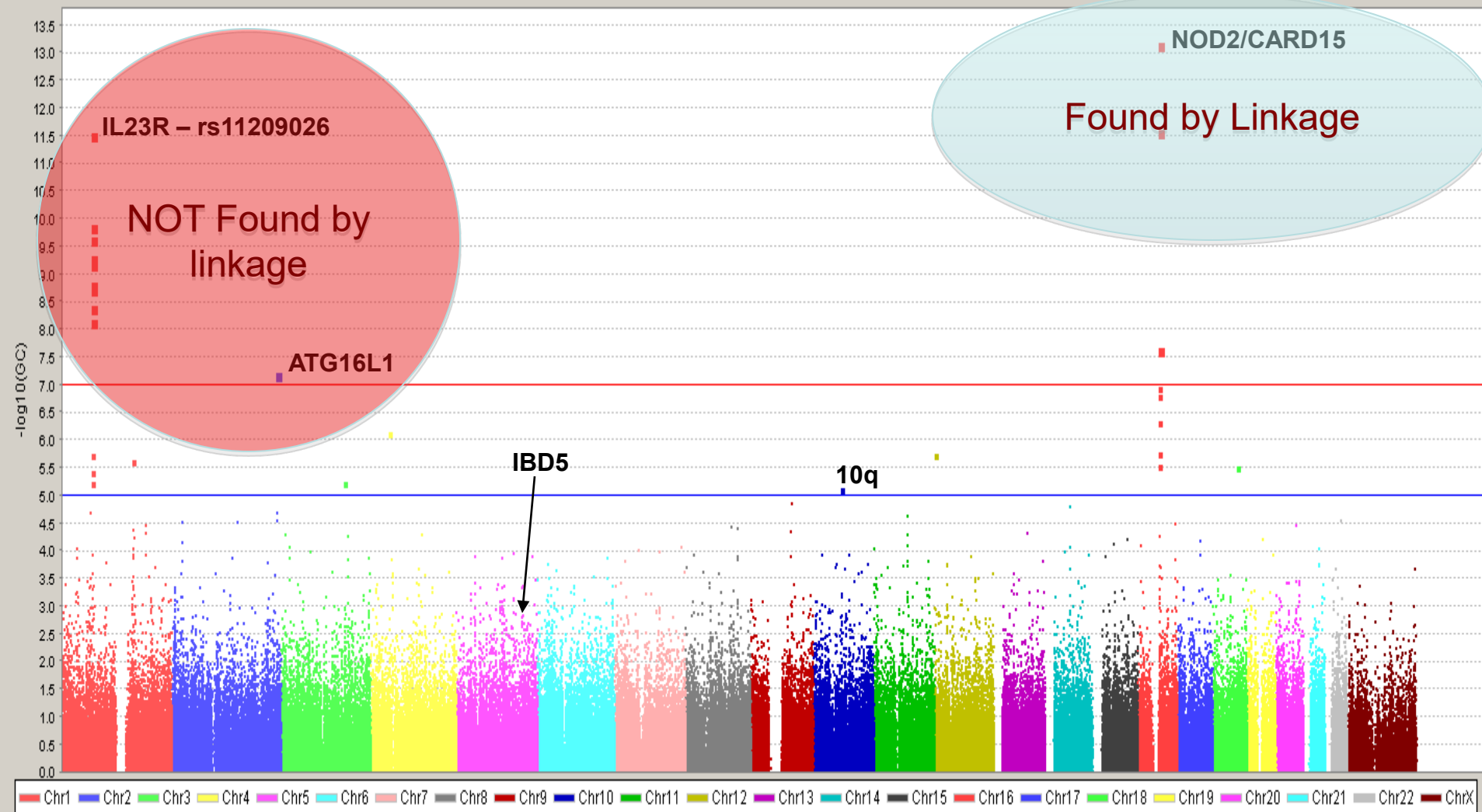
# Genomewide Association



‘Manhattan’ plot

Q-Q plot

# IBDGC Crohn's genome-wide association results



# Linkage vs. Association

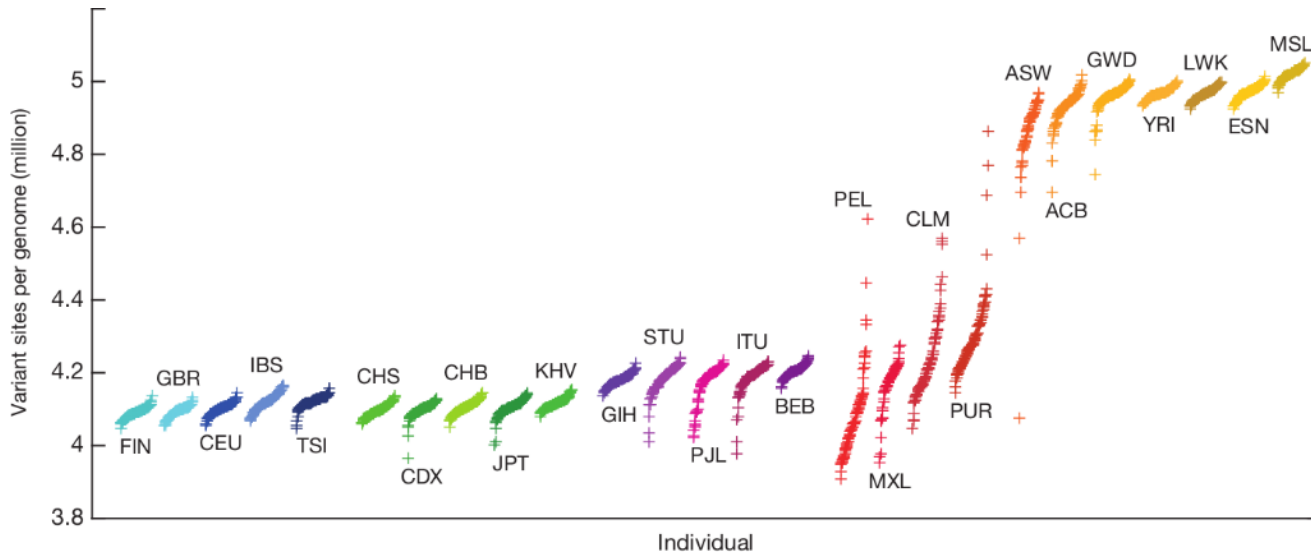
NOD2: low-frequency, strong risk variants

IL23R: low-frequency, strong protective variant

ATG16L1: common associated variant

Locus	Frequency	Odds-ratio	ASSOCIATION cases to achieve GWS	LINKAGE Pedigrees to achieve signif.
NOD2 (3 coding SNPs)	5%	3.0	435	1400
IL23R (Arg381Gln)	7%	0.33	817	~30,000
ATG16L1 (Thr300Ala)	50%	1.4	1360	~40,000

# Number of variants varies greatly by population

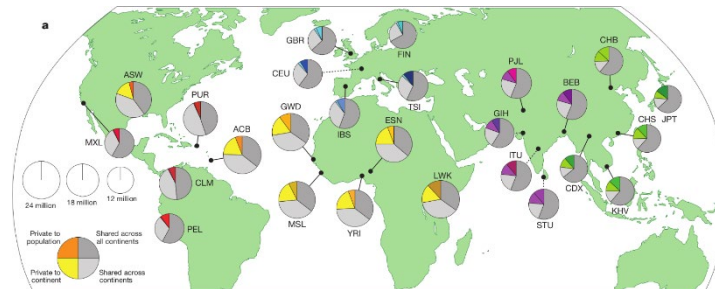
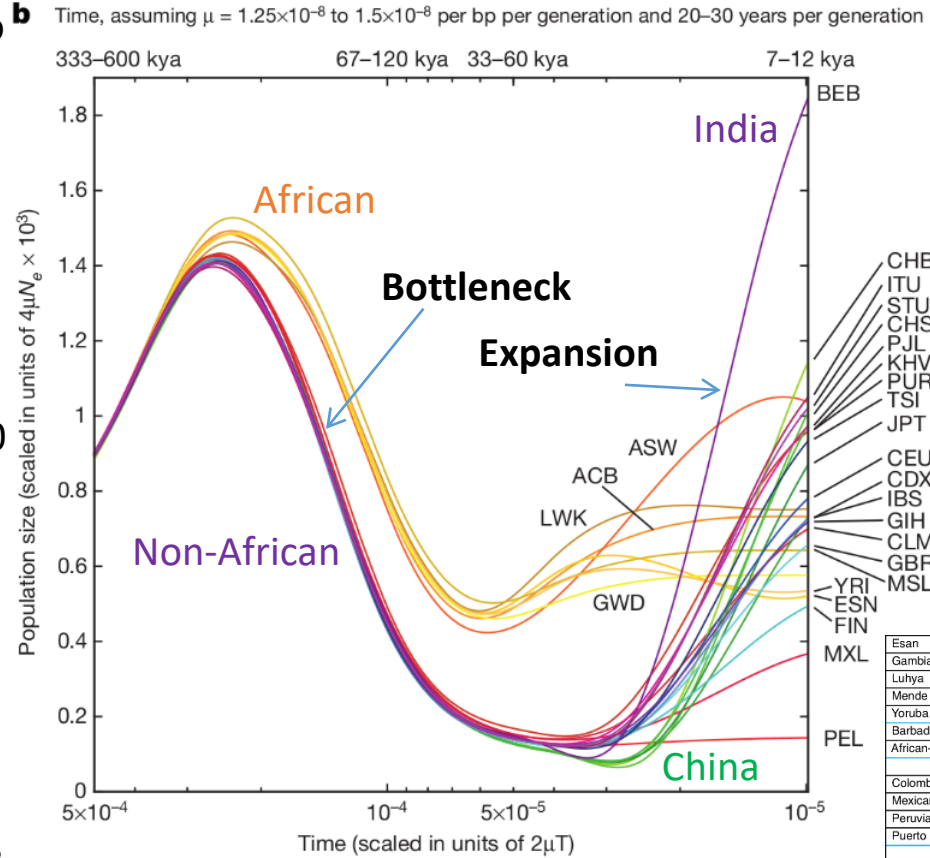


Africa	Esan	ESN	
	Gambian	GWD	
	Luhya	LWK	
	Mende	MSL	
	Yoruba	YRI	
	Barbadian	ACB	
	African-American SW	ASW	
S.America	Colombian	CLM	
	Mexican-American	MXL	
	Peruvian	PEL	
	Puerto Rican	PUR	
E. Asia	Dai Chinese	CDX	
	Han Chinese	CHB	
	Southern Han Chinese	CHS	
	Japanese	JPT	
	Kinh Vietnamese	KHV	
Europe	CEPH	CEU	
	British	GBR	
	Finnish	FIN	
	Spanish	IBS	
	Tuscan	TSI	
India	Bengali	BEB	
	Gujarati	GIH	
	Telugu	ITU	
	Punjabi	PJL	
	Tamil	STU	

- Over 100 million observed variants: 4-5M positions differ between each of us and the human reference
- Each of us carries 2-3K structural variants affecting 20mb of sequence
- Each of us carries hundreds of protein truncating variants, 10Ks of non-synonymous mutations
- African individuals have more variation in their genomes (**why?**)

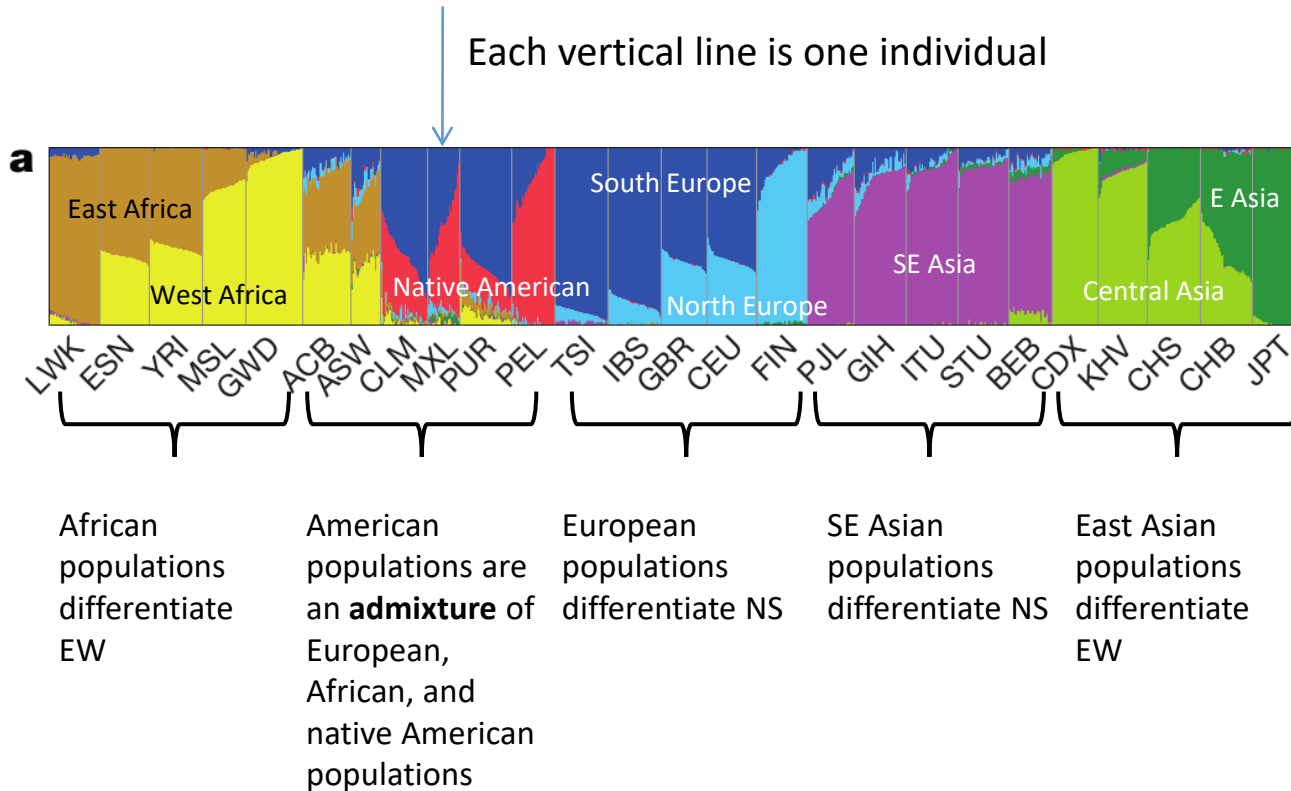
# Population size, bottlenecks and expansion

- **Effective population size:** number of individuals needed in idealized model to recapitulate population properties
- Here, recapitulate the **coalescent time:** time to most recent common ancestor
- **Pairwise Markov sequential coalescent model** with population splits/growth enables comparison within vs. between populations
- 1KG suggests shared history beyond 150 kya
- Non-African population: Loss of heterozygosity, **bottleneck** 15-20 kya (migration out of Africa)
- After migration, rapid population expansion (with interesting exceptions: Finland, Peru, Mexico)
- Bottlenecks/founder effects: rare alleles suddenly rise in frequency due to small population size
- Selective sweeps: rare alleles suddenly rise in frequency due to positive selection
- Admixture between previously isolated populations



Esan	ESN	
Gambian	GWD	
Luhya	LWK	
Mende	MSL	
Yoruba	YRI	
Barbadian	ACB	
African-American SW	ASW	
Colombian	CLM	
Mexican-American	MXL	
Peruvian	PEL	
Puerto Rican	PUR	
Dai Chinese	CDX	
Han Chinese	CHB	
Southern Han Chinese	CHS	
Japanese	JPT	
Kinh Vietnamese	KHV	
CEPH	CEU	
British	GBR	
Finnish	FIN	
Spanish	IBS	
Tuscan	TSI	
Bengali	BEB	
Gujarati	GIH	
Telugu	ITU	
Punjabi	P.JL	
Tamil	STU	

# Ancestry painting: population-level



Esan	ESN	
Gambian	GWD	
Luhya	LWK	
Mende	MSL	
Yoruba	YRI	
Barbadian	ACB	
African-American SW	ASW	
Colombian	CLM	
Mexican-American	MXL	
Peruvian	PEL	
Puerto Rican	PUR	
Dai Chinese	CDX	
Han Chinese	CHB	
Southern Han Chinese	CHS	
Japanese	JPT	
Kinh Vietnamese	KHV	
CEPH	CEU	
British	GBR	
Finnish	FIN	
Spanish	IBS	
Tuscan	TSI	
Bengali	BEB	
Gujarati	GIH	
Telugu	ITU	
Punjabi	PJJ	
Tamil	STU	

- Goal: infer **ancestry** of segments of the genome, **population structure** (patterns of relatedness between ancestry groups)
- Sharing of genetic variants enables **ancestry painting** of individual genomes
- The history of migration, settlement, conquest is written on our genomes

# Ancestry painting (e.g. admixed individual)

Chromosome View ⌵ ⊖ Sub-regional Resolution ⊕



Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. This analysis includes DNA you received from all of your recent ancestors, on both sides of your family. The results reflect where your ancestors lived before the widespread migrations of the past few hundred years.

79.0%	Sub-Saharan African
72.3%	West African
2.9%	Central & South African
3.8%	Broadly Sub-Saharan African
18.4%	European
2.5%	Northern European
0.2%	British & Irish
0.2%	Scandinavian
11.4%	Broadly Northern European
0.6%	Ashkenazi
0.5%	Southern European
3.3%	Broadly Southern European
3.3%	Broadly European
1.9%	East Asian & Native American
0.8%	Native American
0.8%	Southeast Asian
0.2%	Broadly East Asian & Native American
0.7%	Unassigned
100%	<b>TL Dixon</b>

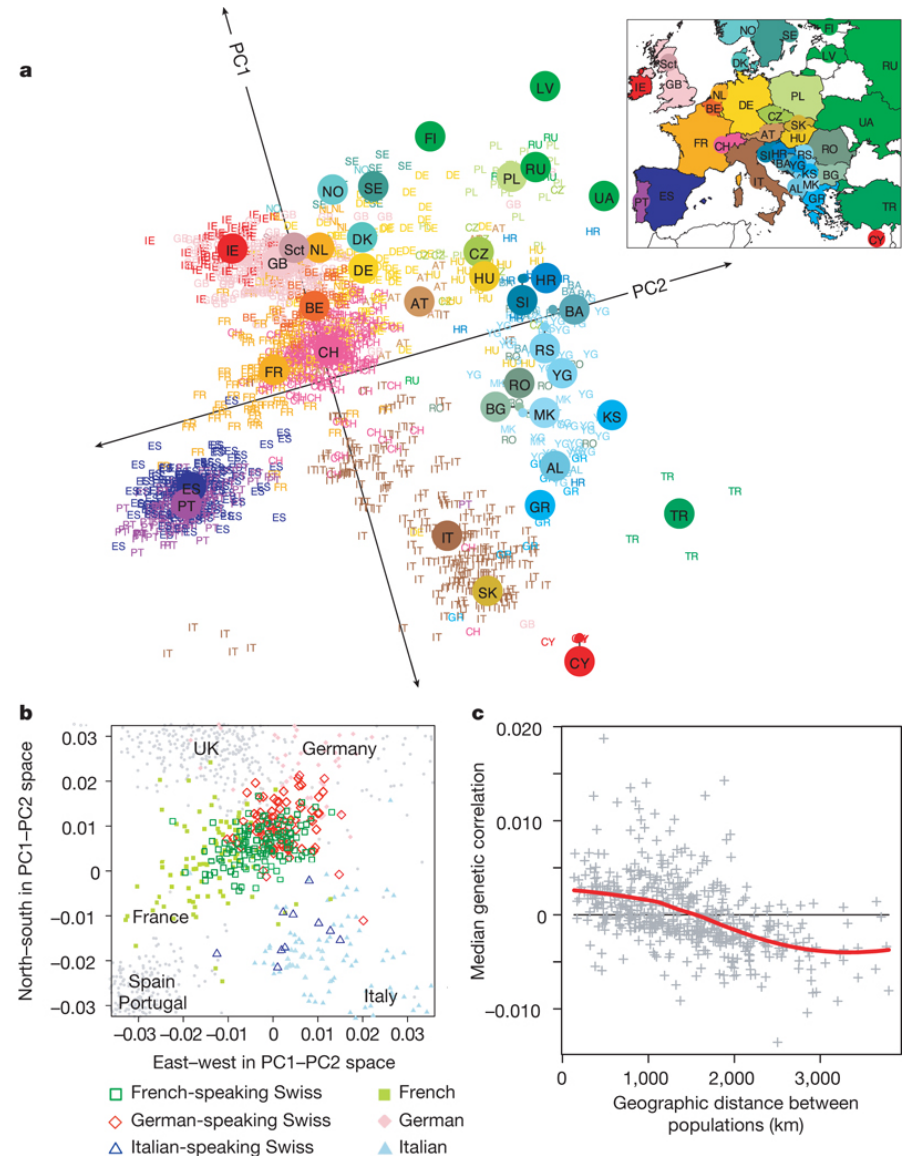
TL Dixon's Ancestry Composition results were updated on December 24, 2014.

[show all populations](#)

## Which segments of a genome are shared with what populations

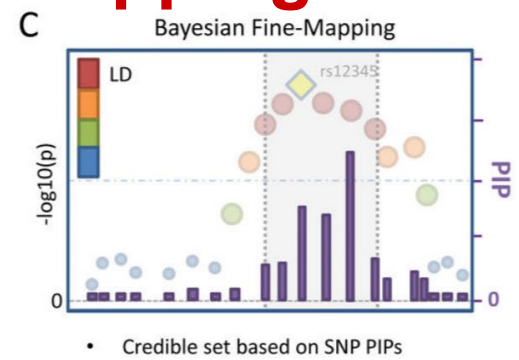
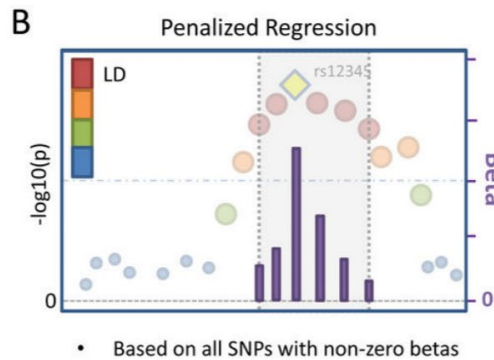
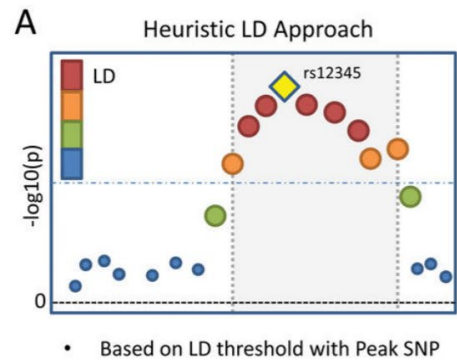
# Genetic relatedness and geography

- Can we decompose genetic variation into the major forces shaping it?
- ➔ PCA/SVD decomposition
- First components correspond to population structure.
- Population structure is shaped by geography! (people near each other are more likely to mate)
- In Europe, First two components correspond to N-S and E-W migration axes
- Country neighbors & borders visible at the genetic level

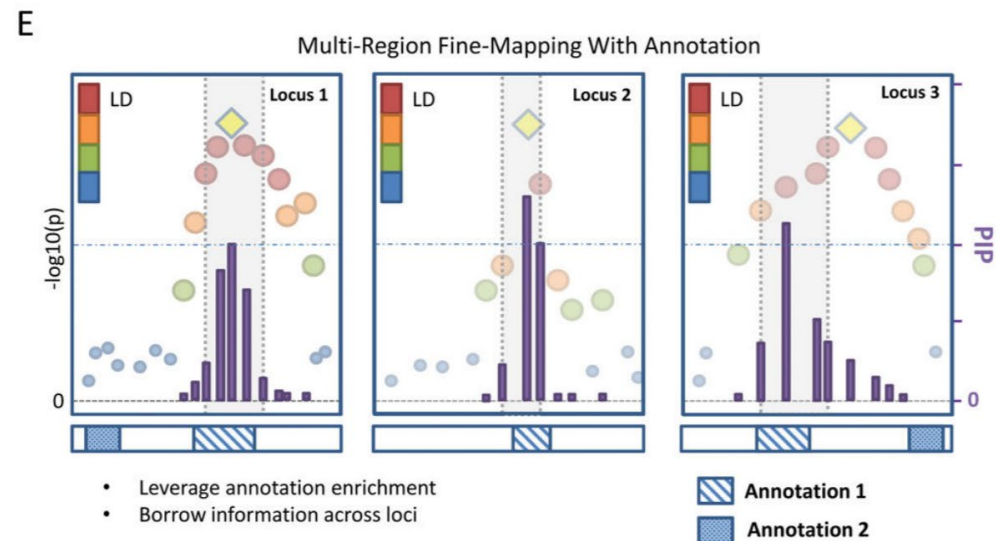
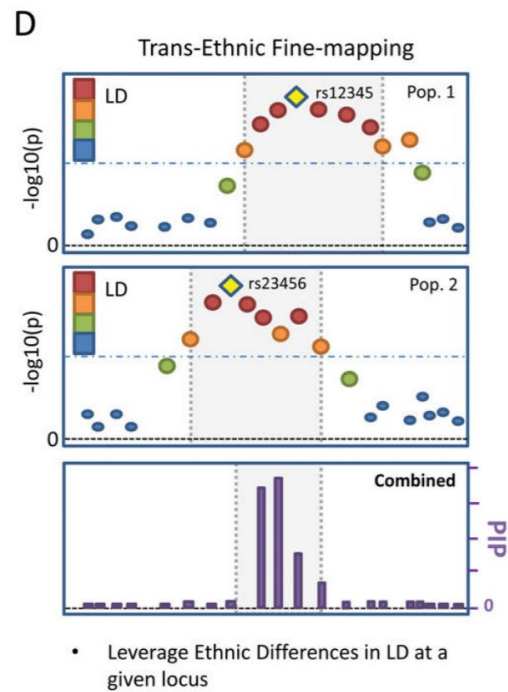




# GWAS fine-mapping

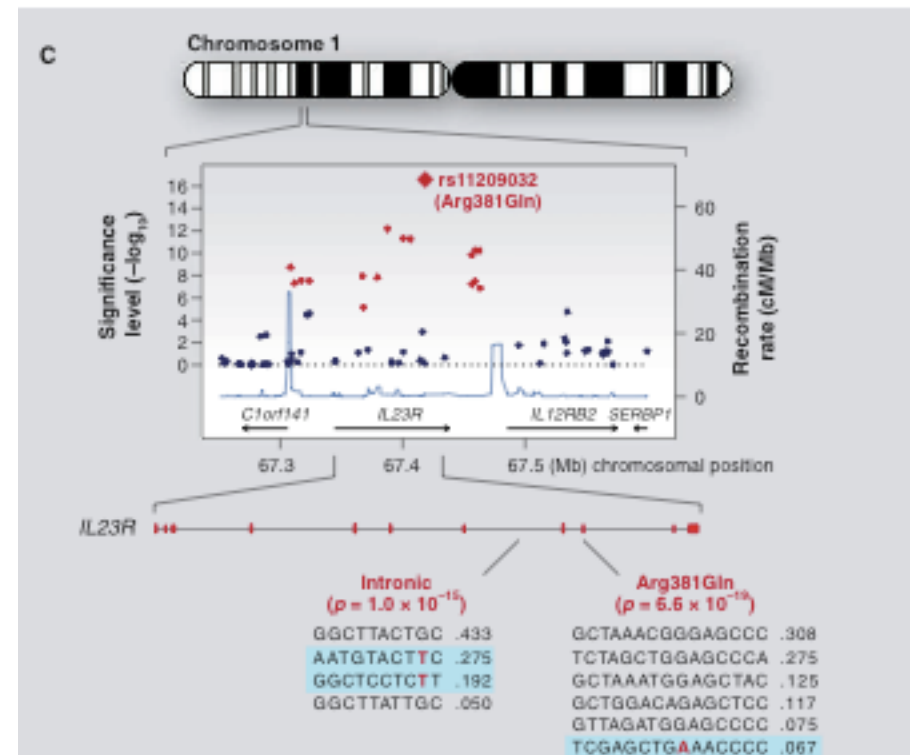
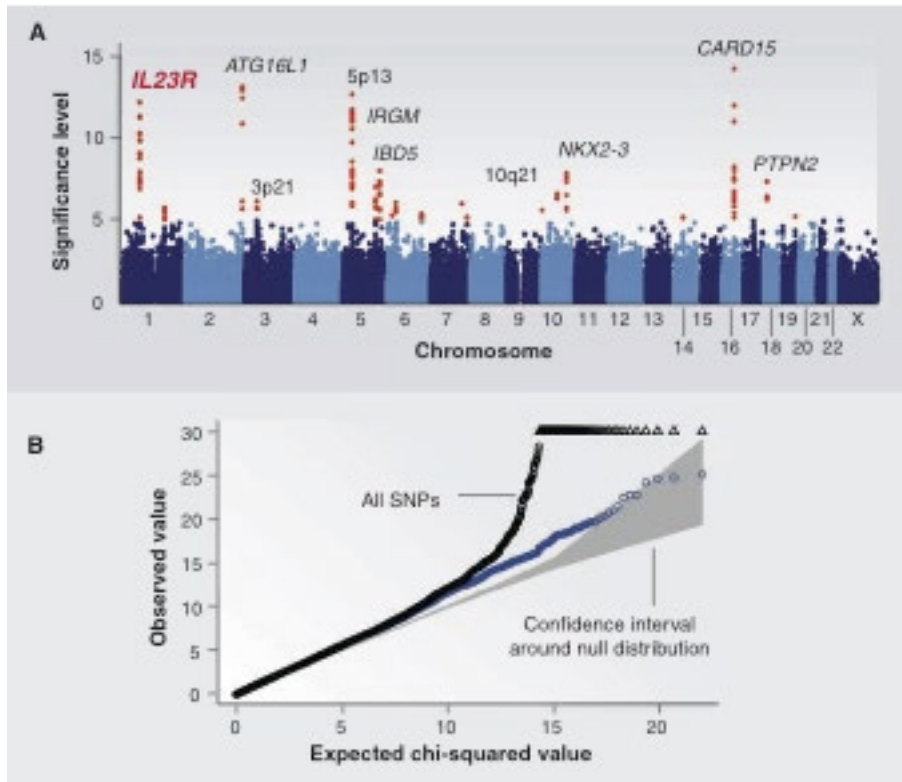


- LocusZoom of marginal SNP associations
- Y-axis:  $-\log_{10}(p\text{-values})$
- X-axis: Variant positions
- Gold: peak SNP
- Other=degree LD w/peak SNP (red, orange, green, blue)
- Purple bars=additional variant-level statistics by fine-mapping
- (Penalized regression=Beta; Bayesian: posterior inclusion probabilities (PIPs))
- Light grey=regions selected by fine-mapping



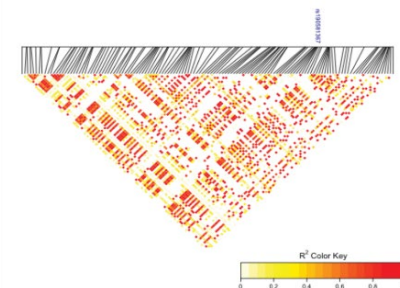
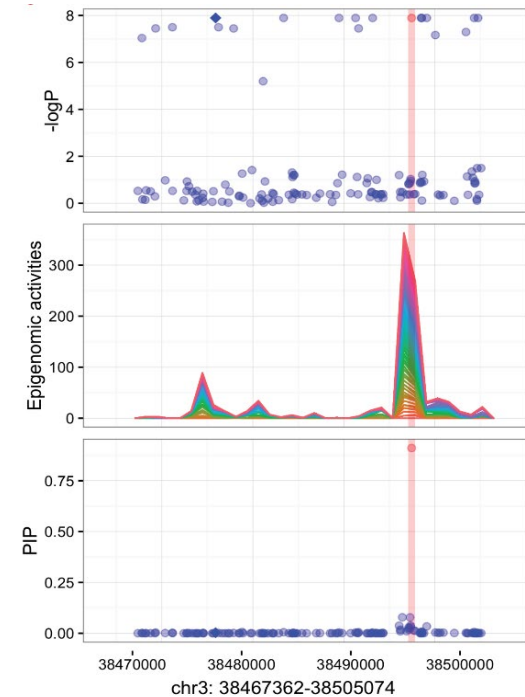
- **A**=heuristic using LD w/ peak SNP (>orange)
- **B**=Penalized regression=Beta not shrunk to zero
- **C**=Bayesian PIPs summed to credible sets using  $P_{\text{coverage}} > 95\%$  (note: peak SNP not always highest PIP ← correlation structure of SNPs in region)
- **D**=2 pops w/ different local LD struct → meta-analysis narrow fine-mapping credible region
- **E**=Anno1 overlap in locus 1 & 2 → predict top-PIP SNP in locus 3 (overlaps anno1)

# Fine Mapping



# Fine-mapping disease associations: (1) Epigenomics / functional data (next lecture)

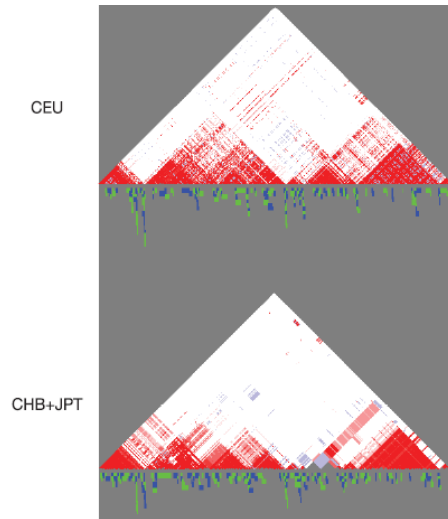
- **Association mapping** refers to identifying variants/gene associated with disease
- This is confounded by LD
- Many variants are strongly correlated to the true causal variant, and will show nearly as strong associations
- Use estimated correlations to explain correlated associations and recover the true underlying effects



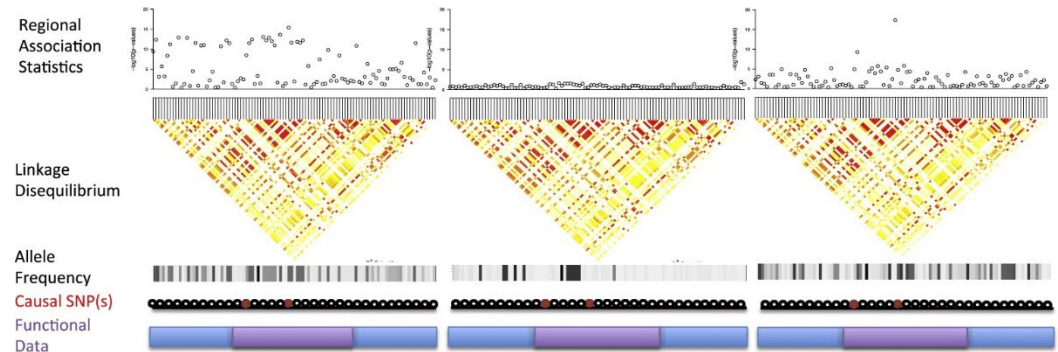
# Fine-mapping disease associations

## (2) Multi-ethnic analysis

Case 1: LD boundaries differ



Case 2: allele frequencies differ



- Allele frequencies and LD patterns can differ between populations
- Currently, disease associations are biased for discovery in European cohorts
- As we begin conducting association studies in Asia/Africa, there is a pressing need to develop statistical methods which can account for population genetic differences

# Overview: Genetic prediction of complex traits

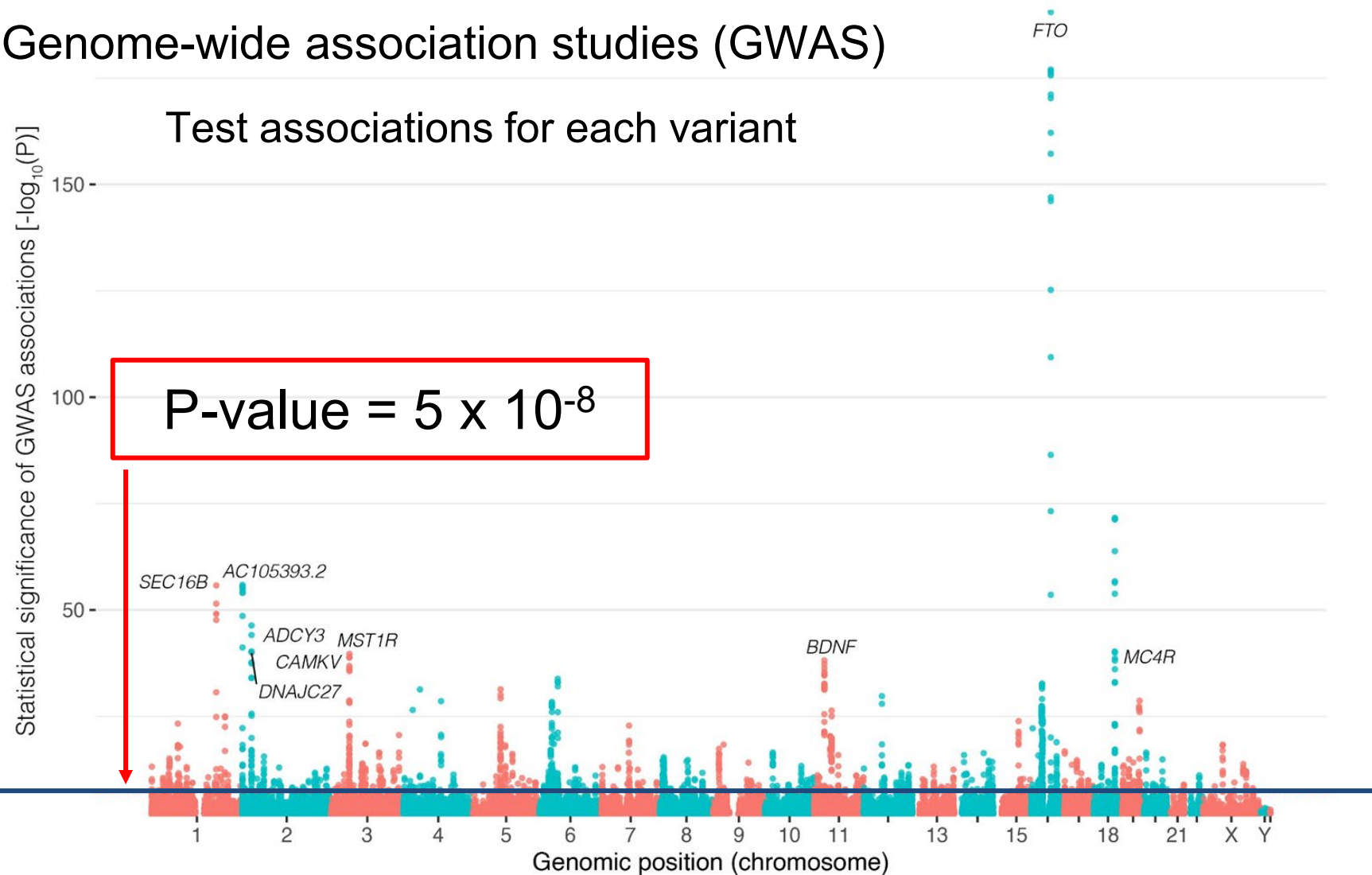
---

1. Foundations of Human Genetic Variation
2. Polygenic score (PGS) introduction
3. PGS Evaluation
4. Methods to fit PGS model
5. Challenges and opportunities in PGS research

# GWAS reveals complex traits are polygenic

## Genome-wide association studies (GWAS)

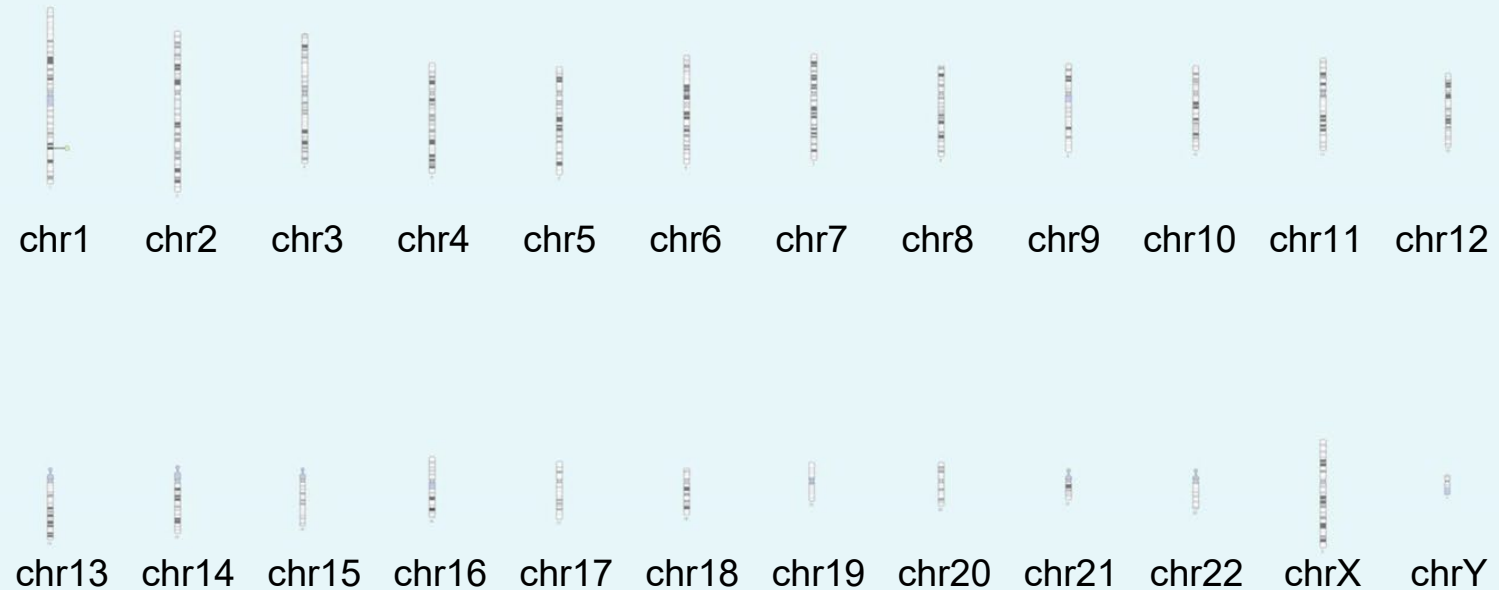
Test associations for each variant



# Mapping disease-associated variants with GWAS

---

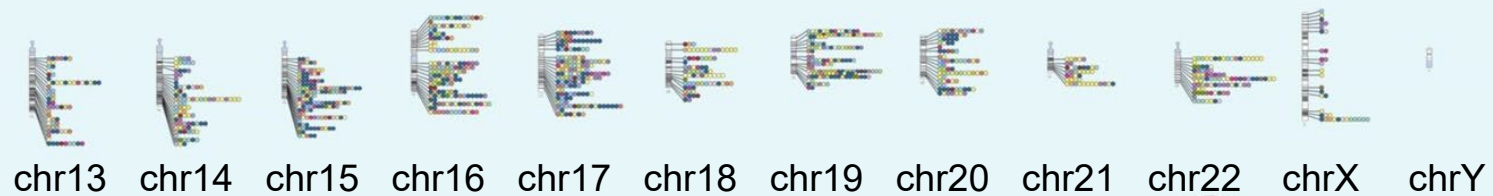
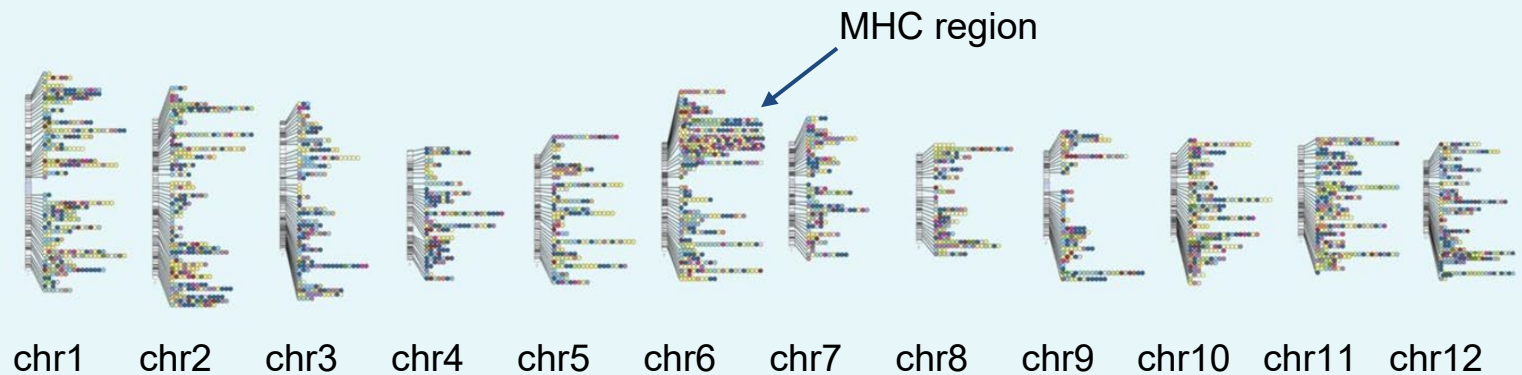
2006 Jan



[www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)

# Mapping disease-associated variants with GWAS

2013 Apr



[www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)



# Mapping disease-associated variants with GWAS

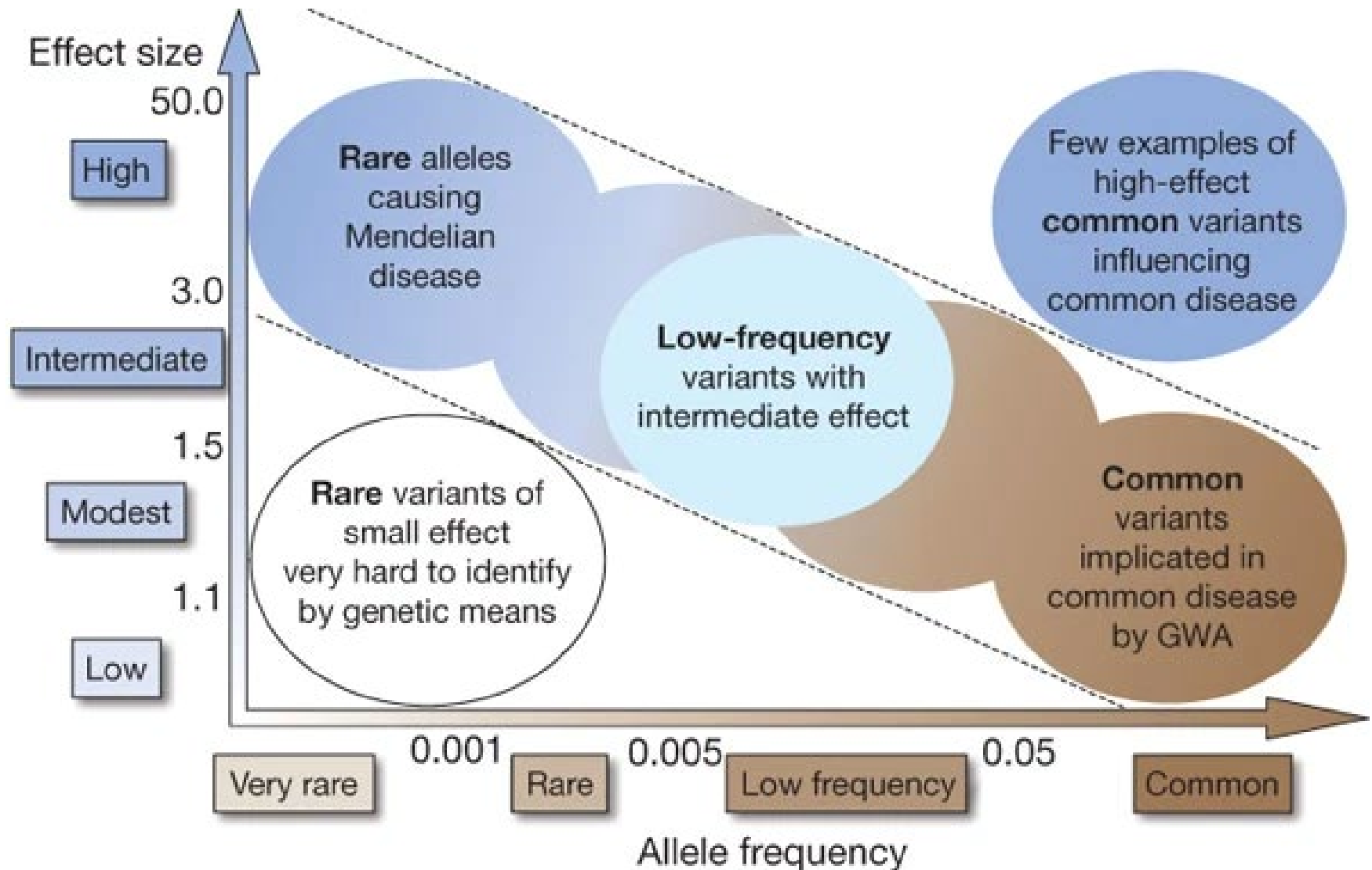
---

2019 July



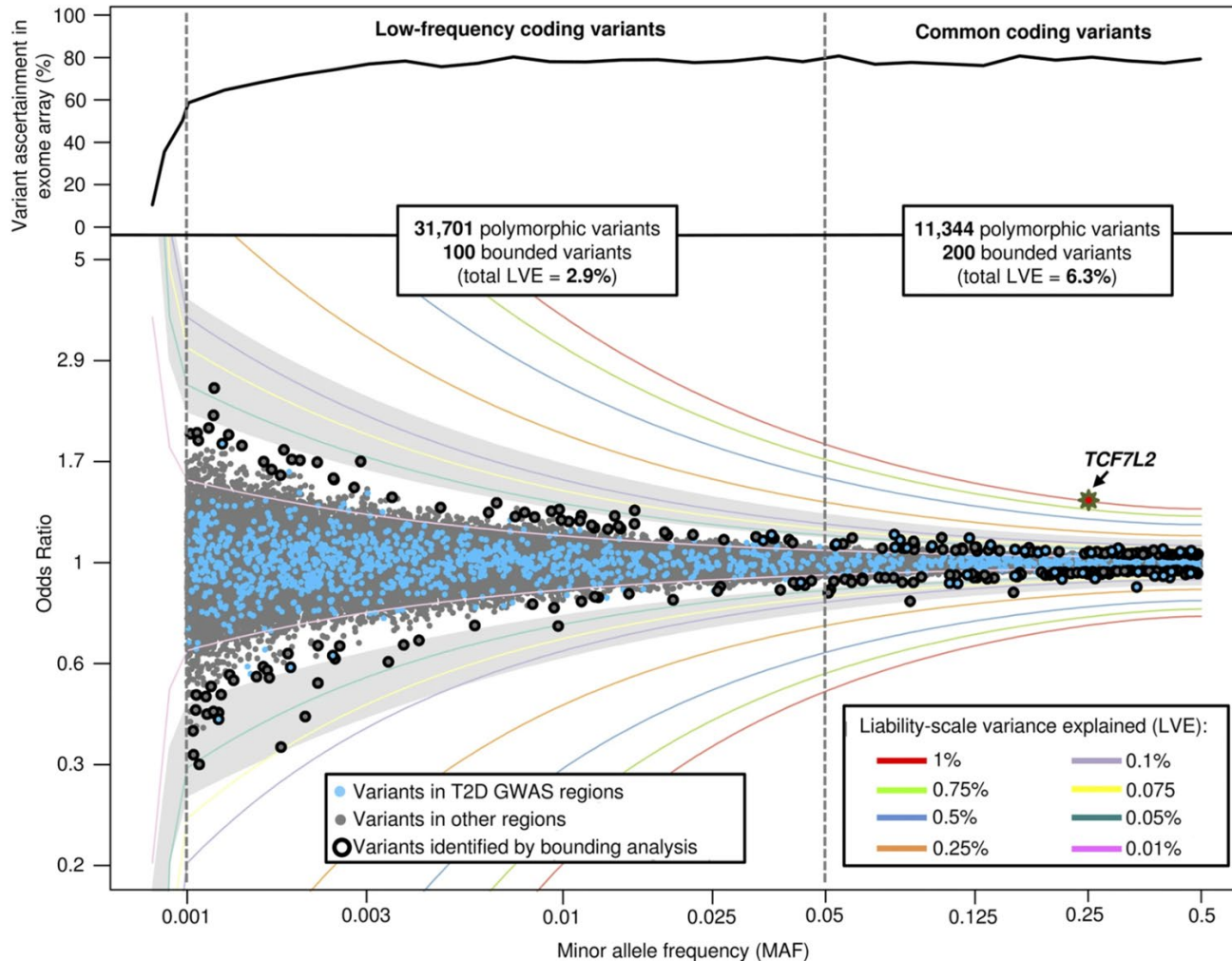
[www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)

# Most common variants have small effects



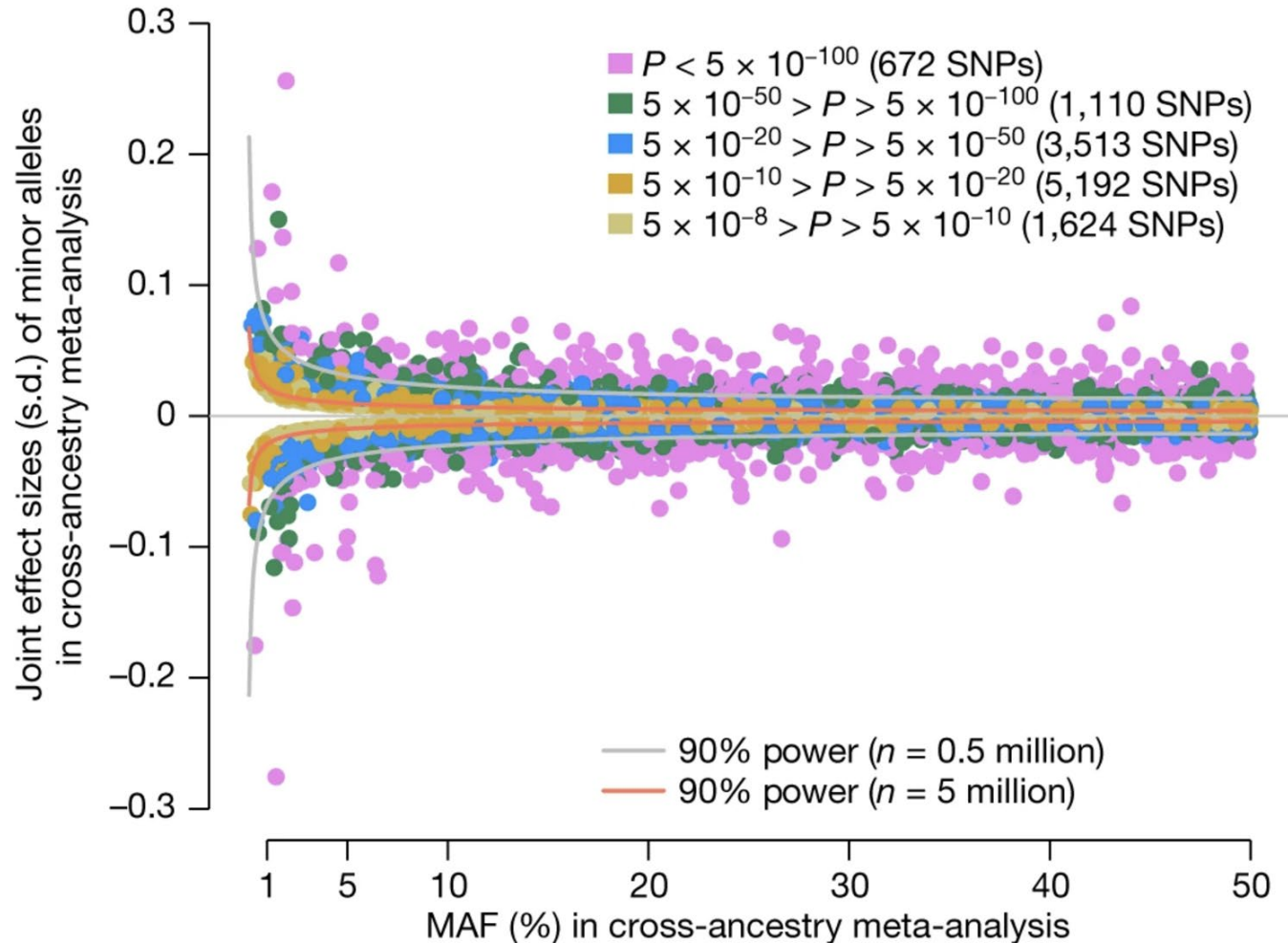
# Most common variants have small effects

## Type 2 diabetes



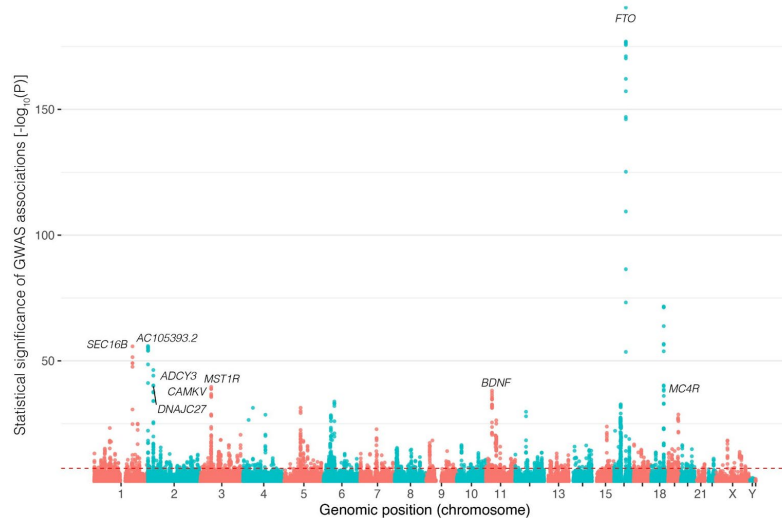
# Most common variants have small effects

Standing height (n = 5 million, 2022)

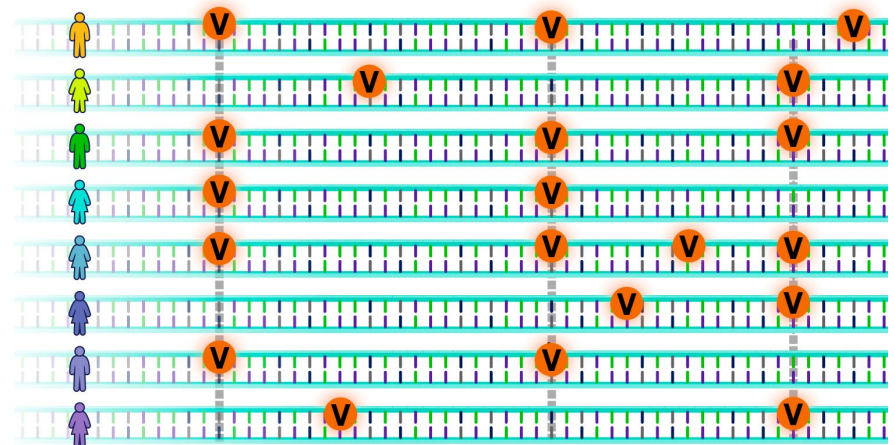


# Estimating individual-level liability of complex traits

Population-level inference vs. individual-level inference



Population-level inference  
(GWAS)

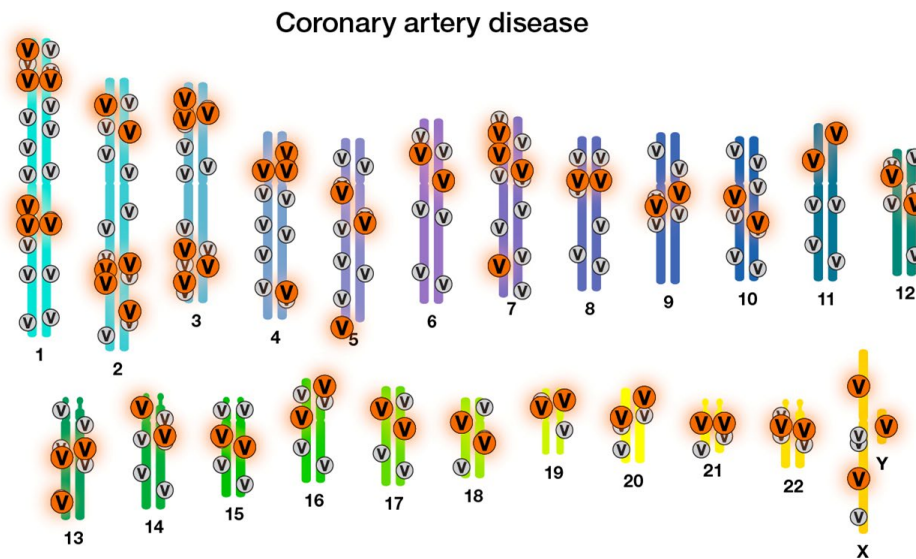


Individual-level inference  
(???)

How do we inform population-level insights into individuals?

# Challenges in polygenic complex traits

- Monogenic traits (e.g. cystic fibrosis)
  - “Carrier” or “non-carrier”
  - *CFTR* (cystic fibrosis transmembrane conductance regulator)
  - high penetrance, high effect size, often coding variants
- Polygenic complex traits (e.g. coronary artery disease, height, etc.)
  - Different individuals have a different subset of “risk” alleles
  - Lower penetrance, lower effect size, many non-coding variants

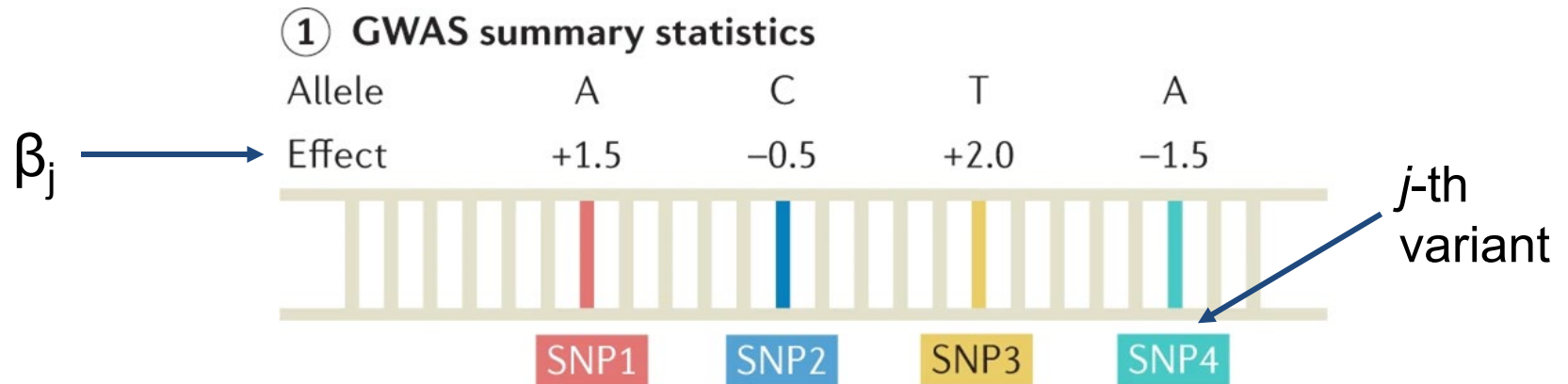


# Polygenic scores combine effects of disease-associated alleles for each individual

- Polygenic scores (PGS)
  - aka. Genetic risk score (GRS), Polygenic risk score (PRS), etc.
    - “risk” → disease risks
    - “Polygenic” → statement of the genetic architecture of a trait
- Polygenic score := weighted sum of disease-associated alleles

$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

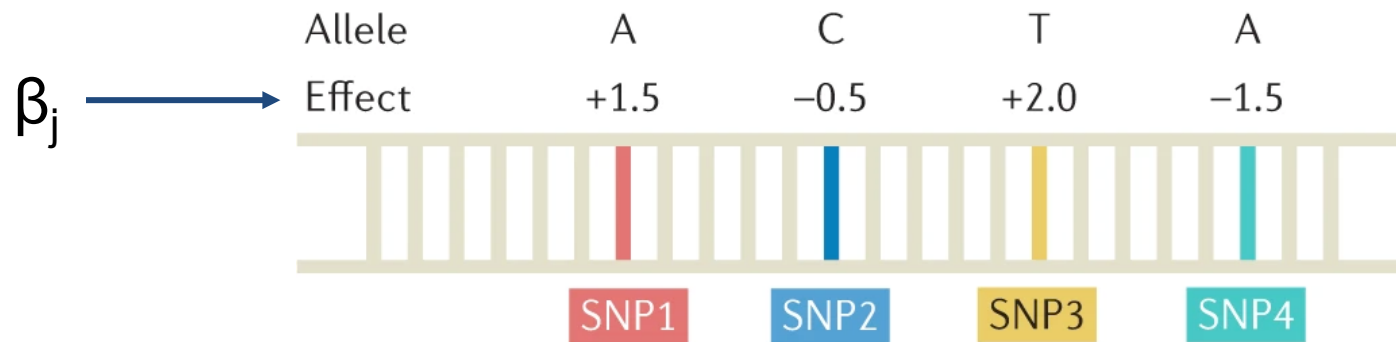
$i$ -th individual       $G$ : genotype  
 $j$ -th variant       $\beta$ : effect size



# Polygenic scores combine effects of disease-associated alleles for each individual

- Polygenic score  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$ 
  - $i$ -th individual
  - $j$ -th variant
  - $G$ : genotype
  - $\beta$ : effect size

## ① GWAS summary statistics



## ② Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

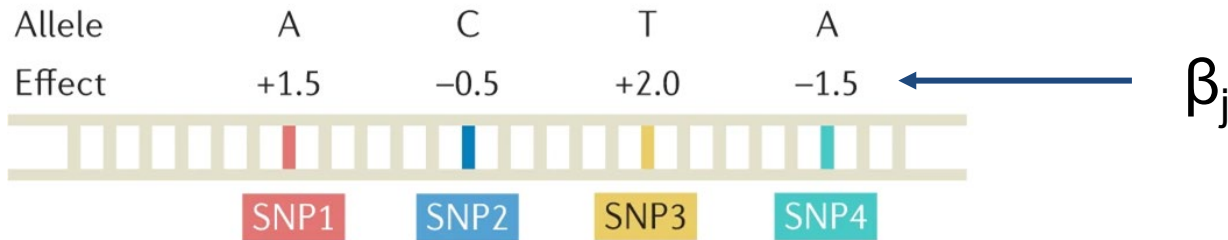
$i$ -th individual

$j$ -th variant



# Polygenic scores combine effects of disease-associated alleles for each individual

## ① GWAS summary statistics



## ② Genotype data

	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

$i$ -th individual  
genotype

$G$ :

$j$ -th variant  
size

$\beta$ : effect  
size

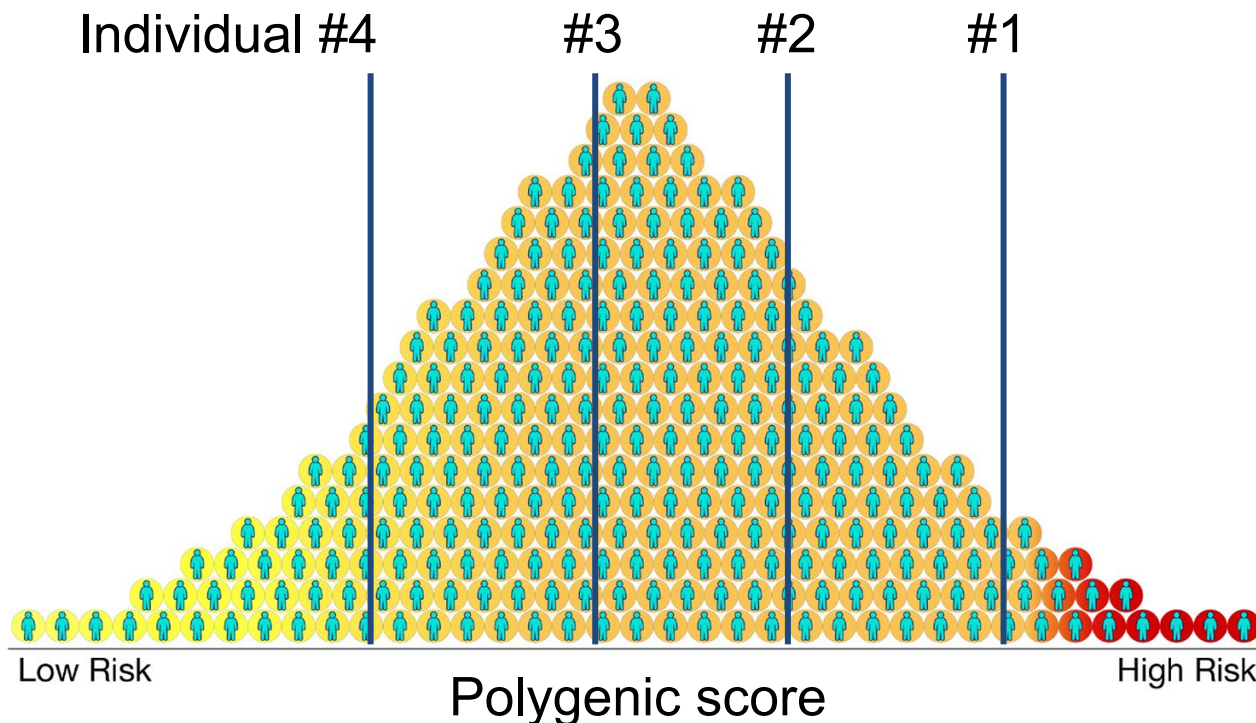
## ③ Polygenic risk score

Individual 1	1.5	-	0.5	+	4.0	-	0.0	=	<b>5.0</b>
Individual 2	1.5	-	0.0	+	2.0	-	1.5	=	<b>2.0</b>
Individual 3	0.0	-	1.0	+	2.0	-	1.5	=	<b>-0.5</b>
Individual 4	0.0	-	1.0	+	0.0	-	3.0	=	<b>-4.0</b>

# Polygenic scores combine effects of disease-associated alleles for each individual

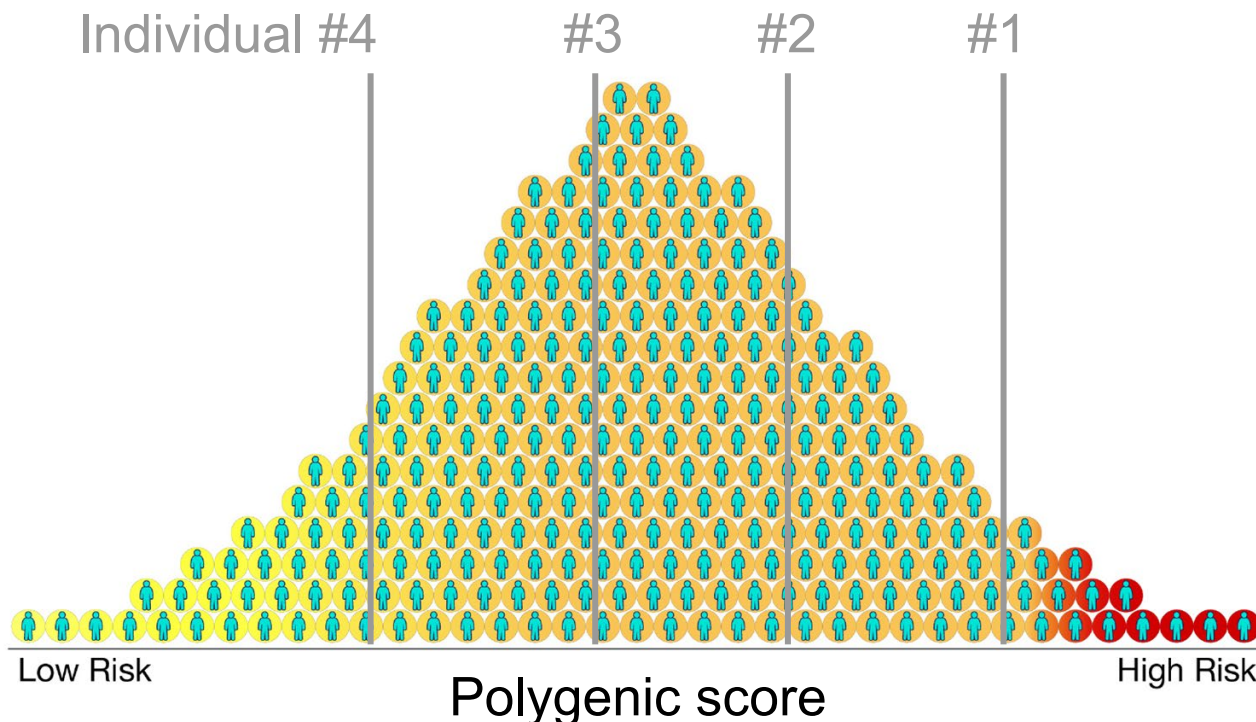
## ③ Polygenic risk score

Individual 1	1.5	-	0.5	+	4.0	-	0.0	=	5.0
Individual 2	1.5	-	0.0	+	2.0	-	1.5	=	2.0
Individual 3	0.0	-	1.0	+	2.0	-	1.5	=	-0.5
Individual 4	0.0	-	1.0	+	0.0	-	3.0	=	-4.0



# Polygenic scores estimate the relative genetic liability of disease

- **Genetic** liability of the disease – complex traits are influenced by genetics, environmental factors, and their interactions
- “**Relative**” – baseline risk factors (age, biological sex, comorbidity, ...) are not part of the picture
- “**Estimate**” – sample size & statistical power, model misspecification



# Potential of PRS in clinical practice

---

## AHA SCIENTIFIC STATEMENT

---

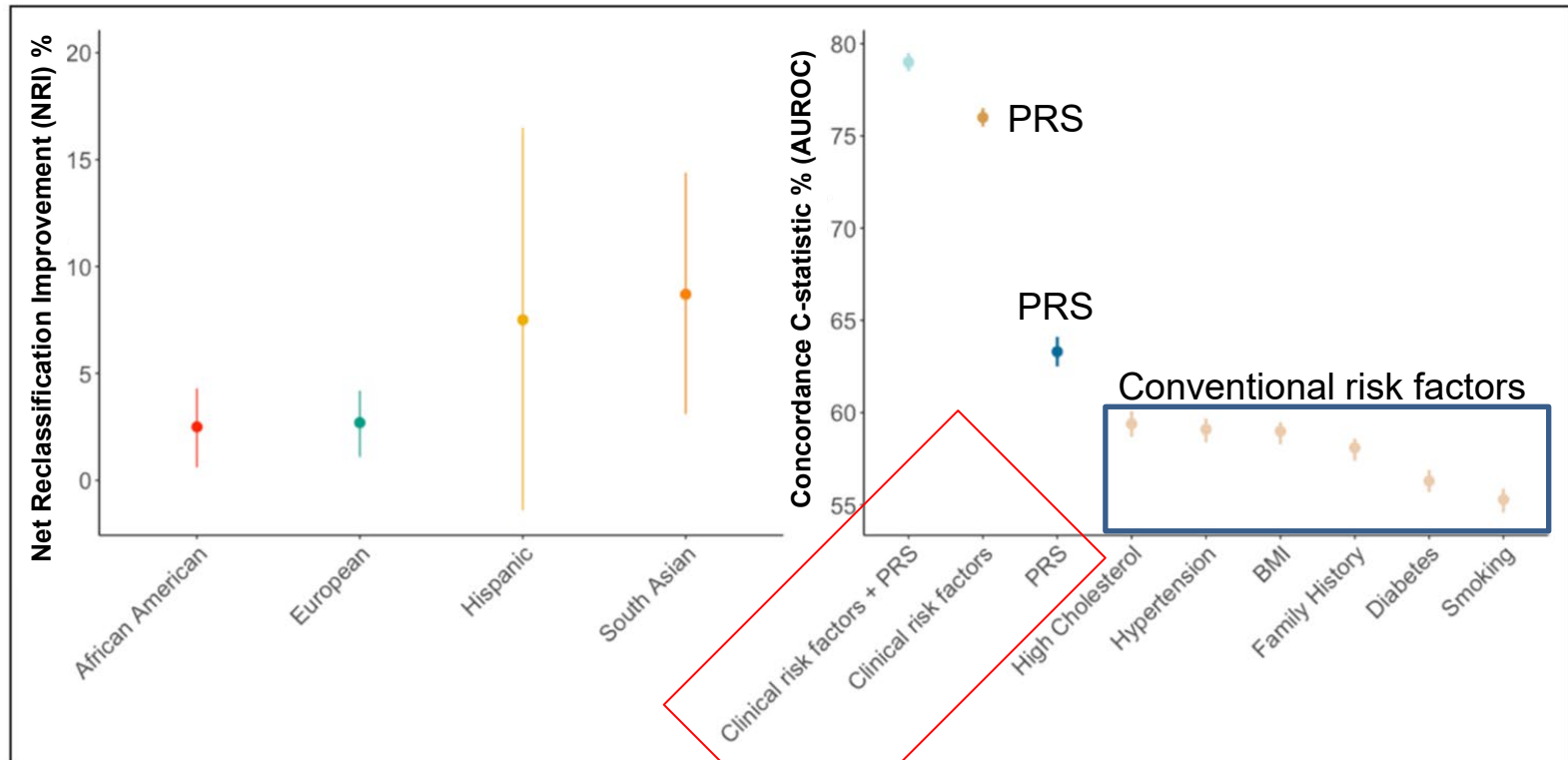
### Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association

Jack W. O'Sullivan, MBBS, DPhil, Chair; Sridharan Raghavan, MD, PhD; Carla Marquez-Luna, PhD; Jasmine A. Luzum, PharmD, PhD; Scott M. Damrauer, MD, FAHA; Euan A. Ashley, MBChB, DPhil, FAHA; Christopher J. O'Donnell, MD, MPH; Cristen J. Willer, DPhil; Pradeep Natarajan, MD, MMSc, Vice Chair; on behalf of the American Heart Association Council on Genomic and Precision Medicine; Council on Clinical Cardiology; Council on Arteriosclerosis, Thrombosis and Vascular Biology; Council on Cardiovascular Radiology and Intervention; Council on Lifestyle and Cardiometabolic Health; and Council on Peripheral Vascular Disease

*“These observations point to the **possibility of using genetic profiling to inform clinical practice** in significantly larger groups of individuals than for whom monogenic cardiovascular variants are considered. As a result of exponential increases in the proportion of individuals with broad genetic profiling, **cardiovascular PRSs are beginning to enter clinical practice**. Such PRSs may be appropriately considered in select scenarios, given the current evidence base.”*

# Potential relevance of PRS in clinical practice

Example: coronary artery disease



**Figure 3.** Predictive ability of polygenic risk scores for coronary artery disease.

- PRS has higher risk stratification ability than conventional risk factors
- PRS & conventional risk factors leads to improvement

# Potential clinical utility of PRS for cardiovascular disease

Disease/risk factor	Potential clinical utility of PRS
CAD	Earlier identification for lifestyle therapies and statins, potentially for those with very high CAD PRSs Earlier screening for subclinical atherosclerosis to time the initiation of pharmacotherapies Use as a risk-enhancing factor for primary prevention in middle-aged patients at borderline-intermediate 10-y ASCVD risk
AF	Earlier AF detection and resultant prophylactic anticoagulation, potentially with monitoring devices Rigorous control of additive clinical risk factors for AF
T2D	Earlier lifestyle modification Potential consideration of prophylactic hypoglycemic medications with concomitant additional T2D clinical risk factors Genomic stratification may optimize hypoglycemic choice
VTE	Rigorous VTE risk-reducing strategies in the context of high-risk scenarios (prolonged travel, major surgery, etc)
Hypercholesterolemia	Earlier institution and earlier uptitration of lipid-lowering pharmacotherapies analogous to FH
Pharmacogenomics	Personalized drug therapy regimens that increase drug efficacy and decrease toxicities, eg, personalized $\beta$ -blocker target dose in patients with HFrEF or the prevention of drug-induced QT prolongation

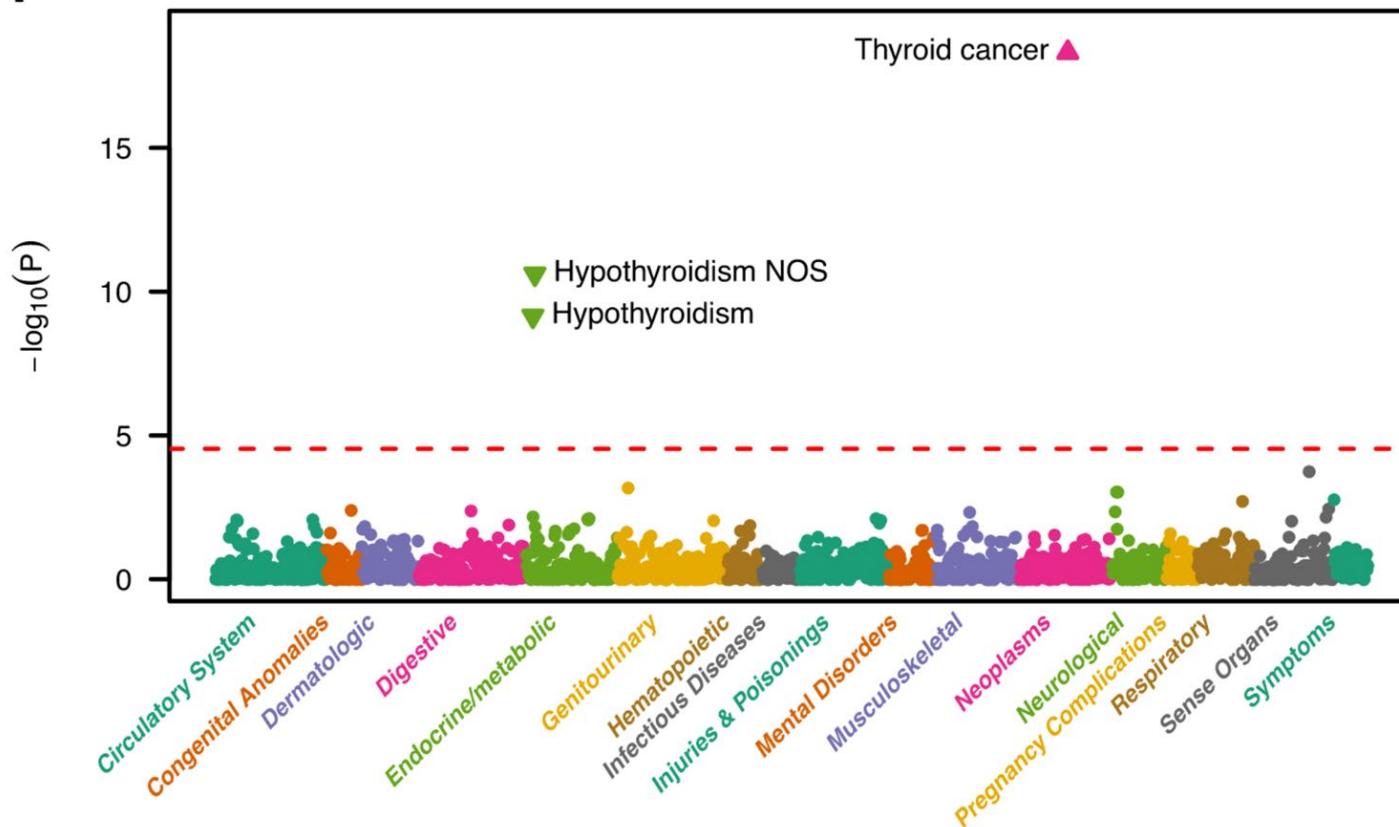
AF indicates atrial fibrillation; ASCVD, atherosclerotic cardiovascular disease; CAD, coronary artery disease; FH, familial hypercholesterolemia; HFrEF, heart failure with reduced ejection fraction; PRS, polygenic risk score; T2D, type 2 diabetes; and VTE, venous thromboembolism. Lone AF refers to AF in the absence of other cardiovascular risk factors (typically in young adults).

- Early-stage identification/intervention, Risk stratification, ...

# PGS is a useful tool for research

Cancer PRS model shows pleiotropic association with non-cancer traits

F



Evaluate the observed phenotypic enrichments of all patients with high cancer PRS

1. Start with PRS score
2. Rank patients
3. Find phenotypic enrichments for those patients
4. Method: ROC  
x-axis: Cancer PRS score  
y-axis: %people with trait
5. Take significance, plot it on this graph here

PRS-PheWAS analysis, assessing genetic correlation between traits

# PGS is a useful tool for research

## LETTERS

<https://doi.org/10.1038/s41591-020-0785-8>

nature  
medicine

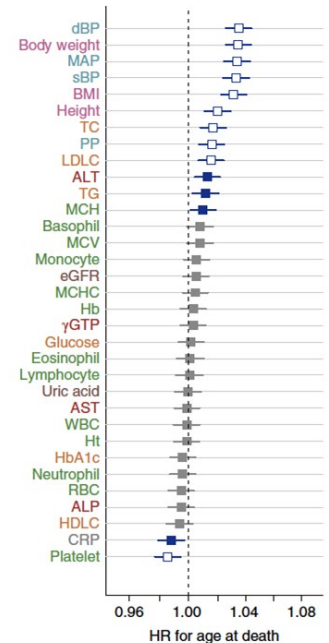
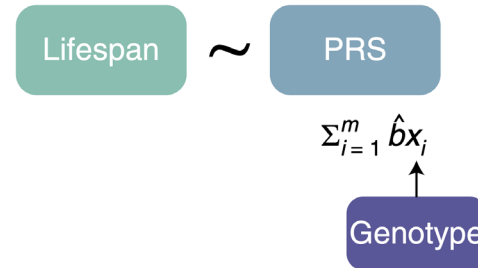
Check for updates

## Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan

Saori Sakaue<sup>1,2,3,18</sup>, Masahiro Kanai<sup>1,2,4,5,6,7,8,18</sup>, Juha Karjalainen<sup>4,5,6,8</sup>, Masato Akiyama<sup>1,9</sup>, Mitja Kurki<sup>14,5,6,8</sup>, Nana Matoba<sup>1</sup>, Atsushi Takahashi<sup>1,10</sup>, Makoto Hirata<sup>11</sup>, Michiaki Kubo<sup>12</sup>, Koichi Matsuda<sup>13</sup>, Yoshinori Murakami<sup>14</sup>, FinnGen, Mark J. Daly<sup>4,5,6,8</sup>, Yoichiro Kamatani<sup>1,15</sup> and Yukinori Okada<sup>2,16,17</sup> ✉

PGS(biomarker) associations with lifespan (age at death)  
Death might affect phenotypes measured, but PRS of those phenotypes can correlate with age at death more 'cleanly'

## Association of PRS with lifespan



nature  
medicine

## ARTICLES

<https://doi.org/10.1038/s41591-022-01957-2>

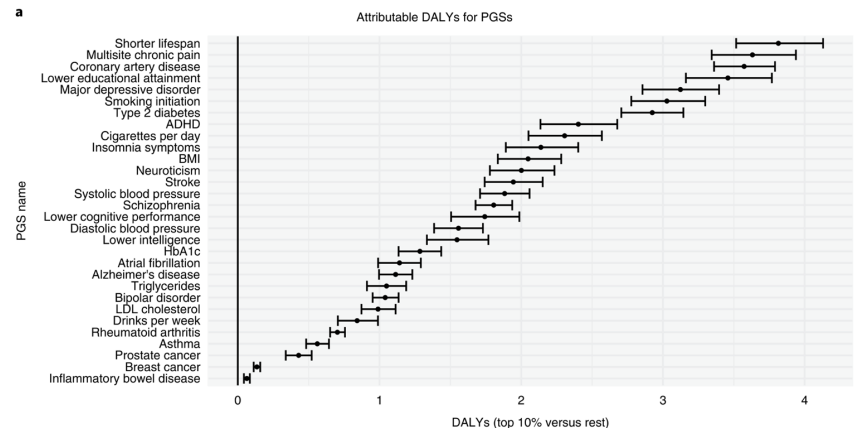
Check for updates

## OPEN

## Genetic risk factors have a substantial impact on healthy life years

Sakari Jukarainen<sup>1</sup> ✉, Tuomo Kiiskinen<sup>1,2,3</sup>, Sara Kuitunen<sup>1,2</sup>, Aki S. Havulinna<sup>1,2</sup>, Juha Karjalainen<sup>1,3,4</sup>, Mattia Cordioli<sup>1</sup>, Joel T. Rämö<sup>1</sup>, Nina Mars<sup>1</sup>, FinnGen<sup>1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100</sup>, Kaitlin E. Samocha<sup>3,5</sup>, Hanna M. Ollila<sup>1,3,5,6</sup>, Matti Pirinen<sup>1,7,8</sup> and Andrea Ganna<sup>1,3,4</sup> ✉

## PGS associations with disability adjusted life years (DALY)

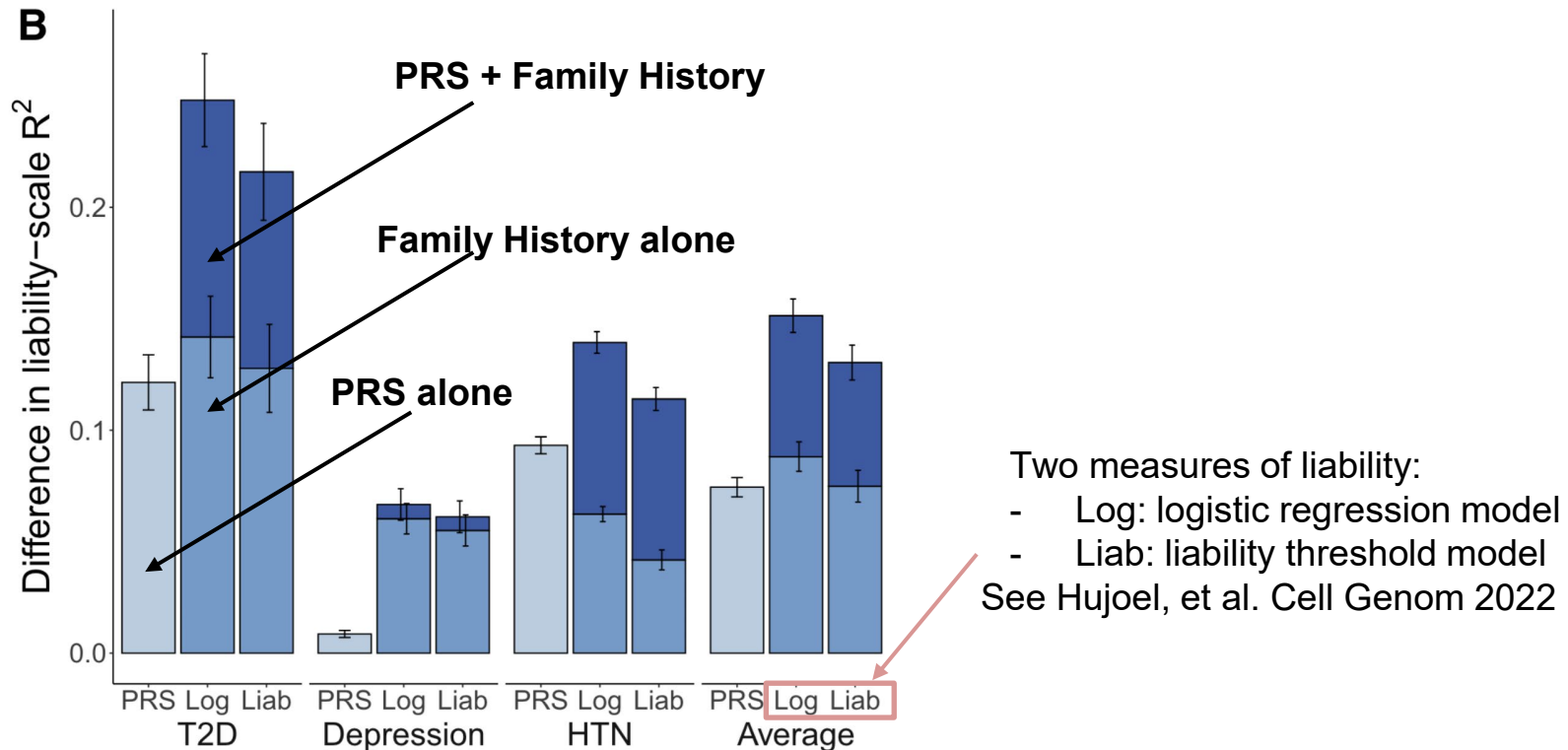




# Family history (FH) complements PGS

## Risk factors not captured in PGS

- Rare variants with large effects
  - Sample size & statistical power limitation in PGS
- Environmental factors

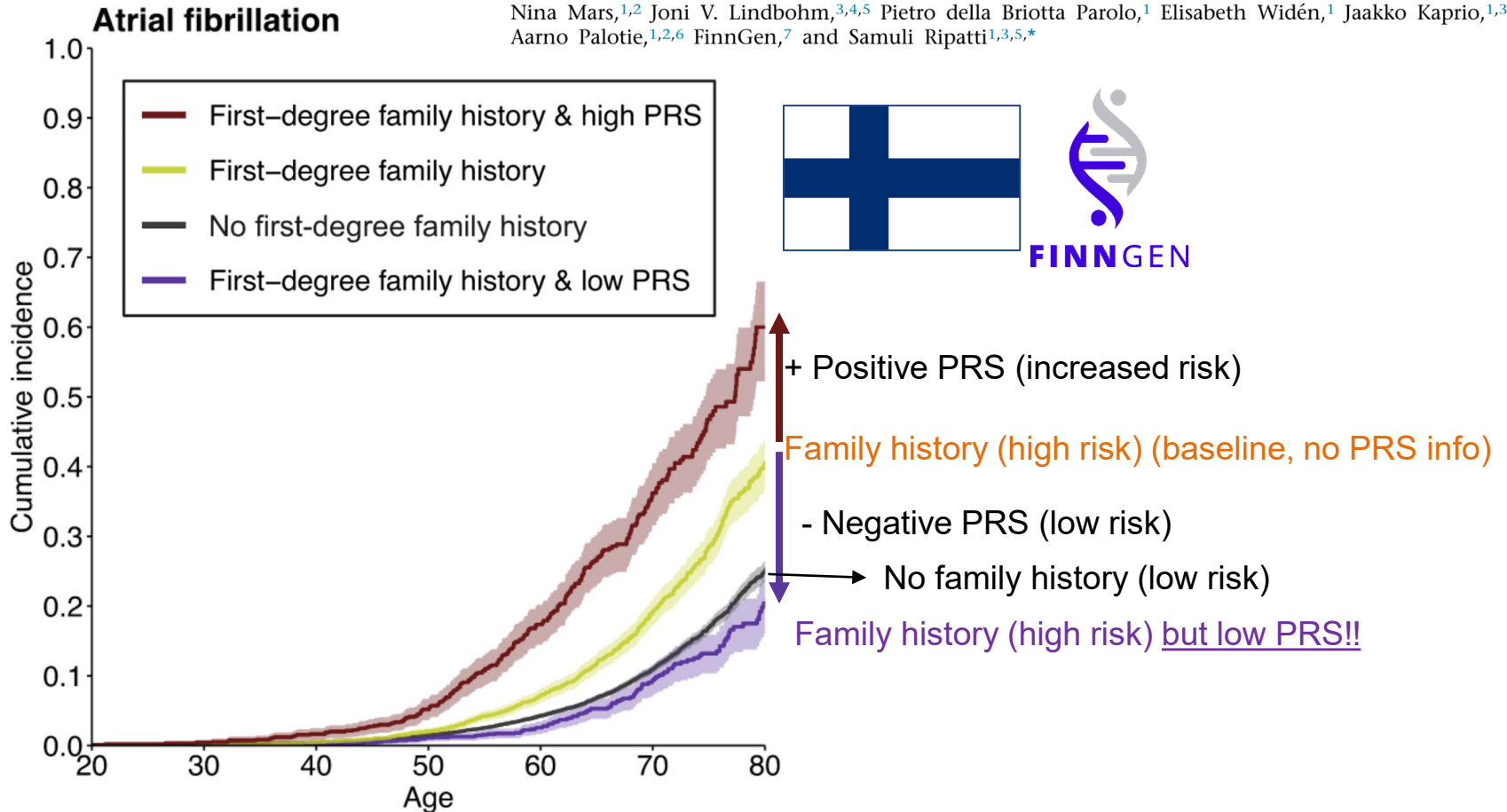


# Family history (FH) complements PGS

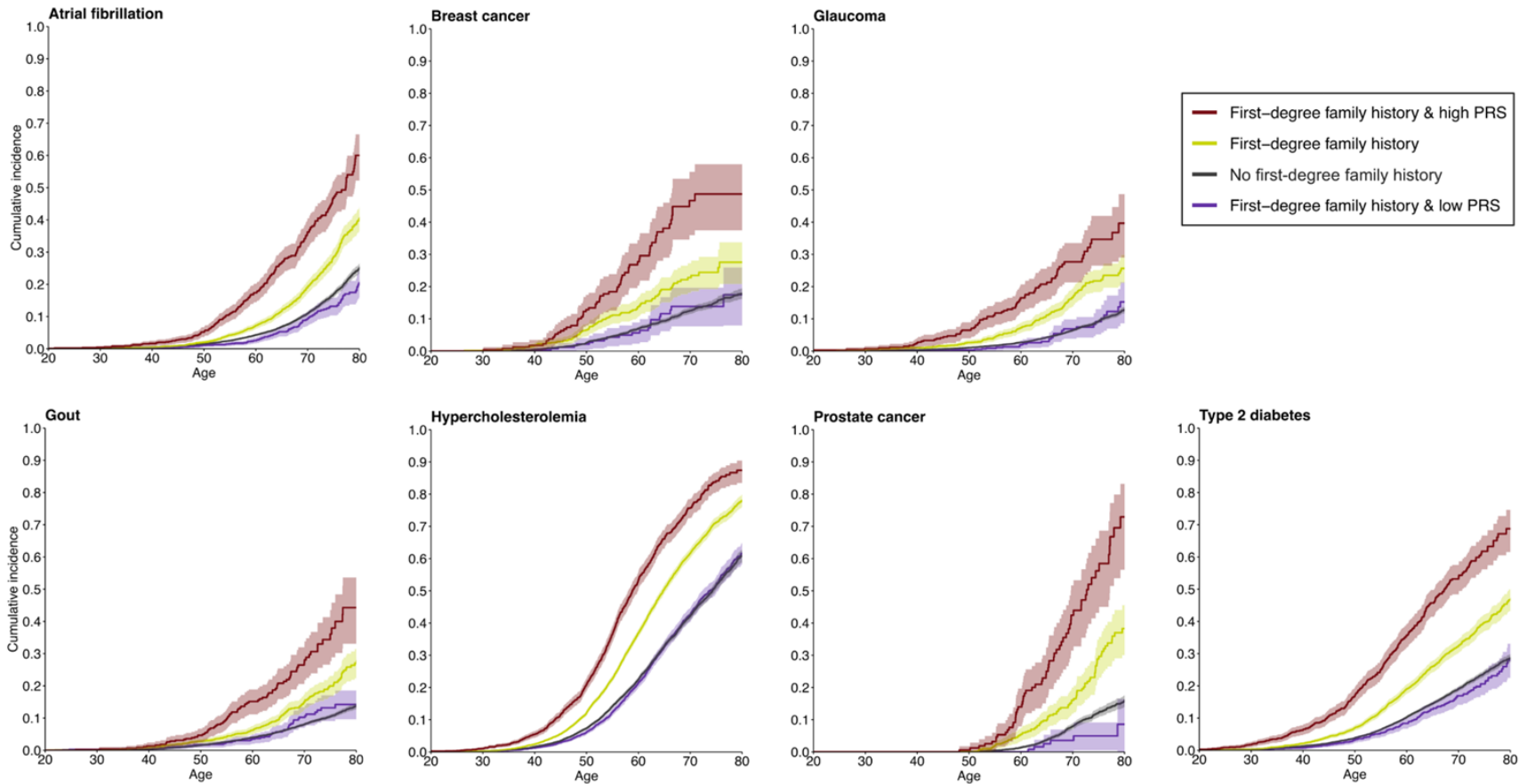
## ARTICLE

### Systematic comparison of family history and polygenic risk across 24 common diseases

Nina Mars,<sup>1,2</sup> Joni V. Lindbohm,<sup>3,4,5</sup> Pietro della Briotta Parolo,<sup>1</sup> Elisabeth Widén,<sup>1</sup> Jaakko Kaprio,<sup>1,3</sup> Aarno Palotie,<sup>1,2,6</sup> FinnGen,<sup>7</sup> and Samuli Ripatti<sup>1,3,5,\*</sup>



# Family history (FH) complements PGS



# Summary 1: Polygenic score (PGS) introduction

---

- GWAS revealed large number of common variants contribute to complex traits; the individual effects of variants are small
- Polygenic scores (PGS) combine effects of disease-associated alleles for each individual
- PGS has potential relevance for clinical applications for some traits and for some populations
- PGS would be useful for research
- Current PGS models captures incomplete genetic liability of disease and PGS and family history are complementary to each other

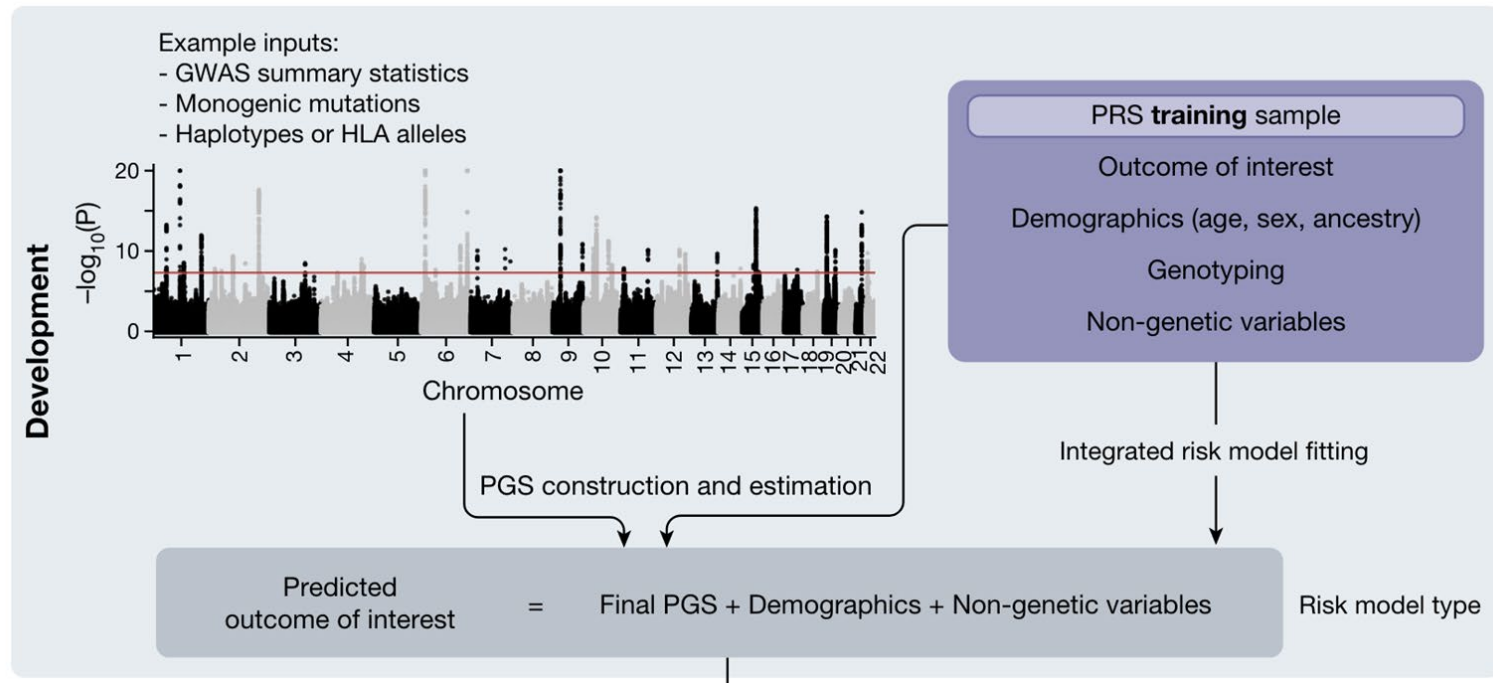
# Overview: Genetic prediction of complex traits

---

1. Foundations of Human Genetic Variation
2. Polygenic score (PGS) introduction
3. PGS Evaluation
4. Methods to fit PGS model
5. Challenges and opportunities in PGS research

# PGS development and validation process

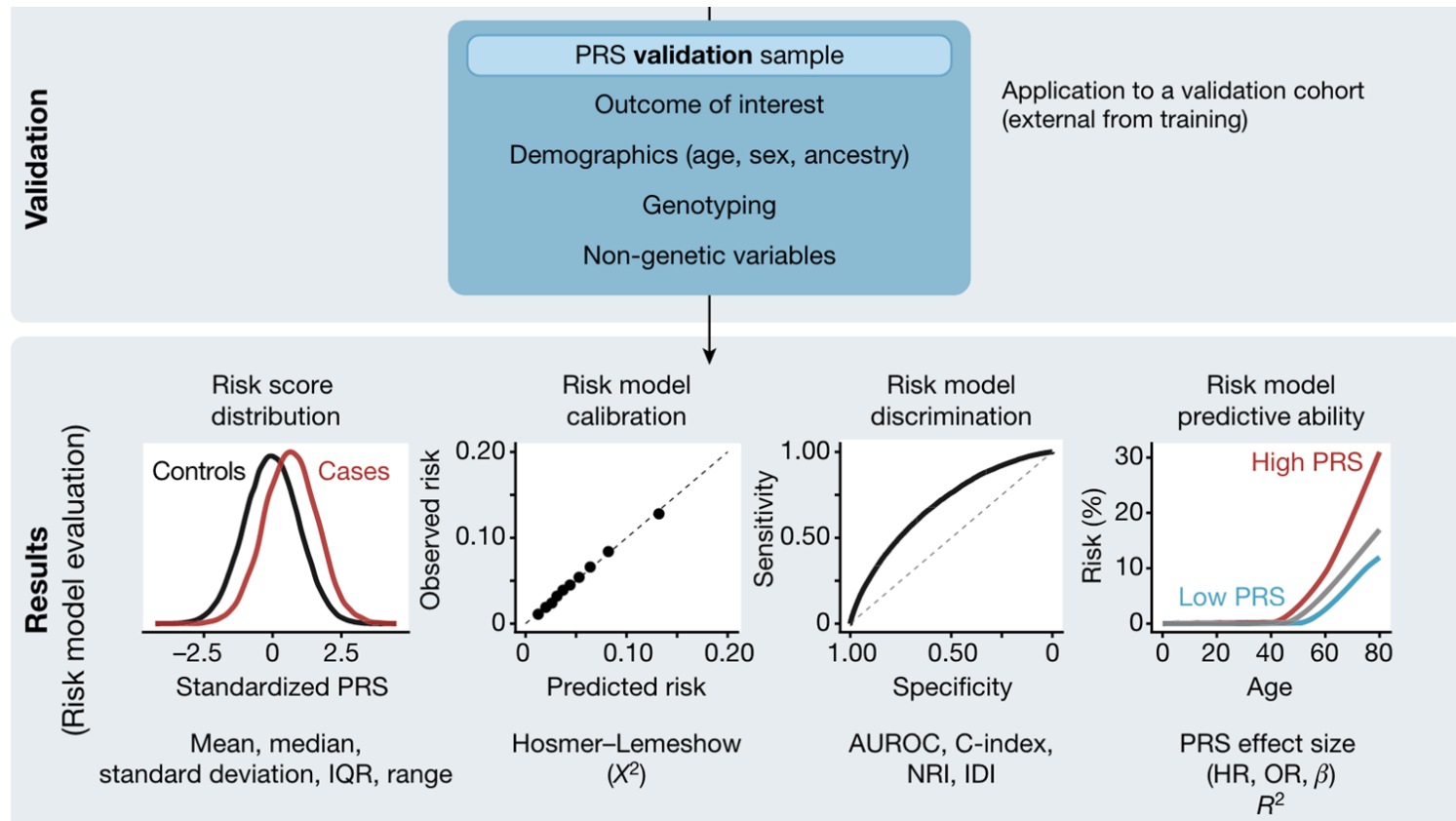
## 1. PGS development



- age, sex, demographics (genotype PCs) are typically considered as covariates

# PGS development and validation process

## 2. Evaluation and validation of the PGS model



PRS alone only gives a relative genetic burden score, but not an absolute risk for an individual

For a new cohort, need to calibrate predictive value of PRS score. can calculate for each person, but need a 'translation table' to get actual risk for that cohort

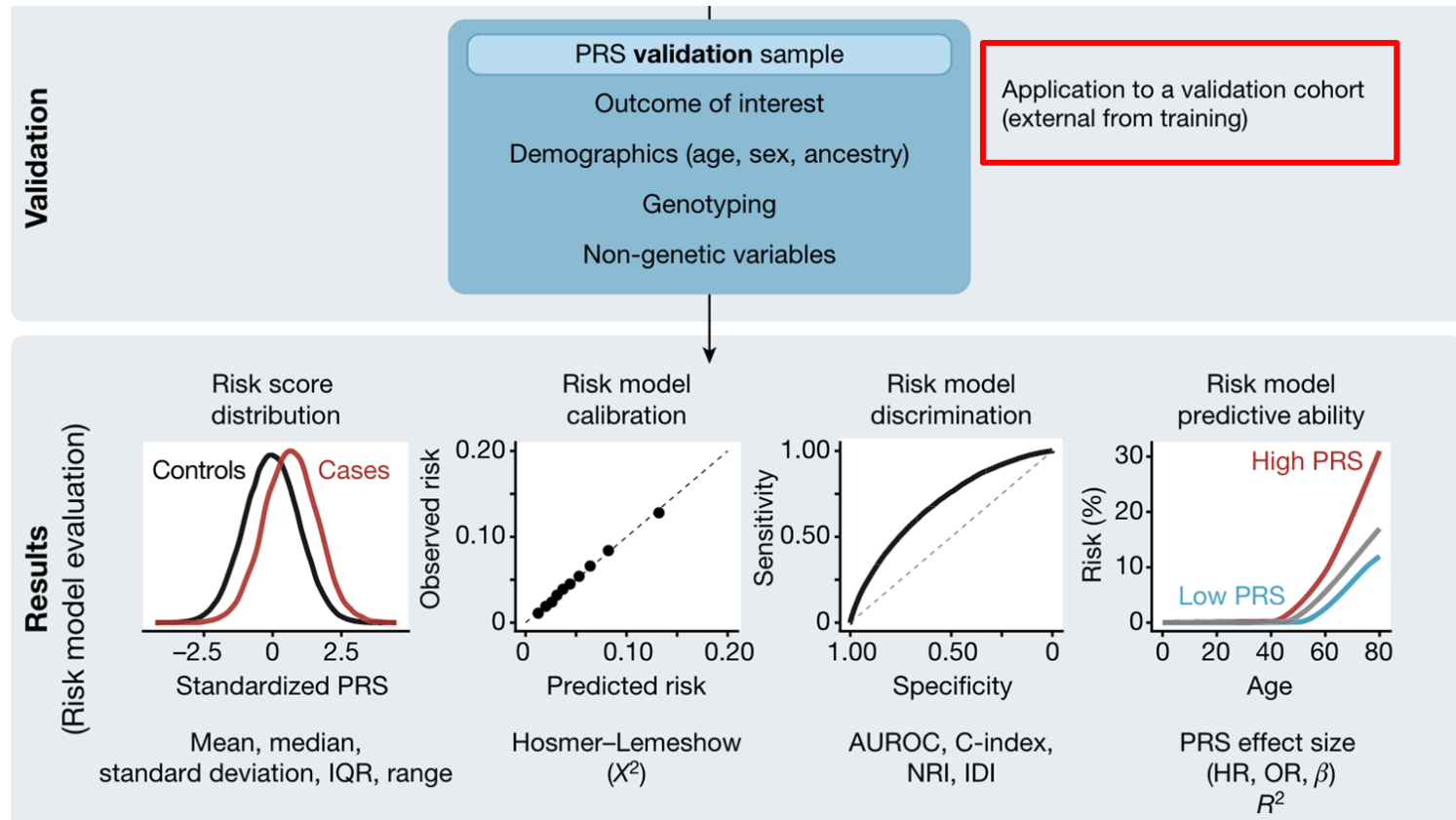
Then for that cohort, you can evaluate the overall discrimination strength of PRS score

You can then combine it with other risk factors (such as age) to get increased predictive power when combining PRS.

→ Age-matched risk from PRS

# PGS development and validation process

## 2. Evaluation and validation of the PGS model



Use **hold-out test set** or **external validation set** when evaluating the predictive performance of PGS models



# **Heritability ( $h^2$ ) – the theoretical upper bound of predictive performance for quantitative traits**

---

- Complex trait (T) = Genetics (G) + Environment (E) + GxE interaction
- Let's consider the variance of the observed trait ( $\sigma_T^2$ )
- Under a simple scenario: T = G + E (no GxE interaction)
  - $\sigma_T^2 = \sigma_G^2 + \sigma_E^2$
  - $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$ 
    - A: additive effects
    - D: non-additive effects (dominance, recessive, etc.)
    - I: interaction effects

# Heritability ( $h^2$ ) – the theoretical upper bound of predictive performance for quantitative traits

- Complex trait (T) = Genetics (G) + Environment (E) + GxE interaction
- Let's consider the variance of the observed trait ( $\sigma_T^2$ )
- Under a simple scenario: T = G + E (no GxE interaction)
  - $\sigma_T^2 = \sigma_G^2 + \sigma_E^2$
  - $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$ 
    - A: additive effects
    - D: non-additive effects (dominance, recessive, etc.)
    - I: interaction effects
- **[Definition] Heritability**
  - $H^2$  (Broad-sense heritability) =  $\sigma_G^2 / \sigma_T^2$
  - $h^2$  (narrow-sense heritability) =  $\sigma_A^2 / \sigma_T^2$
  - Heritability: fraction of phenotypic variance explained by (additive) genetic effects

# Some notes on heritability ( $h^2$ )

---

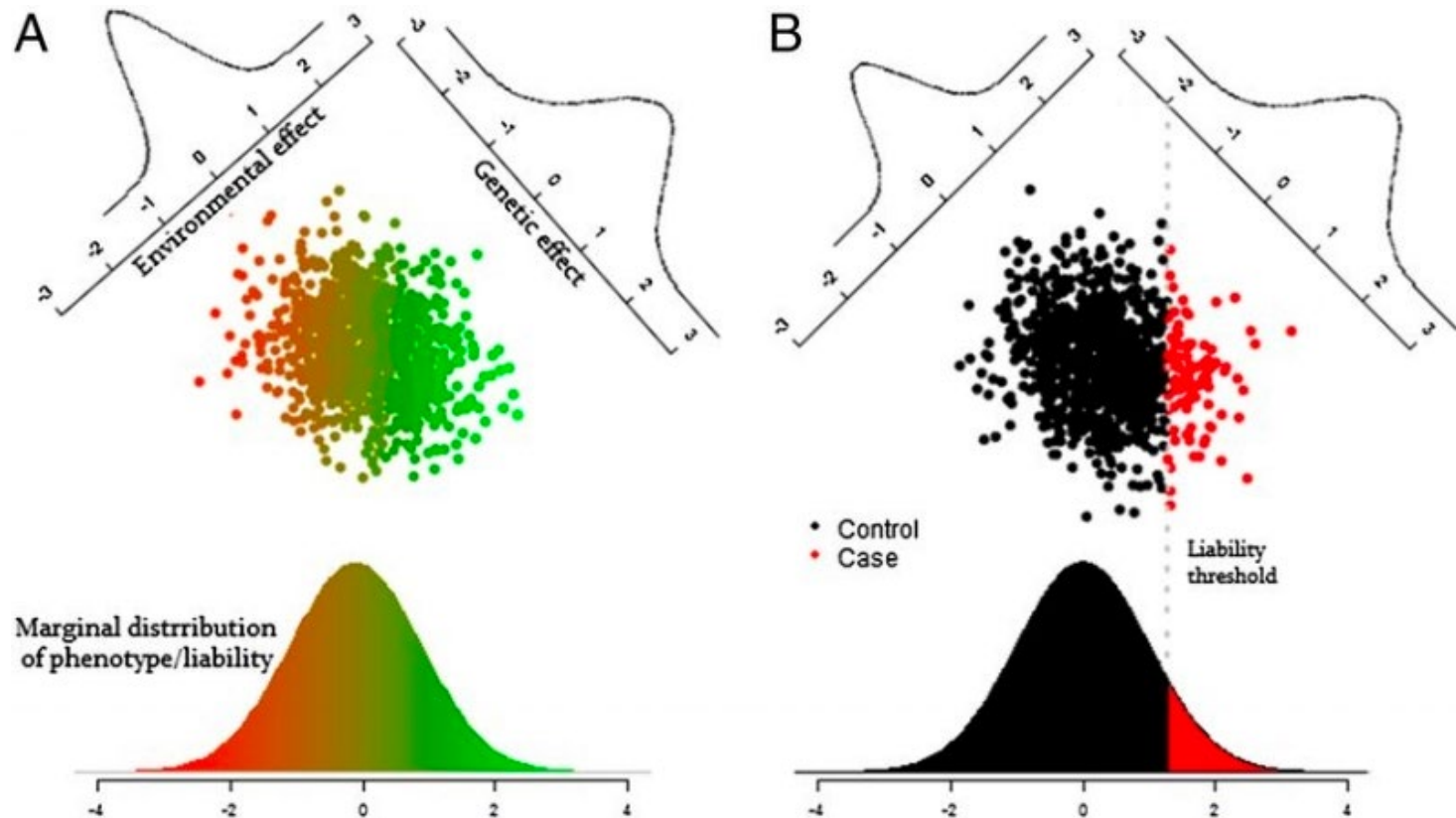
- Heritability is **not directly observable** and is often estimated by statistical model (typically from twin studies, more recently GWAS)
  - Phenotypic variance depends on the population of the study
- Heritability is a population-level, not individual-level, parameter
  - It does NOT inform the level of genetic influence on a trait for one particular individual
  - It does NOT inform the individual-level predictive accuracy/reliability of polygenic prediction
  - See Visscher et al., *Nat Rev Gen* (2008) for common pitfalls
- Heritability estimates for binary traits (observed- vs. liability-scale)
  - Using GWAS data, one can compute observed-scale heritability
  - Observed-scale heritability depends on the fraction of observed cases and disease prevalence. Need to control for ascertainment bias in GWAS discovery cohort = Use cumulative density function + prevalence (next slides)
  - Observed-scale heritability vs. Liability-scale heritability

$$h_{liability}^2 = h_{observed}^2 \frac{K(1-K)}{\varphi(\Phi^{-1}[K])^2} \frac{K(1-K)}{P(1-P)}$$

$K$ =disease prevalence in population  
 $P$ =disease prevalence in GWAS set  
 $\Phi$ =cumulative density of Normal distr.

# Liability and threshold model for binary traits

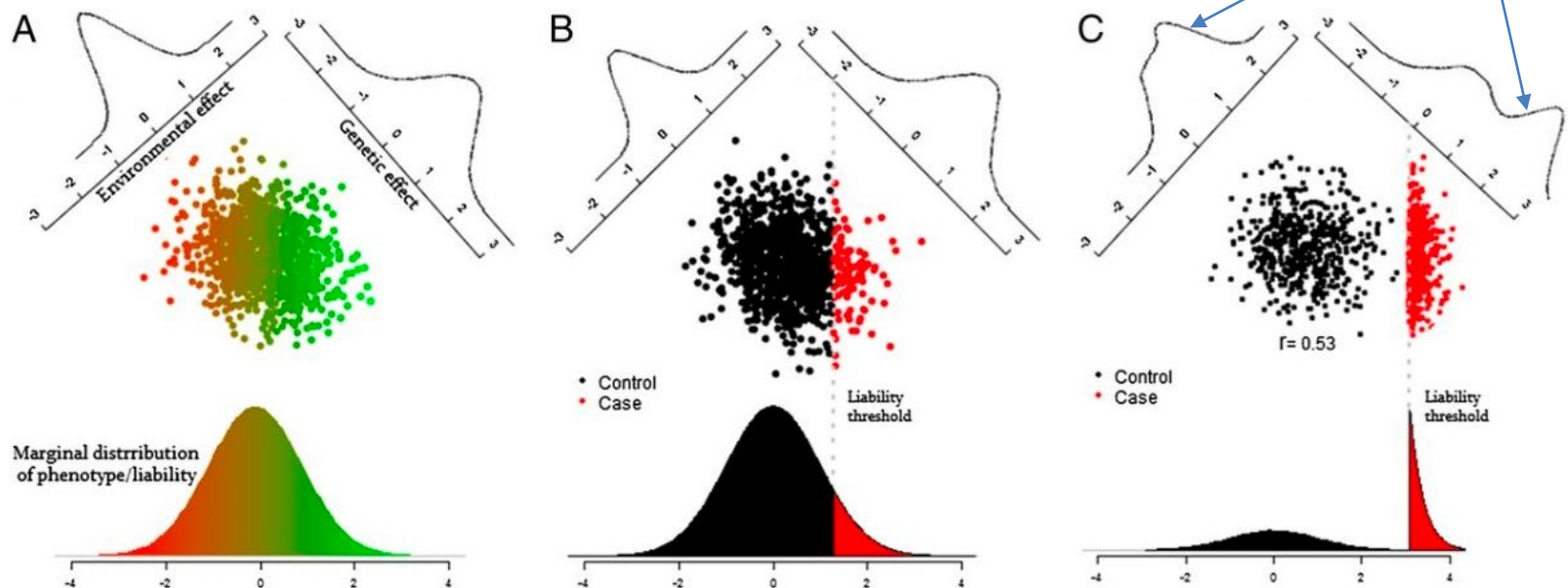
- Assume the continuous distribution of liability. Consider our observed cases are the one passing the liability threshold



# Liability and threshold model for binary traits

- Assume the continuous distribution of liability. Consider our observed cases are the one passing the liability threshold
- We may consider the heritability on the liability scale
  - Observed-scale vs. liability-scale
- In case-control GWAS, we may have overrepresentation of case samples.

This is why we need to adjust observed scale to liability scale

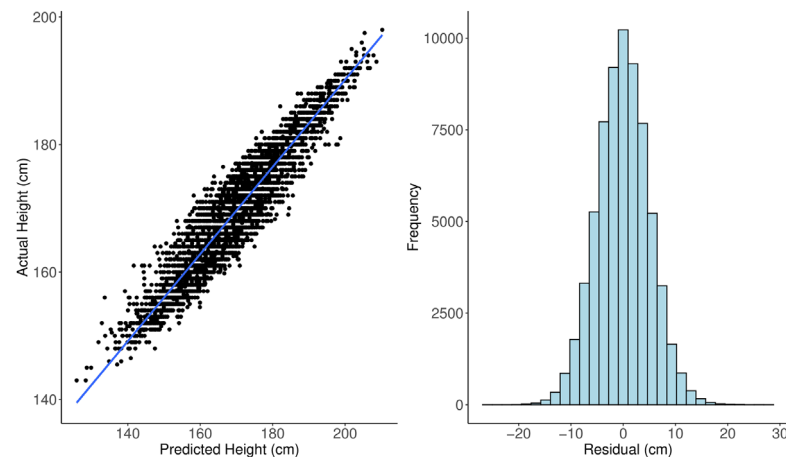
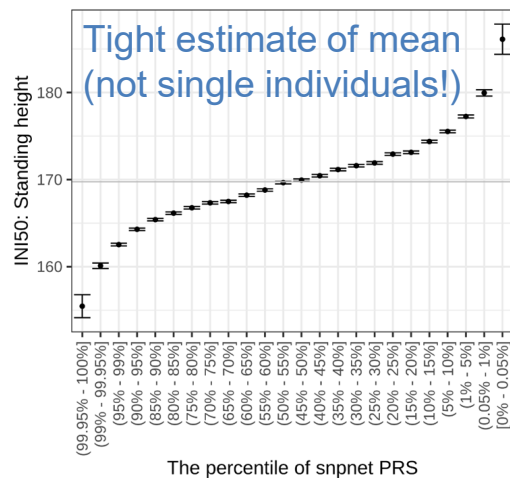
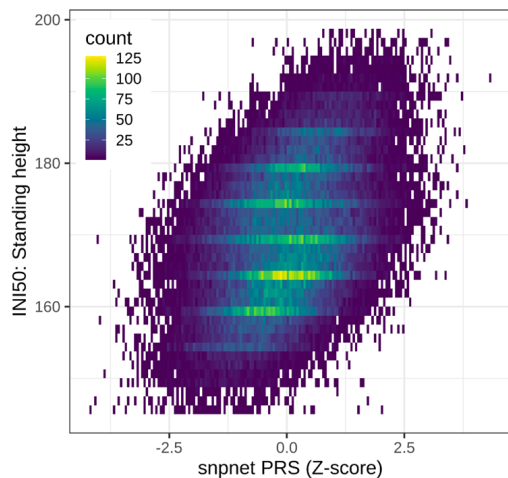


# PGS evaluation - $R^2$ for quantitative traits

PGS evaluation:  $R^2$  is a common metric for quantitative traits

Example: predicting standing height in UK Biobank with snpnet

hold-out test set  $R^2$ : 0.178 (PGS alone), 0.717 (PGS + covariates)



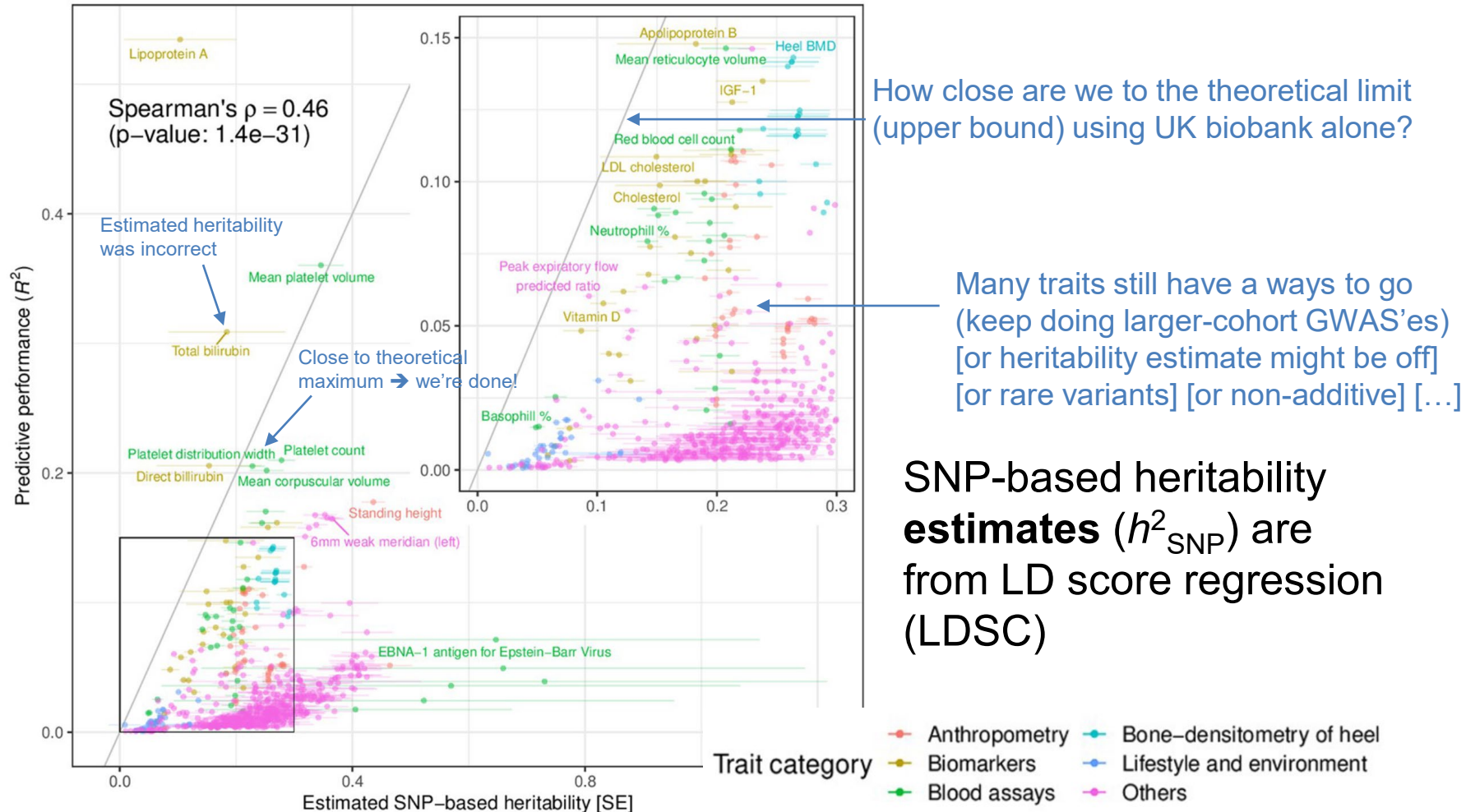
PGS alone gives  $R^2=0.178$

Using sex + age + 10 genotype PCs as covariates  
Subset of 10,000 individuals  
→ Very high accuracy prediction (0.717)

Using 330k people from UK biobank: 270k train + 60k test

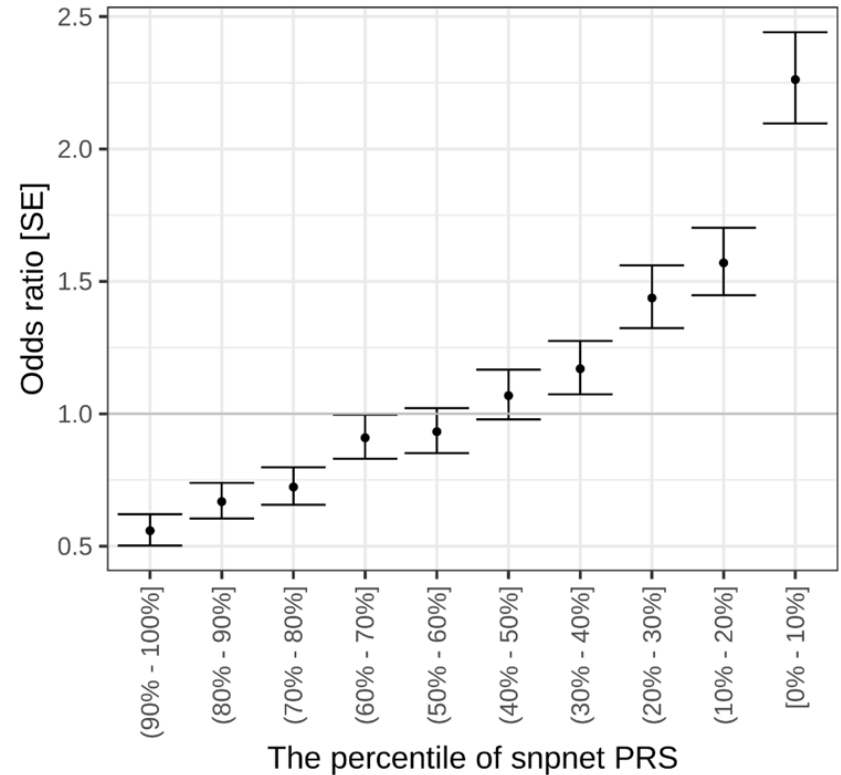
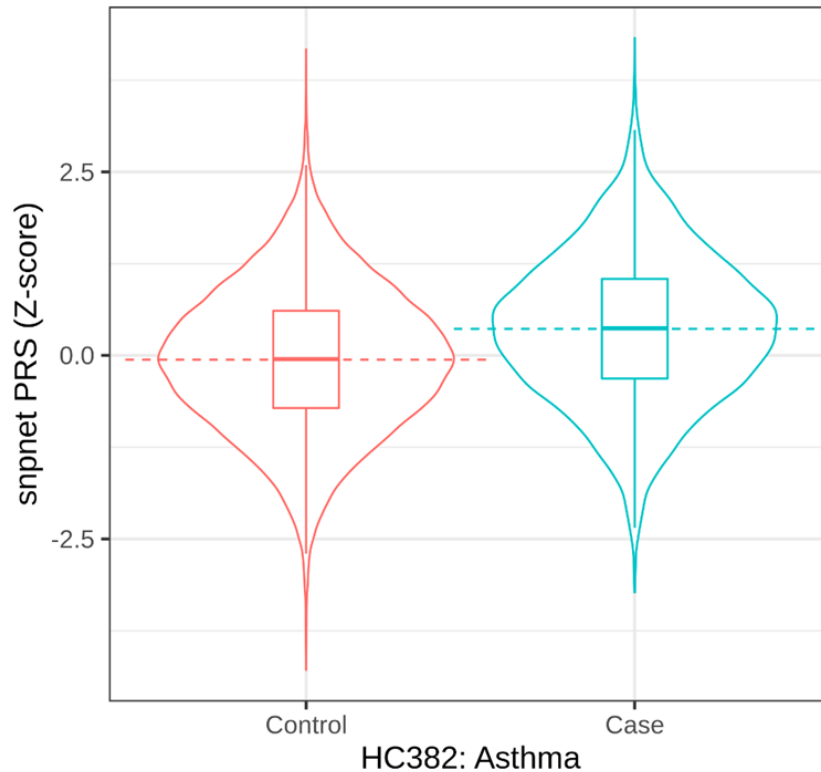
# SNP heritability $h^2$ is the upper bound of the PGS predictive performance

Comparison of  $R^2$  vs.  $h^2_{\text{SNP}}$  for quantitative traits in UK Biobank



# PGS evaluation for binary traits

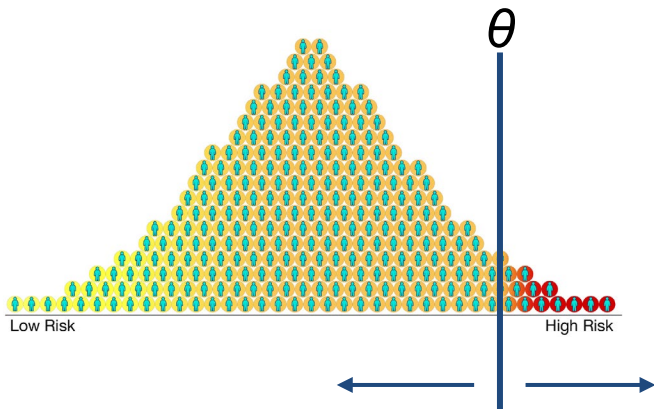
Example: asthma in UK Biobank



- Predictive performance: AUROC, observed-scale pseudo- $R^2$ , liability-scale pseudo- $R^2$ , ...



# Area under the receiver-operator curve (AUC or AUROC)

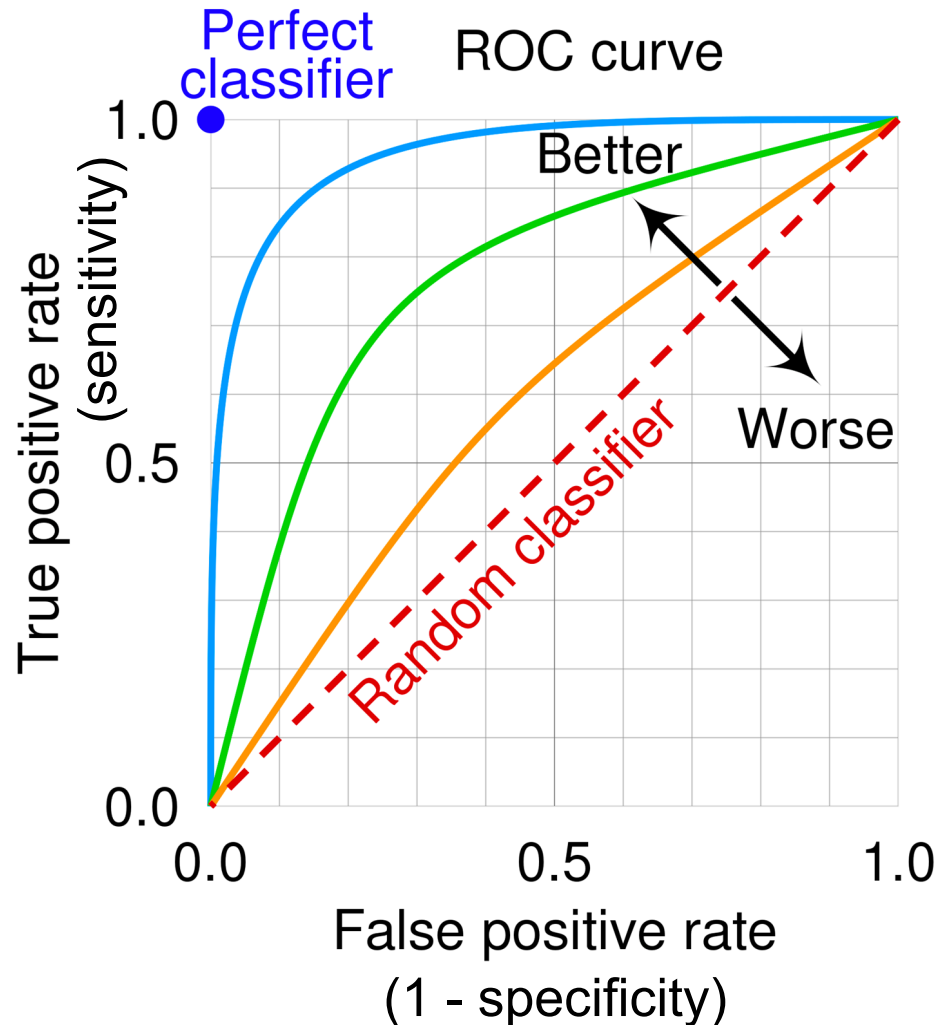


Predicted label

Actual label

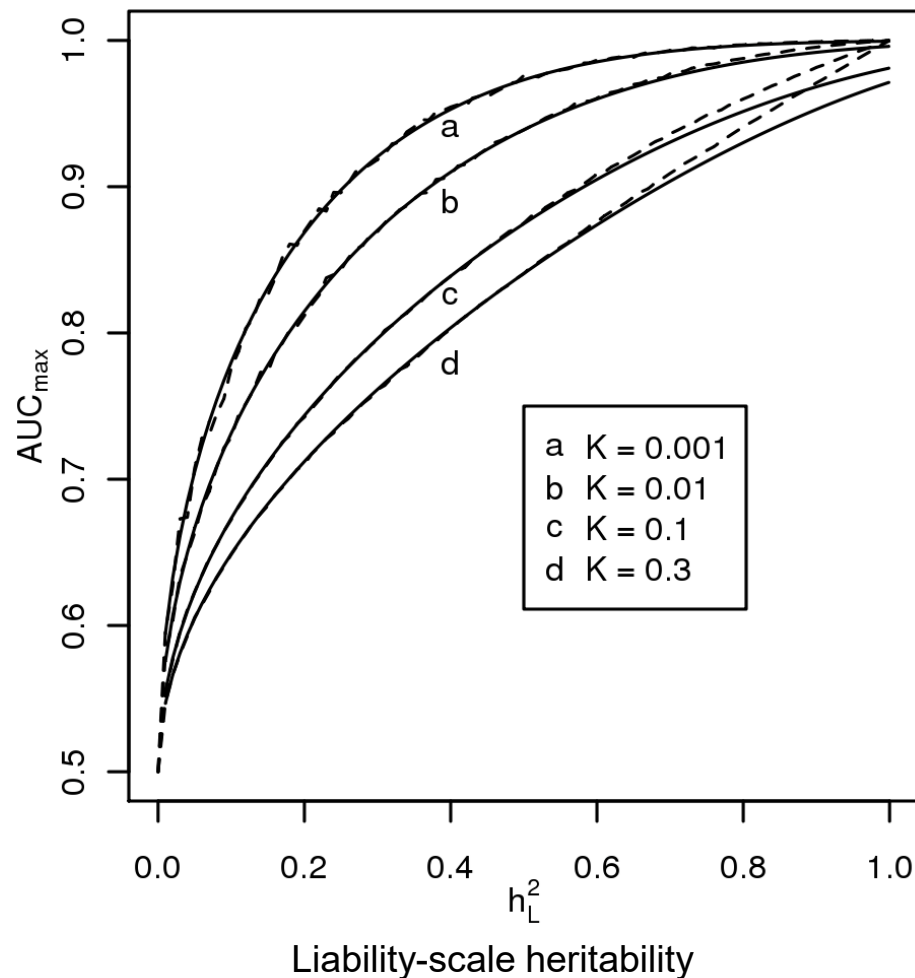
	PP	PN
P	TP	FN
N	FP	TN

- True positive rate =  $TP / P$
- False positive rate =  $FP / N$   
=  $1 - TN / N$



# Max AUC for genetic risk prediction depends on heritability and disease prevalence

AUC is calculated on the observed-scale and depends on disease parameters



# Pseudo- $R^2$ as a goodness of fit for binary traits

---

- AUROC is not the only metric
  - Cox and Shell's pseudo- $R^2$  (based on likelihood)
  - Nagelkerke's pseudo- $R^2$  (aka Cragg and Uhler's pseudo- $R^2$ )
    - Normalized C&S pseudo- $R^2$  so that the maximum reaches 1
- 

Brief description

Notation and formula

---

$R^2$  on the observed scale

$$R_o^2 = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Cox and Snell's  $R^2$  on the observed scale

$$R_{C\&S}^2 = 1 - \left\{ \frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right\}^{2/N}$$

Nagelkerke's  $R^2$  on the observed scale

$$R_N^2 = \frac{R_{C\&S}^2}{1 - (\text{Likelihood}_{\text{null}})^{2/N}}$$

---

# Pseudo- $R^2$ as a goodness of fit for binary traits

---

- AUROC is not the only metric
  - Cox and Shell's pseudo- $R^2$  (based on likelihood)
  - Nagelkerke's pseudo- $R^2$  (aka Cragg and Uhler's pseudo- $R^2$ )
    - Normalized C&S pseudo- $R^2$  so that the maximum reaches 1
- 

Brief description

Notation and formula

---

$R^2$  on the observed scale

$$R_o^2 = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Cox and Snell's  $R^2$  on the observed scale

$$R_{C\&S}^2 = 1 - \left\{ \frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right\}^{2/N}$$

Nagelkerke's  $R^2$  on the observed scale

$$R_N^2 = \frac{R_{C\&S}^2}{1 - (\text{Likelihood}_{\text{null}})^{2/N}}$$

---

# Liability-scale Pseudo- $R^2$ has expectation of $h_l^2$

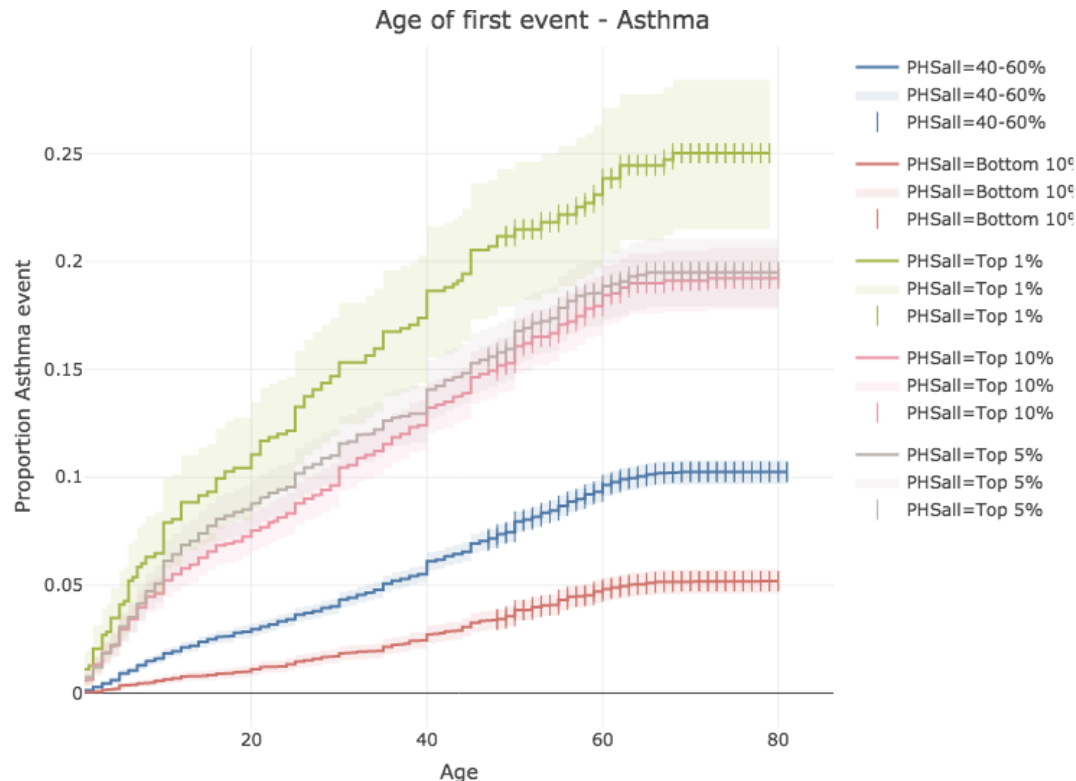
**TABLE I. Brief description of  $R^2$  measures used in this study and their theoretical expectation**

Brief description	Notation and formula	Expectation
$R^2$ on the observed scale	$R_o^2 = 1 - \frac{\sum_i^N (y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$	$h_l^2 \frac{z^2}{K(1-K)}$
Cox and Snell's $R^2$ on the observed scale	$R_{C\&S}^2 = 1 - \left\{ \frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right\}^{2/N}$	$h_l^2 \frac{z^2}{K(1-K)}$
Nagelkerke's $R^2$ on the observed scale	$R_N^2 = \frac{R_{C\&S}^2}{1 - (\text{Likelihood}_{\text{null}})^{2/N}}$	$\frac{R_{C\&S}^2}{1 - K^{2K} \cdot (1-K)^{2(1-K)}}$
$R^2$ on the liability scale	$R_l^2 = R_o^2 \frac{\hat{K}(1-\hat{K})}{z^2}$	$h_l^2$
$R^2$ on the probit liability scale	$R_{\text{probit}}^2 = \frac{\text{var}(\hat{b}_{\text{probit}}g_i)}{\text{var}(\hat{b}_{\text{probit}}g_i)+1}$	$h_l^2$
$R^2$ on the logit liability scale	$R_{\text{logit}}^2 = \frac{\text{var}(\hat{b}_{\text{logit}}g_i)}{\text{var}(\hat{b}_{\text{logit}}g_i)+3.29}$	$h_l^2$
$R^2$ on the liability scale using AUC	$R_{\text{AUC}}^2 = \frac{2Q^2}{(m_2 - m)^2 + Q^2 m(m-t) + m_2(m_2 - t)}$	$h_l^2$
$R^2$ on the liability scale when using ascertained case-control studies	$R_{\text{tcc}}^2 = \frac{R_{\text{cc}}^2 C}{1 + R_{\text{cc}}^2 \theta C}$	$h_l^2$

$y$ , observations that are 0 or 1 for unaffected and affected individuals;  $h_l^2$ , heritability on the liability scale, in this context the proportion of variance on the liability scale explained by the genetic profile;  $K$ , population prevalence;  $z$ , the height of a normal density curve at the point according to  $K$ ;  $g$ , the sum of all additive genetic factors in the estimated genetic predictor;  $b$ , regression coefficient from generalized linear model;  $m$ , the mean liability for cases;  $m_2$ , the mean liability for controls;  $t$ , the threshold on the normal distribution that truncates the proportion of disease prevalence  $K$ ;  $Q$ , the inverse of the cumulative density function of the normal distribution up to values of AUC;  $C$  and  $\theta$ , correcting factors for ascertainment.

# Polygenic hazard score for genetic liability of disease onset prediction (Cox model)

- Cox proportional Hazard ratio model
- Hazard ratio or C-index are commonly used metric for evaluation
- C-index: fraction of the accurately predicted ordering of the events. See Harrell, et al. (1982), Li and Tibshirani (2019)



## Summary 2: PGS Evaluation

---

- Genetics plays a partial role: Complex trait (T) = Genetics (G) + Environment (E) + GxE interaction
- Heritability := fraction of phenotypic variation explained by genetics in a population
- Use hold-out test set or external validation set to evaluate the predictive performance of PGS
- Commonly used metrics:
  - Quantitative traits:  $R^2$
  - Binary traits: pseudo- $R^2$  (observed, liability), AUROC (observed)
  - Time-to-event traits: Hazard ratio, C-index

# Genetic prediction of complex traits

---

1. Foundations of Human Genetic Variation
2. Polygenic score (PGS) introduction
3. PGS Evaluation
4. Methods to fit PGS model
5. Challenges and opportunities in PGS research



# How to train PGS models?

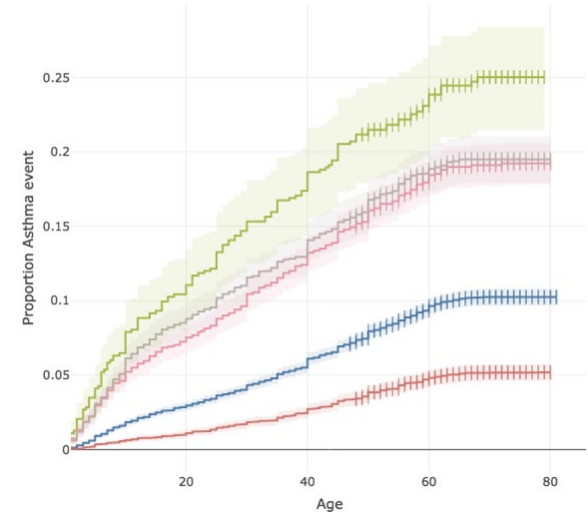
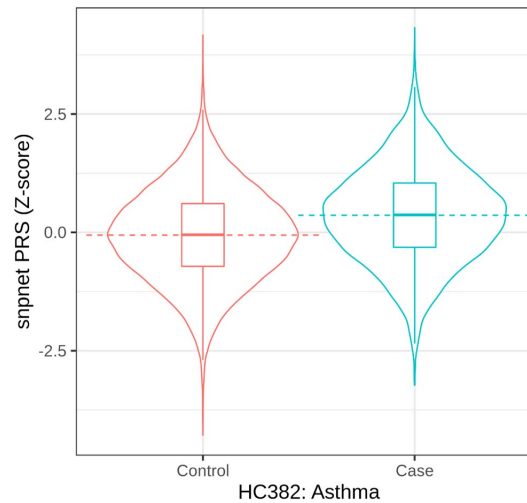
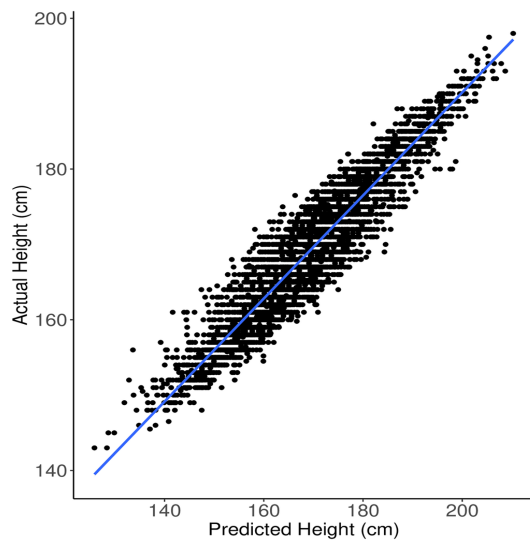
---

- Polygenic score:  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$ 
  - $i$ -th individual
  - $j$ -th variant
  - $G$ : genotype
  - $\beta$ : effect size
- Types of traits
  - Quantitative traits (e.g. biomarkers, anthropometry)
  - Binary traits (e.g. case-control)
  - Time-to-event traits (e.g. disease onset)

# How to train PGS models?

- Polygenic score:  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$   $i$ -th individual  $G$ : genotype  
 $j$ -th variant  $\beta$ : effect size

- Types of traits
  - Quantitative traits (e.g. biomarkers, anthropometry): **linear regression**
  - Binary traits (e.g. case-control): **logistic regression**
  - Time-to-event traits (e.g. disease onset): **Cox model** (time to event, proportional hazard ratio model)



# How to train PGS models?

---

- Polygenic score:  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$ 
  - $i$ -th individual
  - $j$ -th variant
  - $G$ : genotype
  - $\beta$ : effect size
- Types of traits
  - Quantitative traits (e.g. biomarkers, anthropometry)
  - Binary traits (e.g. case-control)
  - Time-to-event traits (e.g. disease onset)
- To train PGS models:
  - Identify set of genetic variants in the model
  - Estimate effect size ( $\beta$ ) for each

# How to train PGS models?

---

- Polygenic score:  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$ 
  - $i$ -th individual
  - $j$ -th variant
  - $G$ : genotype
  - $\beta$ : effect size
- Types of traits
  - Quantitative traits (e.g. biomarkers, anthropometry)
  - Binary traits (e.g. case-control)
  - Time-to-event traits (e.g. disease onset)
- To train PGS models:
  - Identify set of genetic variants in the model
  - Estimate effect size ( $\beta$ ) for each
- PGS modeling approaches:
  - PGS model with genome-wide significant ( $p < 5e-8$ ) SNPs
  - P-value thresholding (P + T)
  - Bayesian approach that considers LD
  - PGS methods on individual-level data (BULP, snpnet, ...)

# How to train PGS models?

---

- Polygenic score:  $PRS_i = \sum_{j \in J} \beta_j G_{ij}$  *i*-th individual G: genotype  
*j*-th variant  $\beta$ : effect size

- Types of traits

- Quantitative traits (e.g. biomarkers, anthropometry)
- Binary traits (e.g. case-control)
- Time-to-event traits (e.g. disease onset)

- To train PGS models:

- Identify set of genetic variants in the model
- Estimate effect size ( $\beta$ ) for each

Active area of research with many proposed methods

- PGS modeling approaches:

- PGS model with genome-wide significant ( $p < 5e-8$ ) SNPs
- P-value thresholding (P + T)
- Bayesian approach that considers LD
- PGS methods on individual-level data (BULP, snpnet, ...)

# Genetic risk scores from GWAS significant SNPs

## Methods

### Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray,<sup>1,4</sup> Michael E. Goddard,<sup>2,3</sup> and Peter M. Visscher<sup>1</sup>

- Wray et al. proposed a method to predict disease risk with GWAS selected loci using simulation data.

ally on the simulated genotype ( $G$ ). For each of these individuals, we knew the true disease probability and estimated disease probability from the selected SNPs, calculated as,

$$P(D_i|G_i) = f_0 \prod_{j=1}^n \lambda_j^{*x_{ij}} \quad \text{and} \quad \hat{P}(D_i|G_i) = f_0 \prod_{j=1}^m \hat{\lambda}_j^{x_{ij}}$$

with  $n$  the total number of true risk loci,  $m$  the number of selected loci (both true and false),  $\hat{\lambda}_j$  the estimated RR for locus  $j$  from the case-control study, and  $x_{ij}$  the number of risk alleles for individual  $i$  at locus  $j$ . Note that the estimated risk will deviate

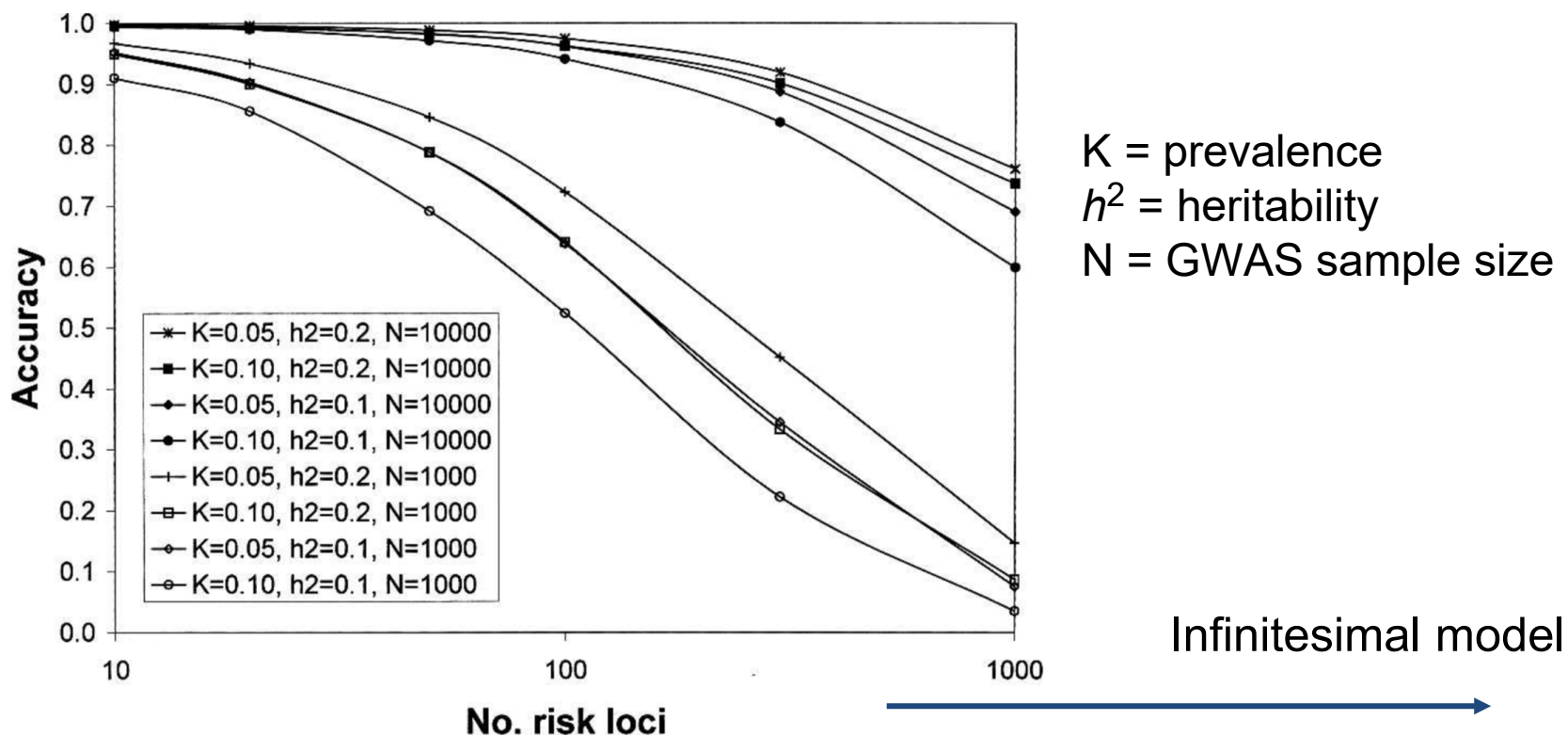
- Investigated how genetic architecture and disease parameters (prevalence and heritability) influence power

# Predictive accuracy of GWAS significant SNPs depends on genetic architecture

Methods

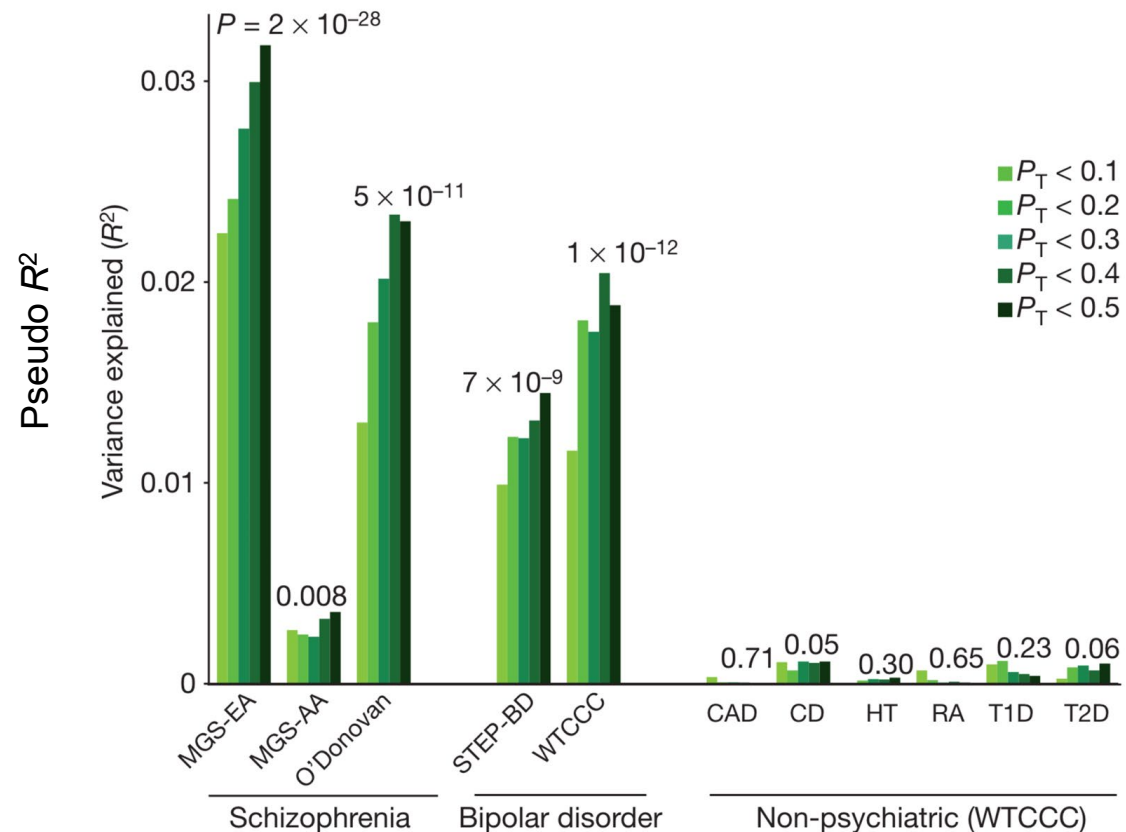
Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray,<sup>1,4</sup> Michael E. Goddard,<sup>2,3</sup> and Peter M. Visscher<sup>1</sup>



# Polygenic scores from GWAS ‘significant’ SNPs

- Schizophrenia GWAS meta-analysis (European, ~3300 cases)
- Tested “polygenic inheritance” hypothesis (Gottesman & Shields, 1967)
- Polygenic component with liberal significance threshold ( $P_T$ ) predicts disease risks





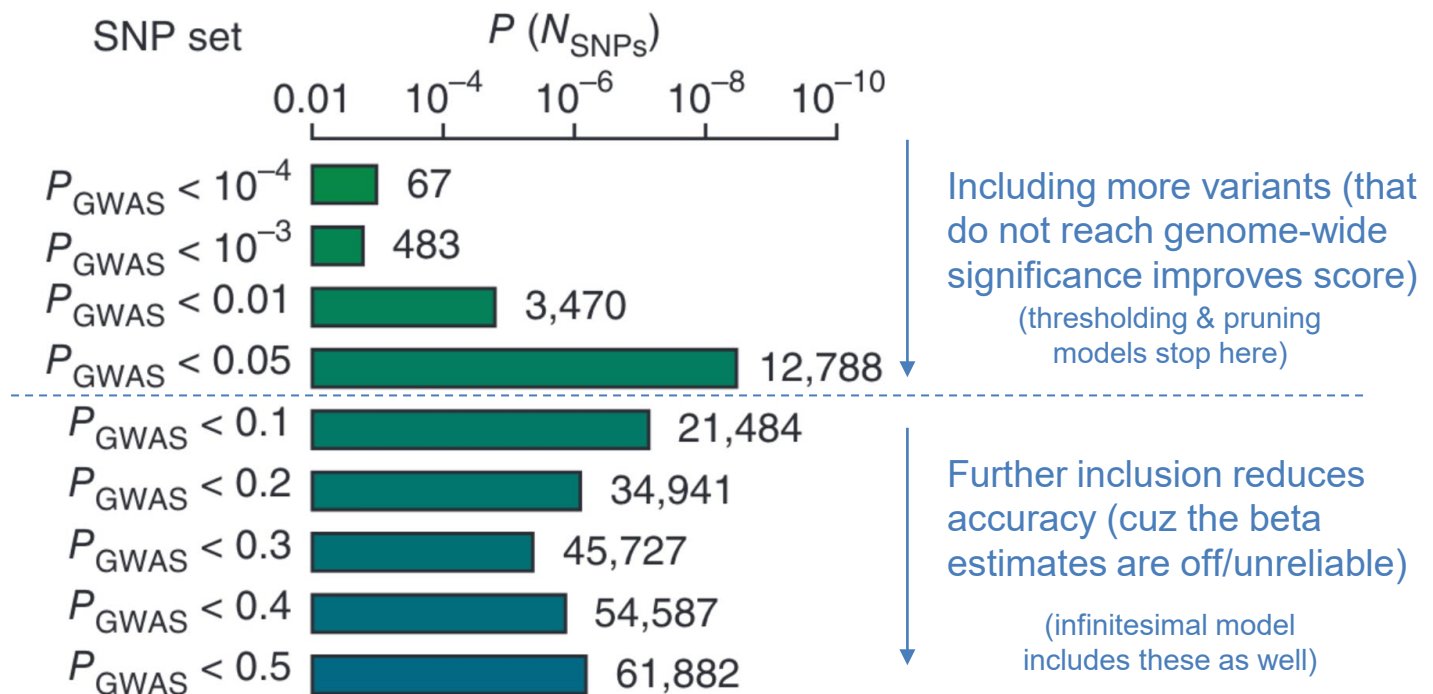
# Genetic architecture and PGS models

---

- Challenge:
  - How to estimate the polygenic effect sizes from GWAS effect size
- PGS accuracy depends on genetic architecture
- Genetic architecture is trait-specific
- **Infinitesimal model**: all independent SNPs have non-zero effects on traits
  - Use all LD-independent SNPs in GWAS
  - Equivalent to P + T model with  $P_T = 1$
- **Non-infinitesimal model**:
  - Mixture of components (zero effects, non-zero effects, ...)

# Pruning and thresholding (P + T) approach improves prediction over infinitesimal model

- Pruning and thresholding
  - Assume genetic architecture where a subset of GWAS SNPs contribute to the disease risk
  - Apply shrinkage of the estimates by P-value thresholding and clumping
- For Rheumatoid Arthritis,  $P_T = 0.05$  was the best in Stahl, et al. 2012



# Pruning and thresholding (P + T) is commonly used PGS model

---

- User-friendly software packages are available for P+T



GigaScience, 8, 2019, 1–6

---

doi: [10.1093/gigascience/giz082](https://doi.org/10.1093/gigascience/giz082)

Technical Note

---

TECHNICAL NOTE

## PRsice-2: Polygenic Risk Score software for biobank-scale data

Shing Wan Choi <sup>1,2,\*</sup> and Paul F. O'Reilly <sup>1,2,\*</sup>

<sup>1</sup>MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London, UK, SE5 8AF; and

<sup>2</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, 1 Gustave L. Levy Pl, New York City, NY 10029, USA

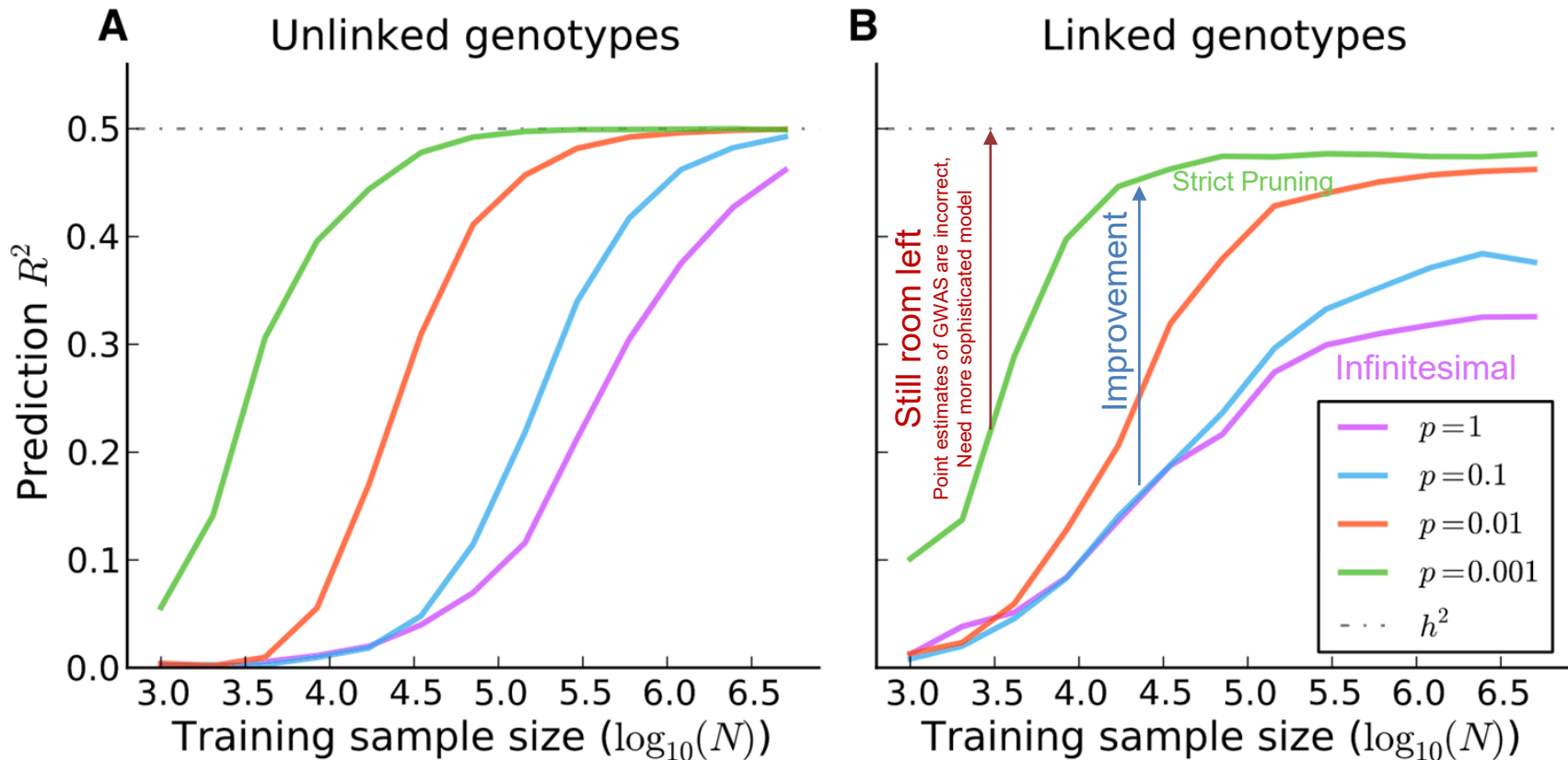
\*Correspondence address. Shing Wan Choi, Icahn School of Medicine, Mount Sinai, New York, USA. E-mail:

[choishingwan@gmail.com](mailto:choishingwan@gmail.com)  <http://orcid.org/0000-0003-2215-3238>; Paul F. O'Reilly, Icahn School of Medicine, Mount Sinai, New York, USA. E-mail:

[paul.oreilly@mssm.edu](mailto:paul.oreilly@mssm.edu)

# Pruning and thresholding (P + T) does not reach maximum predictive performance

- P + T does not model the LD structure between SNPs



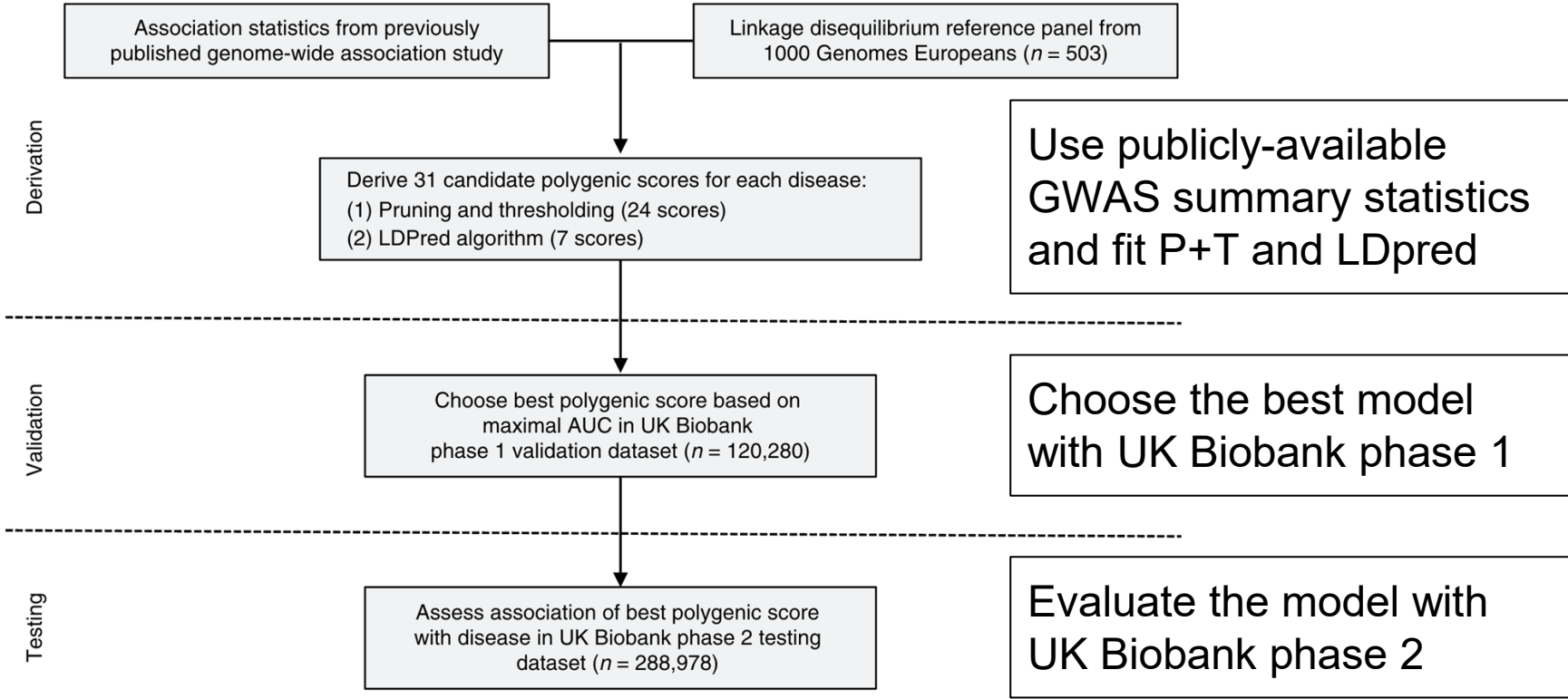
- LDpred (Vilhjálmsón et al 2015) models LD and improved prediction

# Modeling LD structure with LDpred shows prediction improvements over pruning+Thresholding



## Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations

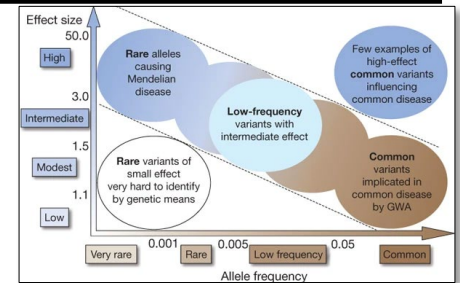
Amit V. Khera<sup>1,2,3,4,5</sup>, Mark Chaffin<sup>4,5</sup>, Krishna G. Aragam<sup>1,2,3,4</sup>, Mary E. Haas<sup>4</sup>, Carolina Roselli<sup>4</sup>, Seung Hoan Choi<sup>4</sup>, Pradeep Natarajan<sup>2,3,4</sup>, Eric S. Lander<sup>4</sup>, Steven A. Lubitz<sup>2,3,4</sup>, Patrick T. Ellinor<sup>2,3,4</sup> and Sekar Kathiresan<sup>1,2,3,4\*</sup>



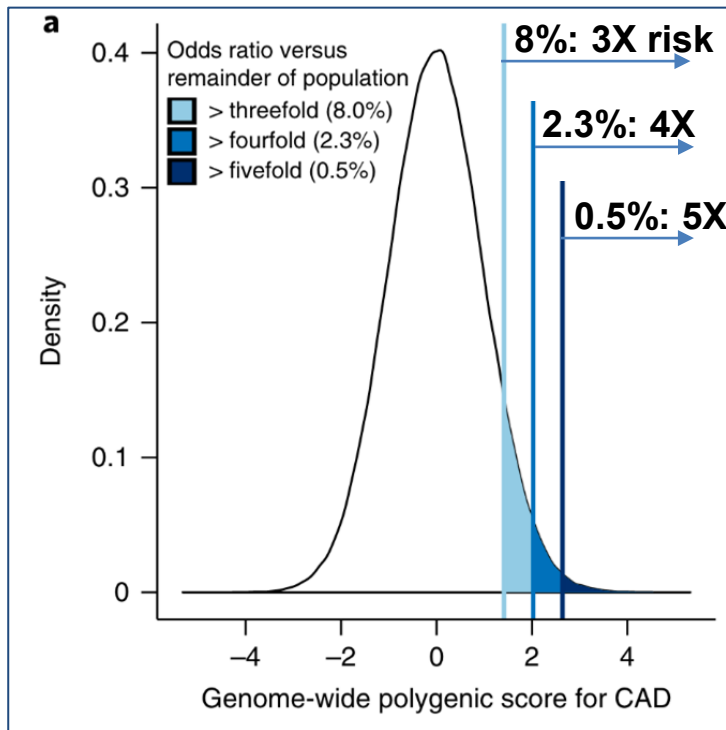
# LDpred: Modeling LD structure shows improvements in prediction over P+T

## Coronary artery disease (CAD)

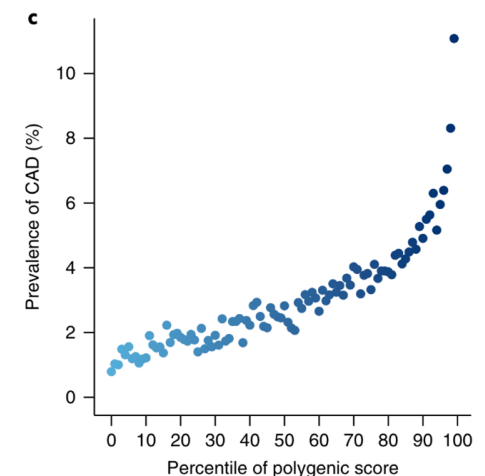
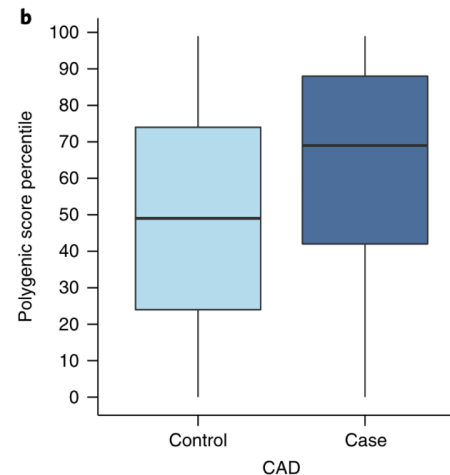
- **Rare** variants associated to familial hypercholesterolemia identified **0.4% of individuals** have odds ratio > 3.0
- **PGS** identified **8% of individuals** with odds ratio > 3.0



Yes, rare variants are \*individually\* very predictive for those individuals that carry them, but for the general population, PGS has now matched this predictive power (+applies to general population!)



Highlight the potential of PGS to identify **large-number of individuals** with high genetic liability of the disease



# Many Bayesian PGS methods report improvements over P+T

---







## SBayesR (Lloyd-Jones, et al. Nat Comm. 2019)

ARTICLE

<https://doi.org/10.1038/s41467-019-12653-0>

OPEN

### Improved polygenic prediction by Bayesian multiple regression on summary statistics

Luke R. Lloyd-Jones <sup>1,9\*</sup>, Jian Zeng <sup>1,9\*</sup>, Julia Sidorenko<sup>1,2</sup>, Loïc Yengo<sup>1</sup>, Gerhard Moser<sup>3,4</sup>, Kathryn E. Kemper<sup>1</sup>, Huanwei Wang <sup>1</sup>, Zhili Zheng<sup>1</sup>, Reedik Magi<sup>2</sup>, Tõnu Esko<sup>2</sup>, Andres Metspalu<sup>2,5</sup>, Naomi R. Wray <sup>1,6</sup>, Michael E. Goddard<sup>7</sup>, Jian Yang <sup>1,8\*</sup> & Peter M. Visscher <sup>1\*</sup>

Continuous thresholds with decaying contribution strengths (instead of single-threshold)

BayesR model (Gaussian mixture):

$$\beta_j | \pi, \sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2 \sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C \sigma_\beta^2) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$$



## PRS-CS (Ge, et al. Nat Comm. 2019)

ARTICLE

<https://doi.org/10.1038/s41467-019-09718-5>

OPEN

### Polygenic prediction via Bayesian regression and continuous shrinkage priors

Tian Ge<sup>1,2,3</sup>, Chia-Yen Chen <sup>1,2,3,4</sup>, Yang Ni <sup>5</sup>, Yen-Chen Anne Feng<sup>1,2,3,4</sup> & Jordan W. Smoller<sup>1,2,3</sup>

Local shrinkage parameter, applied based on GWAS estimate

Global-local scale mixtures of Gaussians:  $\beta_j \sim N\left(0, \frac{\sigma^2}{N} \phi \psi_j\right)$ ,  $\psi_j \sim G(a, \delta_j)$ ,  $\delta_j \sim G(b, 1)$ ,

# Sparse PGS using penalized regression

- Bayesian PGS approaches (LDpred, SBayesR, PRS-CS, etc.) show improvements over P+T
- The resulting model have **millions of SNVs** included in the model

GPS=Genome-wide polygenic score

**Table 1 | GPS derivation and testing for five common, complex diseases**

Disease	Discovery GWAS (n)	Prevalence in validation dataset	Prevalence in testing dataset	Polymorphisms in GPS	Tuning parameter	AUC (95% CI) in validation dataset	AUC (95% CI) in testing dataset
CAD	60,801 cases; 123,504 controls <sup>16</sup>	3,963/120,280 (3.4%)	8,676/288,978 (3.0%)	6,630,150	LDPred ( $\rho = 0.001$ )	0.81 (0.80–0.81)	0.81 (0.81–0.81)
Atrial fibrillation	17,931 cases; 115,142 controls <sup>30</sup>	2,024/120,280 (1.7%)	4,576/288,978 (1.6%)	6,730,541	LDPred ( $\rho = 0.003$ )	0.77 (0.76–0.78)	0.77 (0.76–0.77)
Type 2 diabetes	26,676 cases; 132,532 controls <sup>31</sup>	2,785/120,280 (2.4%)	5,853/288,978 (2.0%)	6,917,436	LDPred ( $\rho = 0.01$ )	0.72 (0.72–0.73)	0.73 (0.72–0.73)
Inflammatory bowel disease	12,882 cases; 21,770 controls <sup>32</sup>	1,360/120,280 (1.1%)	3,102/288,978 (1.1%)	6,907,112	LDPred ( $\rho = 0.1$ )	0.63 (0.62–0.65)	0.63 (0.62–0.64)
Breast cancer	122,977 cases; 105,974 controls <sup>33</sup>	2,576/63,347 (4.1%)	6,586/157,895 (4.2%)	5,218	Pruning and thresholding ( $r^2 < 0.2$ ; $P < 5 \times 10^{-4}$ )	0.68 (0.67–0.69)	0.69 (0.68–0.69)

AUC was determined using a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. The breast cancer analysis was restricted to female participants. For the LDpred algorithm, the tuning parameter  $\rho$  reflects the proportion of polymorphisms assumed to be causal for the disease. For the pruning and thresholding strategy,  $r^2$  reflects the degree of independence from other variants in the linkage disequilibrium reference panel, and  $P$  reflects the  $P$  value noted for a given variant in the discovery GWAS. CI, confidence interval.



# Sparse PGS using penalized regression

---

- Bayesian PGS approaches (LDpred, SBayesR, PRS-CS, etc.) show improvements over P+T
- The resulting model have millions of SNVs included in the model
  - Potential overfit and challenges in interpretation
- Penalized regression (Ridge/Lasso/Elastic Net) for sparse PGS

Lassosum (Mak, et al. *Genet Epidemiol.* 2017.)

Received: 14 June 2016 | Revised: 20 February 2017 | Accepted: 14 March 2017

DOI: 10.1002/gepi.22050

**RESEARCH ARTICLE**

WILEY Genetic  
Epidemiology

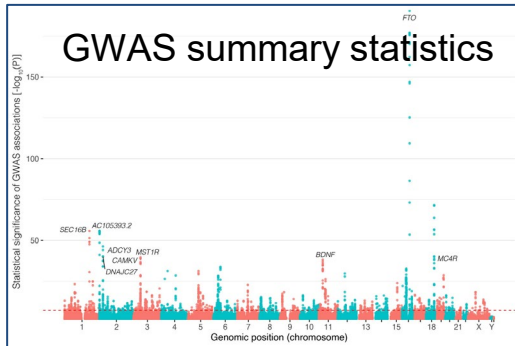
OFFICIAL JOURNAL  
INTERNATIONAL GENETIC  
EPIDEMIOLOGY SOCIETY  
www.geneticepi.org

## Polygenic scores via penalized regression on summary statistics

Timothy Shin Heng Mak<sup>1</sup>  | Robert Milan Porsch<sup>2</sup> | Shing Wan Choi<sup>2</sup> | Xueya Zhou<sup>2</sup> |  
Pak Chung Sham<sup>1,2,3</sup>

# PGS models on individual-level data

- Many PGS approaches start with GWAS summary statistics



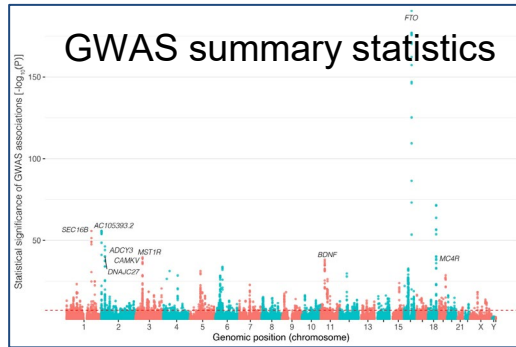
+ LD reference

PGS modeling

$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

# PGS models on individual-level data

- Many PGS approaches start with GWAS summary statistics



→

+ LD reference

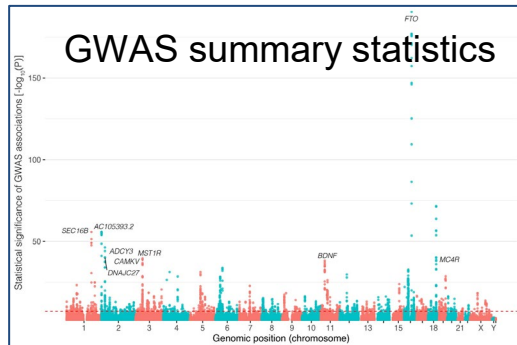
PGS modeling

$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

- We can consider fitting PGS directly on individual-level data
  - Multivariate model that consider multiple SNVs simultaneously
  - (GWAS: fitting univariate effects for each SNVs independently)

# PGS models on individual-level data

- Many PGS approaches start with GWAS summary statistics



PGS modeling

+ LD reference

$$PRS_i = \sum_{j \in J} \beta_j G_{ij}$$

- We can consider fitting PGS directly on individual-level data
  - Multivariate model that consider multiple SNVs simultaneously
  - (GWAS: fitting univariate effects for each SNVs independently)
- Example: BULP (Best Unbiased Linear Predictor)
  - Fit mixed model associations: Model all SNPs jointly instead of individually
  - Accounts for relatedness → Improves when some individuals related
  - Accounts for other SNPs → Improves even if all individuals are unrelated
  - Review: de los Campos et al. *Nat Rev Genet* (2010)
- Example: BASIL (batch screening iterative lasso) and *snpnet*

# Learning PGS on individual-level data with BASIL (Batch Screening Iterative Lasso) and *snpnet*

Polygenic risk score (PRS)

$$\text{PRS}_i = \sum_{j \in J} \beta_j G_{ij}$$

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

$L_1$  penalized regression w/ Lasso  
BASIL algorithm & R *snpnet* package

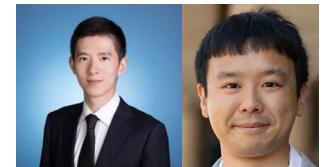
## PLOS GENETICS

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

### A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank

Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A. Rivas, Trevor Hastie 



Junyang Qian  
Yosuke Tanigawa

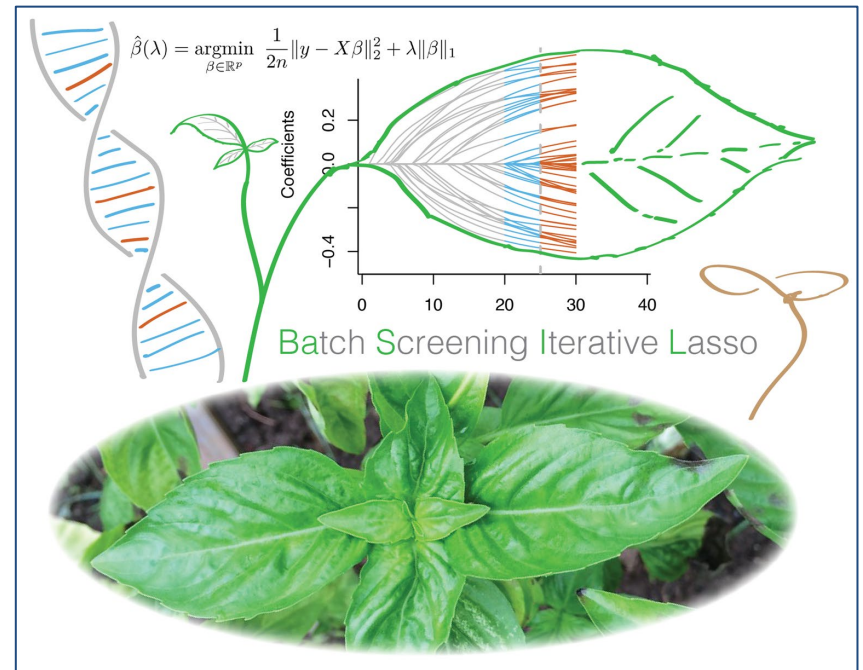
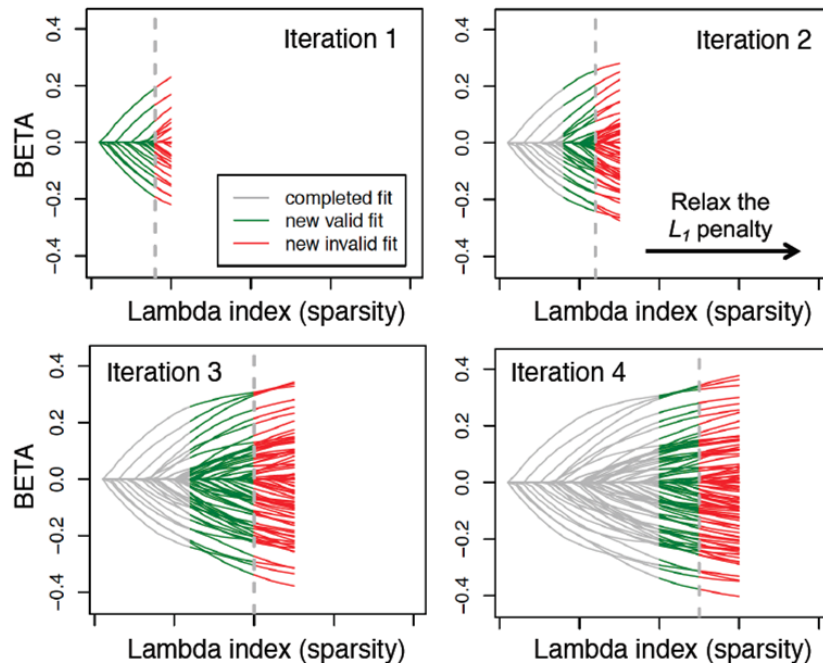
# Batch screening iterative Lasso (BASIL)

BASIL (= Batch Screening Iterative Lasso) in R *snpnet* package

## 3 steps per iteration

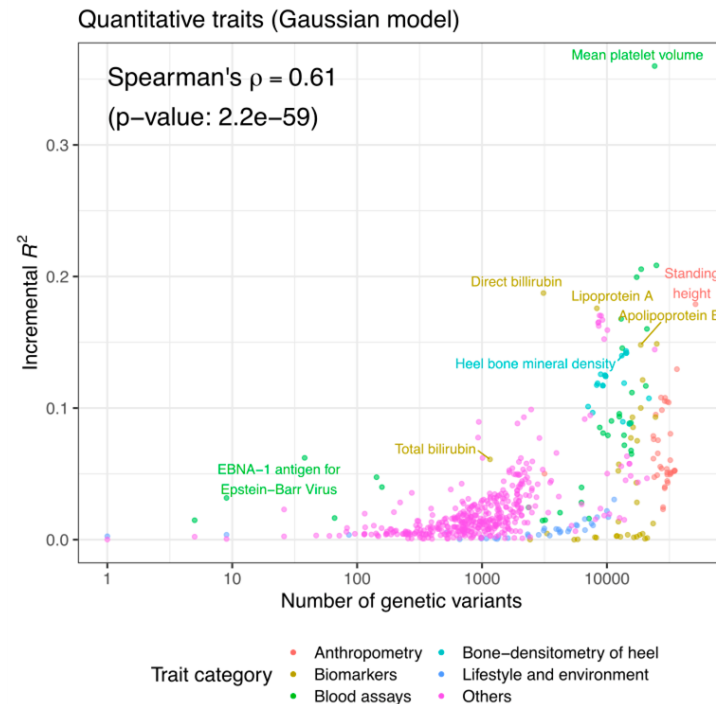
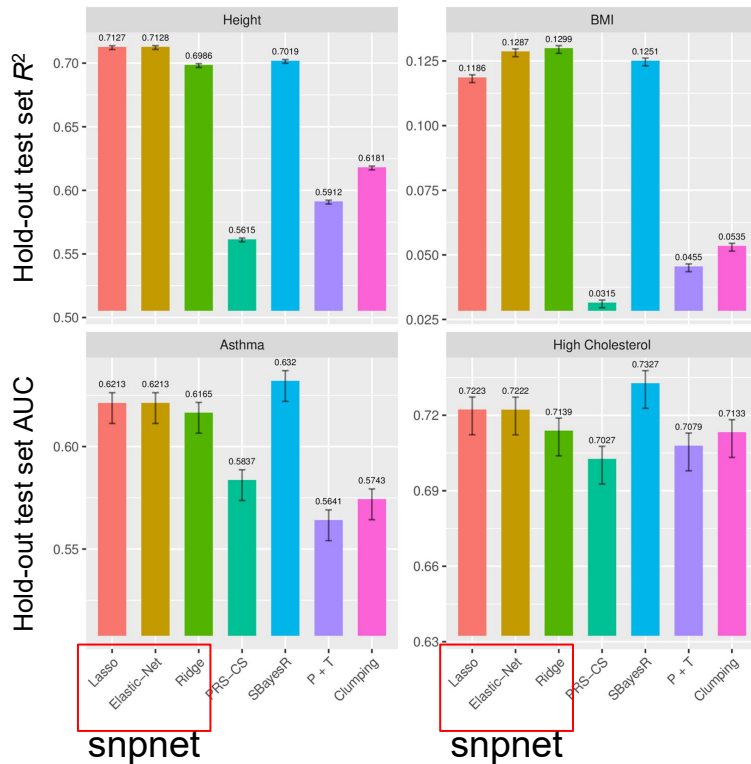
1. Screening
2. Lasso Fit (glmnet)
3. KKT Check

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$



# BASIL/snpnet model are sparse, yet have comparable predictive performance

- The snpnet PRS models (Lasso & Elastic-Net) have comparable predictive performance with SBayesR



- Standing height was one of the most polygenic traits.
- High PRS model has 47k variants (5% of non-zero BETAs)

# Summary 3: Methods to fit PGS model

---

- PGS model: set of variants and their weights
- Predictive performance of “GWAS top hits” depends on genetic architecture of the trait
- PGS methodology: active area of research
- Well-known methodology:
  - Pruning and thresholding (P + T)
  - Bayesian modeling accounts for LD and showed improvements over P + T (LDpred, SBayesR, PRS-CS)
- New approaches:
  - Sparse PGS
  - PGS directly from individual-level data



# Overview: Genetic prediction of complex traits

---

1. Foundations of Human Genetic Variation
2. Polygenic score (PGS) introduction
3. PGS Evaluation
4. Methods to fit PGS model
5. Challenges and opportunities in PGS research

# Limited transferability of polygenic scores (PGS)

Limited predictive performance in non-European cohorts

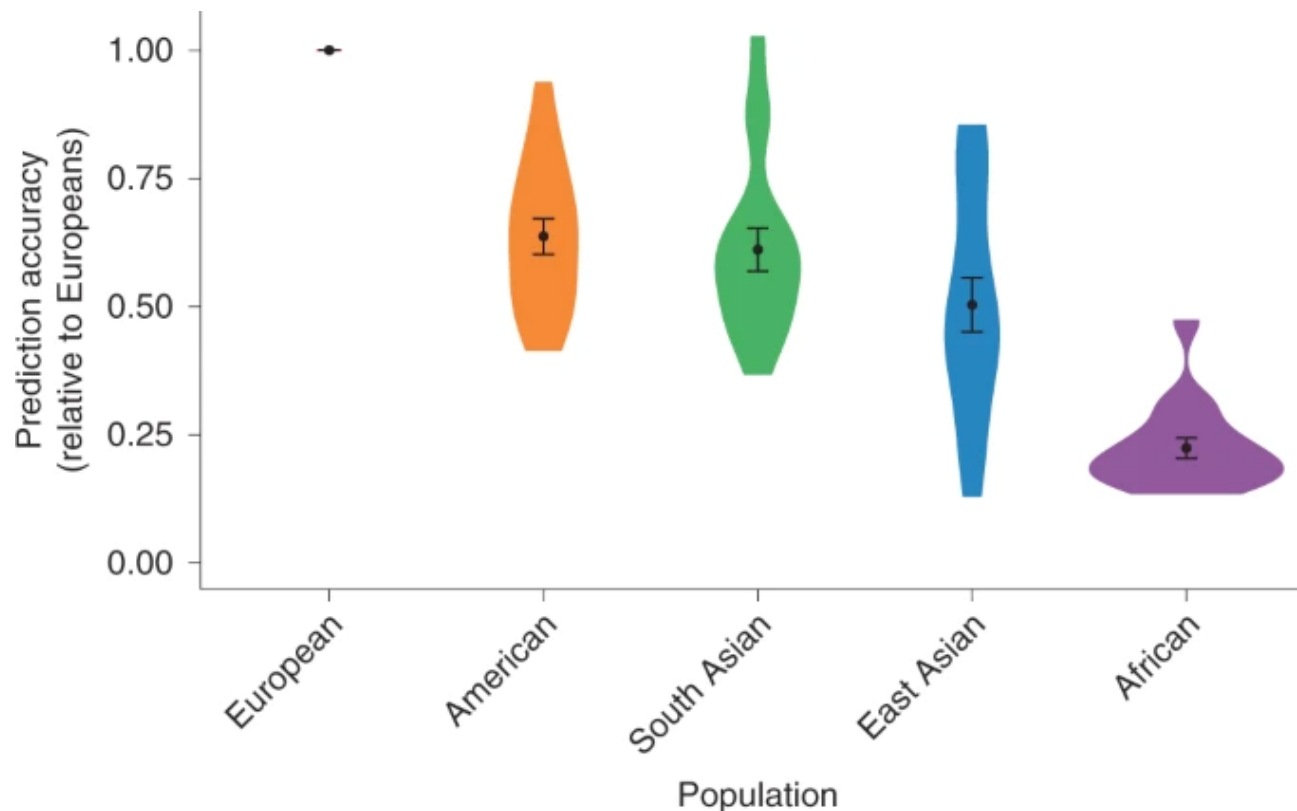
PERSPECTIVE

<https://doi.org/10.1038/s41588-019-0379-x>

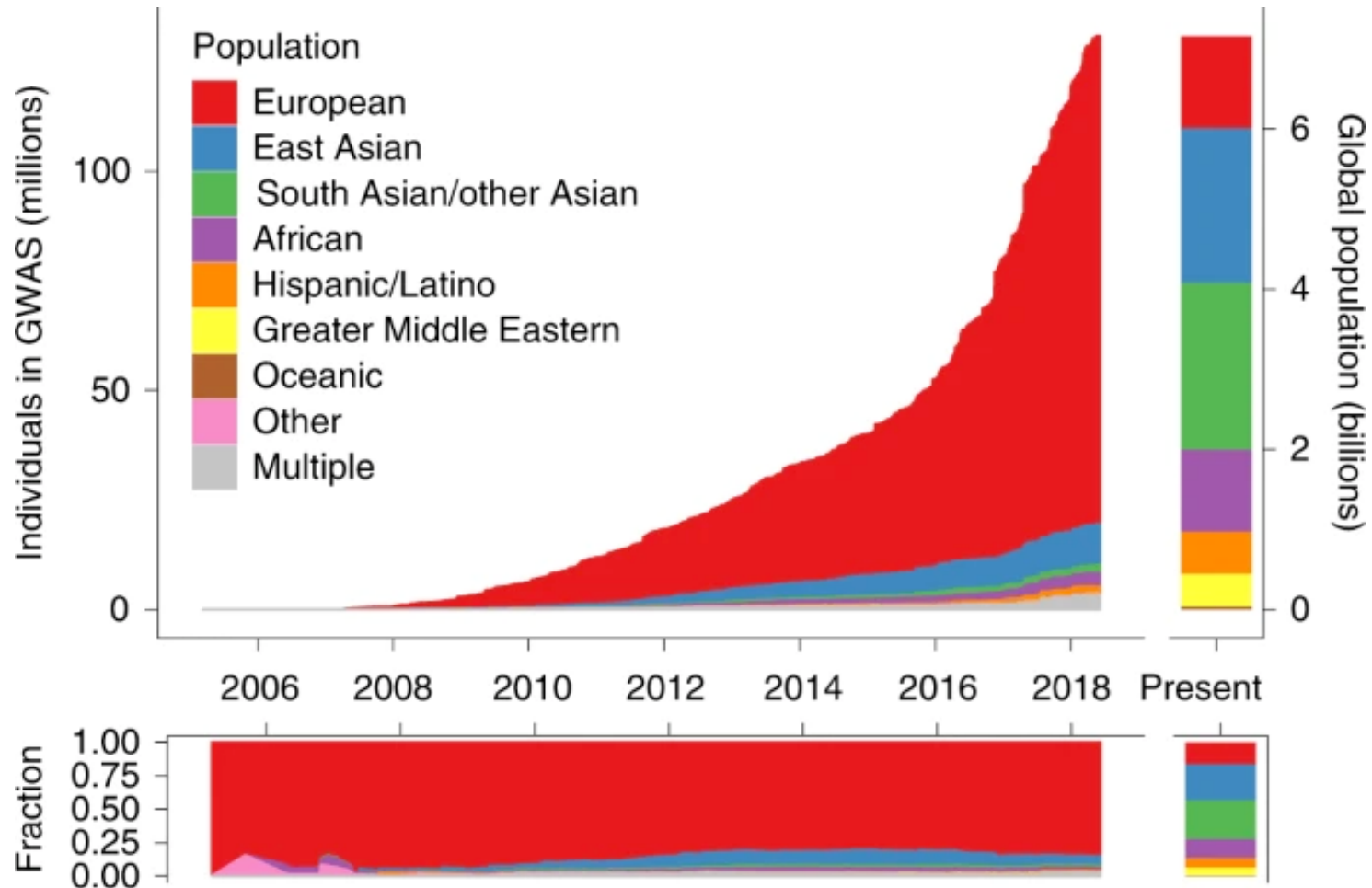
nature  
genetics

**Clinical use of current polygenic risk scores may exacerbate health disparities**

Alicia R. Martin<sup>1,2,3\*</sup>, Masahiro Kanai<sup>1,2,3,4,5</sup>, Yoichiro Kamatani<sup>5,6</sup>, Yukinori Okada<sup>5,7,8</sup>, Benjamin M. Neale<sup>1,2,3</sup> and Mark J. Daly<sup>1,2,3,9</sup>

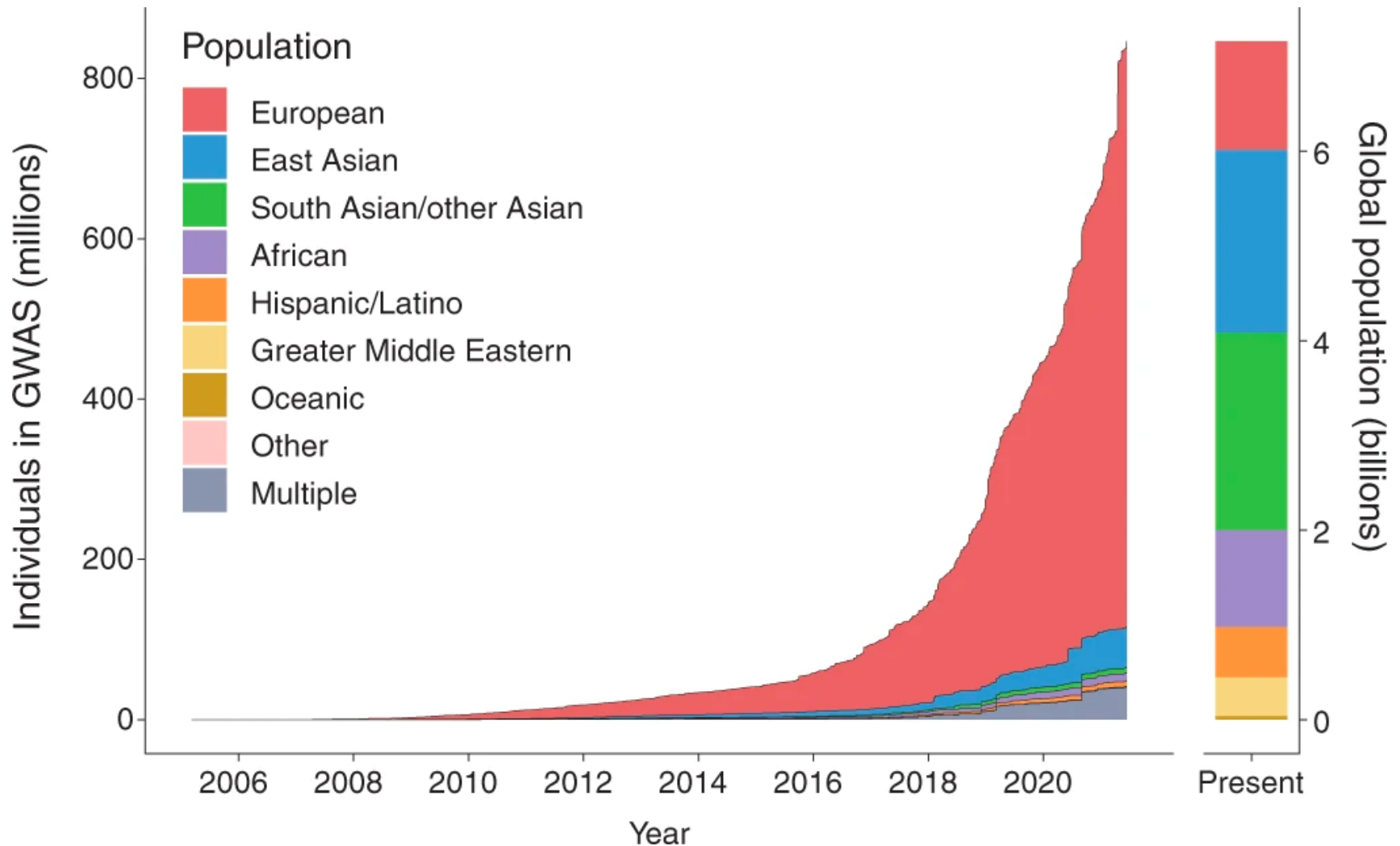


# Underrepresentation of non-European samples in GWAS studies



The challenge is well recognized in 2019 (Martin, et al. 2019)

# We still see lack of diversity today

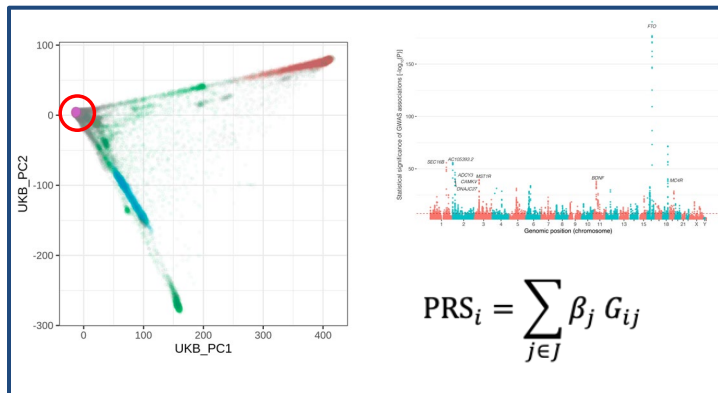


The proportion of samples from individuals cumulatively reported by the GWAS Catalog as of 8 July 2021

# Multi-ancestry polygenic score models combine multiple PGS predictors

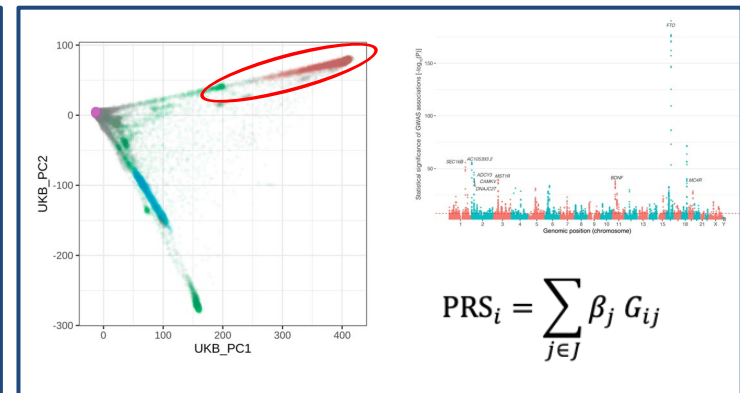
## 1. Fit $\text{PGS}_{\text{EUR}}$ and $\text{PGS}_{\text{AFR}}$ independently

PGS model for European ancestry



Large sample size, statistical power

PGS model for African ancestry



Relevant LD structure and MAF

## 1. Consider linear combination of the two

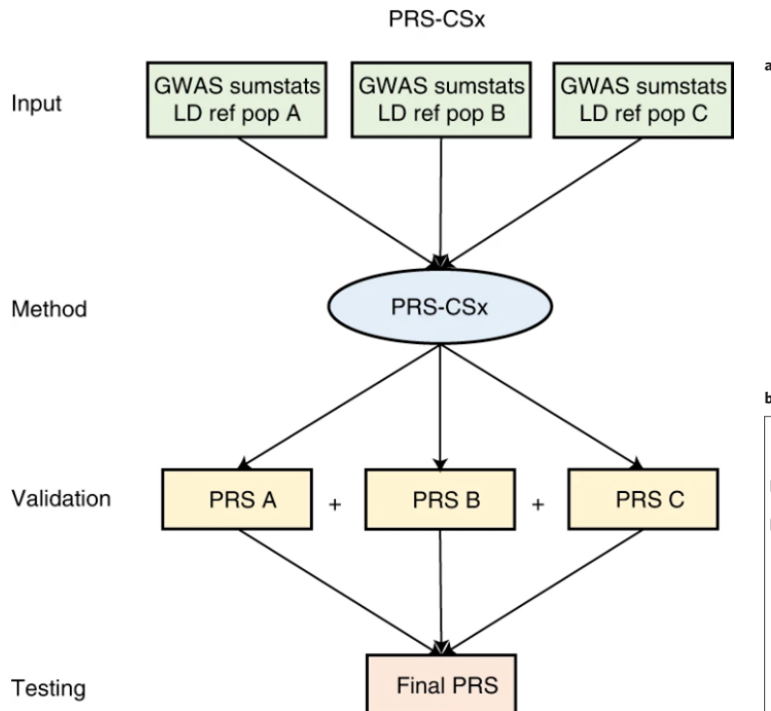
$$w_{\text{EUR}} \text{PGS}_{\text{EUR}} + w_{\text{AFR}} \text{PGS}_{\text{AFR}}$$

Marquez-Luna et al. 2017 Genet Epidemiol

# Multi-ancestry polygenic score models combines multiple PGS predictors

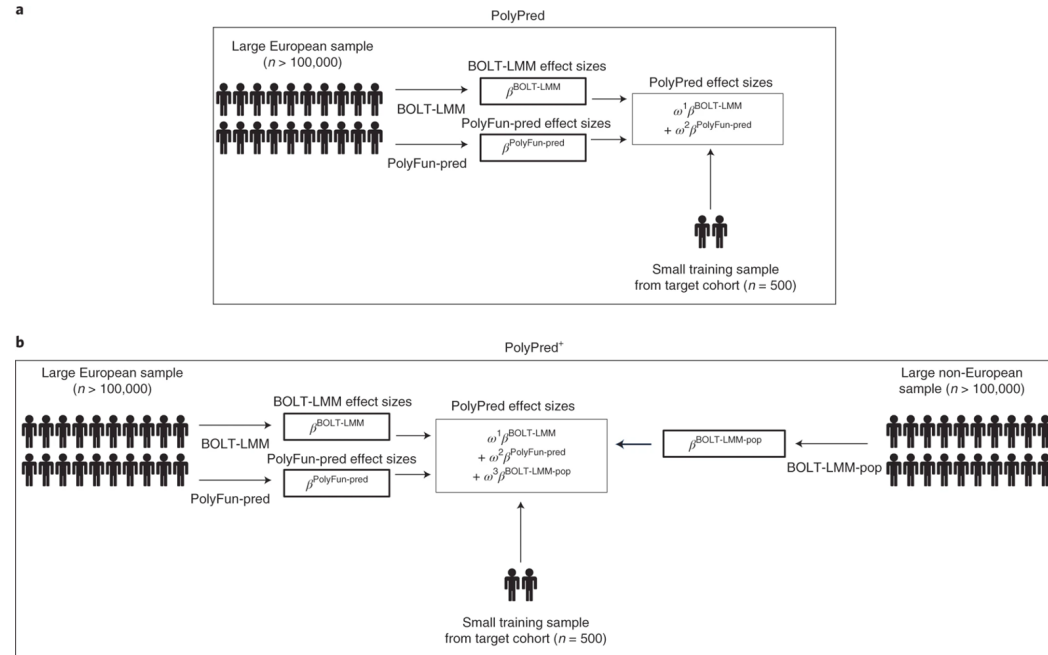
## Improving polygenic prediction in ancestrally diverse populations

Yunfeng Ruan<sup>1,2</sup>, Yen-Feng Lin<sup>3,4,5</sup>, Yen-Chen Anne Feng<sup>1,6,7,8,9,10</sup>, Chia-Yen Chen<sup>11</sup>, Max Lam<sup>1,8,12,13,14</sup>, Zhenglin Guo<sup>1</sup>, Stanley Global Asia Initiatives\*, Lin He<sup>2</sup>, Akira Sawa<sup>15</sup>, Alicia R. Martin<sup>1,8,16</sup>, Shengying Qin<sup>2,60</sup>, Hailiang Huang<sup>1,8,16,60</sup> and Tian Ge<sup>1,6,7,17,60</sup>

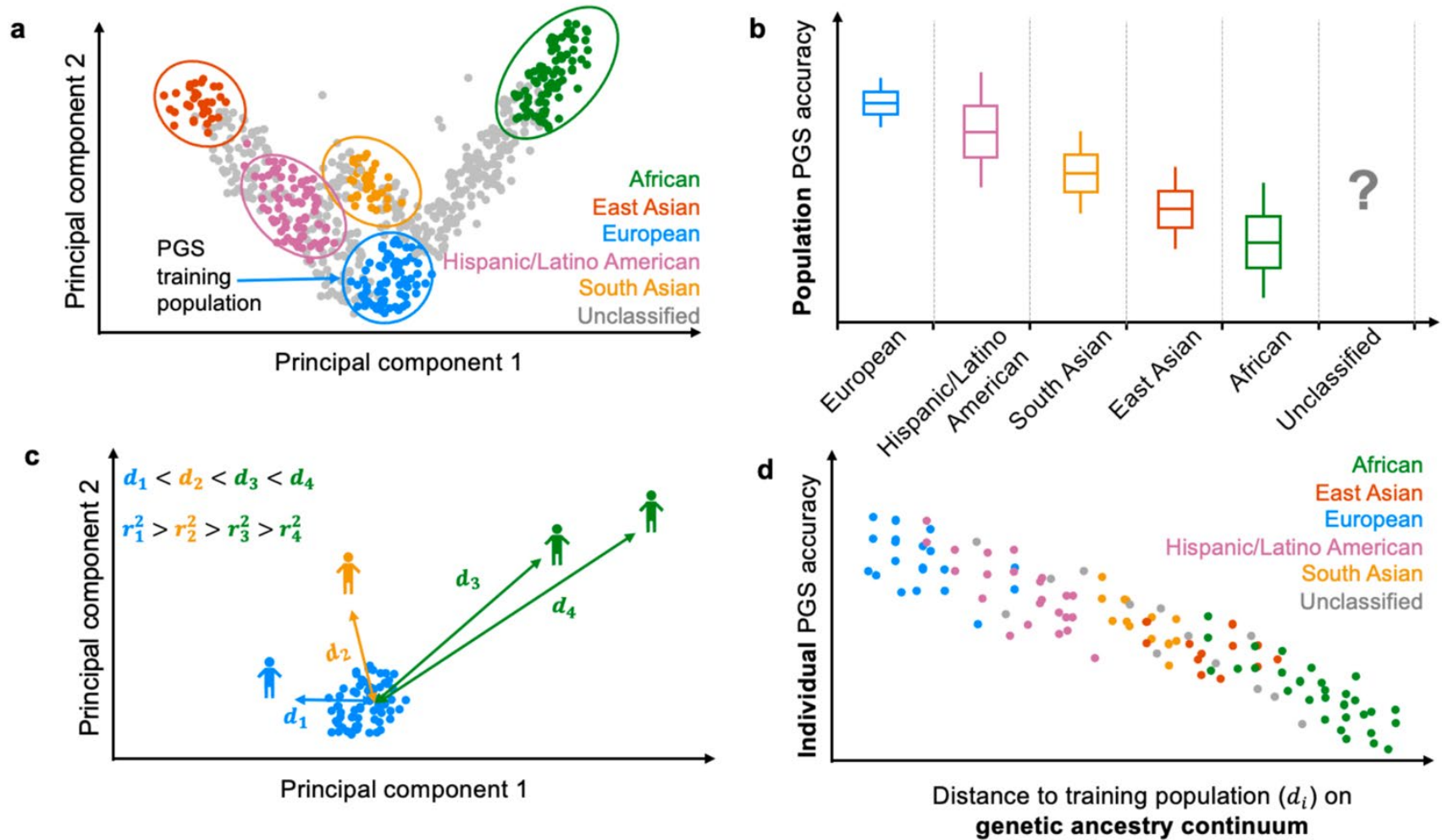


## Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores

Omer Weissbrod<sup>1,34</sup>, Masahiro Kanai<sup>2,3,34</sup>, Huwenbo Shi<sup>1,4,34</sup>, Steven Gazal<sup>1,5,6</sup>, Wouter J. Peyrot<sup>1,7</sup>, Amit V. Khera<sup>2,8</sup>, Yukinori Okada<sup>3,9</sup>, The Biobank Japan Project\*, Alicia R. Martin<sup>1,2</sup>, Hilary K. Finucane<sup>2,10</sup> and Alkes L. Price<sup>1,2</sup>



# Linear decay of the PGS predictive performance across genome-wide genetic ancestry

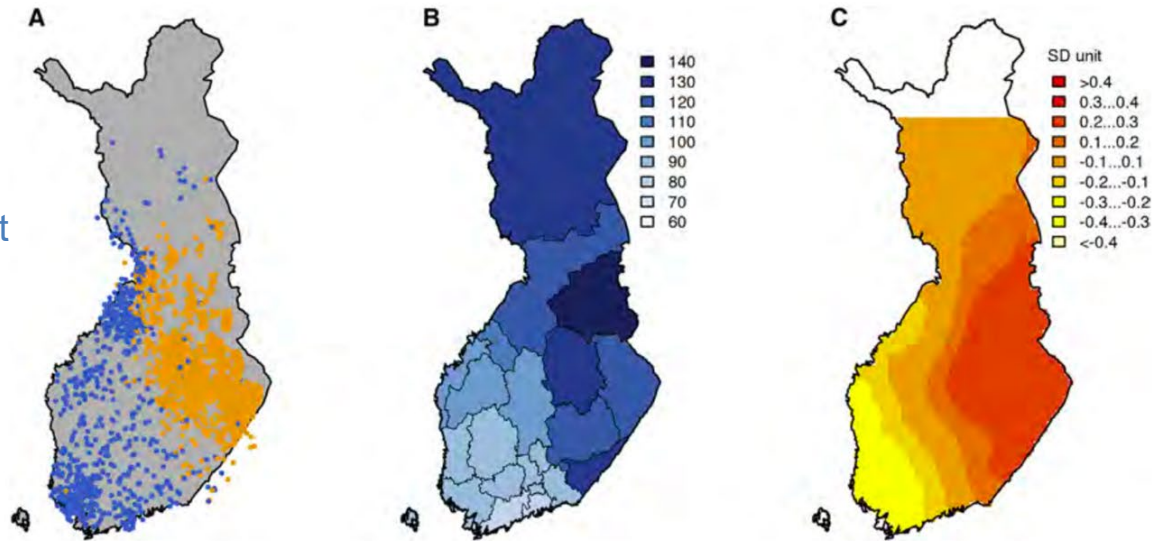


# Effects of some of the population structure remain unadjusted in PGS models

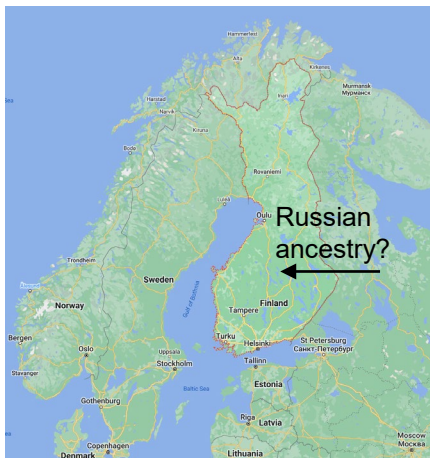
## Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland

Sini Kerminen,<sup>1</sup> Alicia R. Martin,<sup>2,3,4</sup> Jukka Koskela,<sup>1</sup> Sanni E. Ruotsalainen,<sup>1</sup> Aki S. Havulinna,<sup>1,5</sup> Ida Surakka,<sup>1,6</sup> Aarno Palotie,<sup>1,2,3,7,8</sup> Markus Perola,<sup>1,5</sup> Veikko Salomaa,<sup>5</sup> Mark J. Daly,<sup>1,2,3,4</sup> Samuli Ripatti,<sup>1,9</sup> and Matti Pirinen<sup>1,9,10,\*</sup>

Adjusting for PCA of population Structure captures continent-level population stratification, but residual remains within country

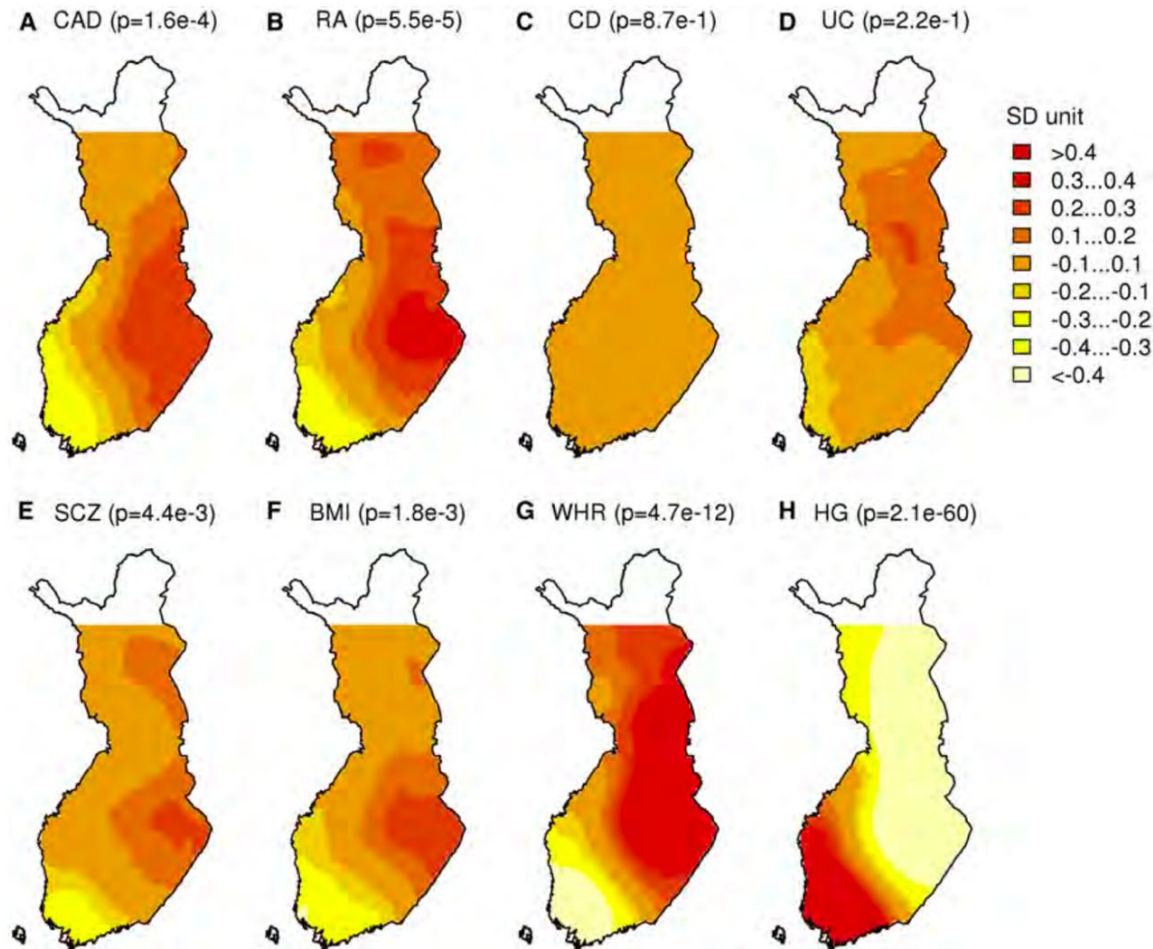


**Figure 1. A Comparison of Genetic Population Structure, Incidence Rates, and Distribution of the Polygenic Score of Coronary Artery Disease in Finland** (A–C) Main genetic population structure (A), the incidence rate for age-adjusted coronary artery disease (CAD) in 2013–2015 (Sepelvaltimotauti-indeksi, see [Web Resources](#)) (B), and the distribution of the polygenic score (PS) for CAD (C) in Finland. The population structure was estimated by clustering 2,376 samples into two groups.<sup>13</sup> The incidence rate is scaled to have a mean = 100. The PS distribution is shown in units of standard deviation.





# Effects of some of the population structure remain unadjusted in PGS models



**Figure 2. Distribution of Polygenic Scores in Finland**

(A–H) Distribution of polygenic scores for (A) coronary artery disease, (B) rheumatoid arthritis, (C) Crohn disease, (D) ulcerative colitis, (E) schizophrenia, (F) body-mass index, (G) waist-hip ratio adjusted for body-mass index, and (H) height. P values correspond to the association with longitude presented in [Table 2](#).

# How best to incorporate rare variants into PGS?

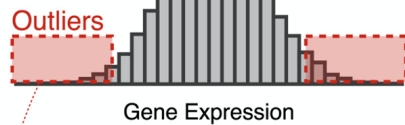
- Active area of research
- One approach: use expression outliers from eQTLs

## Integration of rare expression outlier-associated variants improves polygenic risk prediction

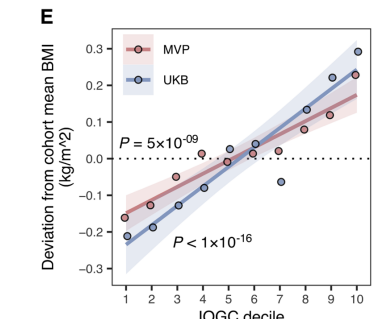
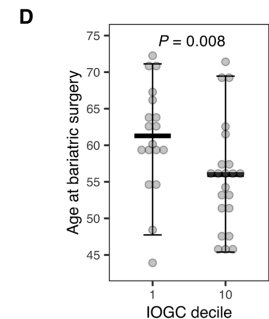
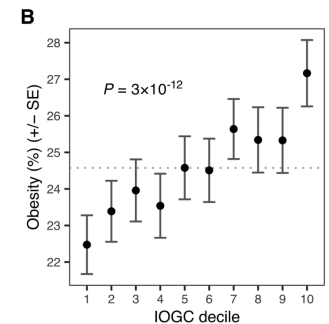
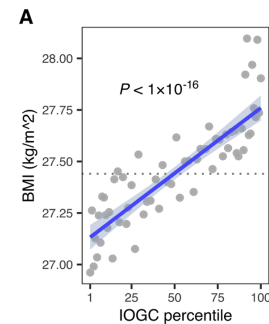
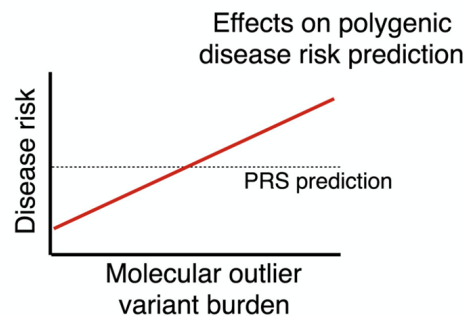
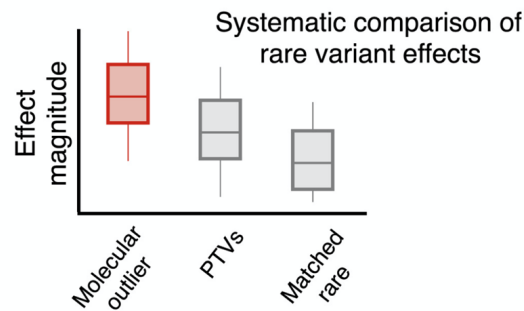
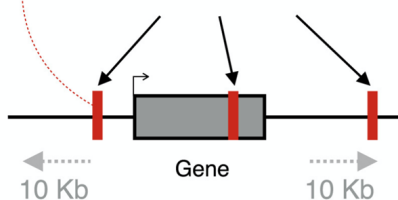
Craig Smail,<sup>1,2,\*</sup> Nicole M. Ferraro,<sup>1</sup> Qin Hui,<sup>3,4</sup> Matthew G. Durrant,<sup>5</sup> Matthew Aguirre,<sup>1</sup> Yosuke Tanigawa,<sup>1</sup> Marissa R. Keever-Keigher,<sup>2</sup> Abhiram S. Rao,<sup>6,7</sup> Johanne M. Justesen,<sup>1</sup> Xin Li,<sup>8</sup> Michael J. Gloudemans,<sup>1</sup> Themistocles L. Assimes,<sup>9,10</sup> Charles Kooperberg,<sup>11</sup> Alexander P. Reiner,<sup>12</sup> Jie Huang,<sup>13</sup> Christopher J. O'Donnell,<sup>14,15,16</sup> Yan V. Sun,<sup>3,4</sup> Million Veteran Program, Manuel A. Rivas,<sup>1</sup> and Stephen B. Montgomery<sup>5,6,\*</sup>

Variants with extreme expression effects also have stronger phenotypic consequences

Rare variants linked to dysregulated gene expression



Rare molecular outlier SNVs



# Uncertainty in assigning “Top X% genetic liability” from PGS

- PGS effect size estimates are from Bayesian inference
- We should consider uncertainties in individual-level PGS estimates

Article | [Published: 20 December 2021](#)

## Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification

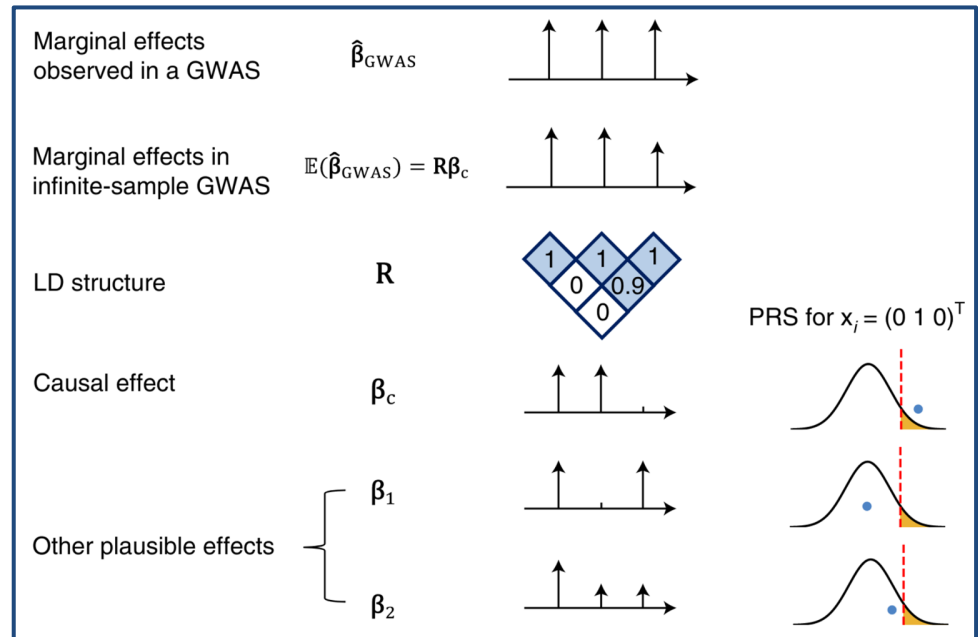
[Yi Ding](#) ✉, [Kangcheng Hou](#) ✉, [Kathryn S. Burch](#), [Sandra Lapinska](#), [Florian Privé](#), [Bjarni Vilhjálmsson](#), [Sriram Sankararaman](#) & [Bogdan Pasaniuc](#) ✉

[Nature Genetics](#) **54**, 30–39 (2022) | [Cite this article](#)

9367 Accesses | 13 Citations | 76 Altmetric | [Metrics](#)

“we observe large variances in individual PRS estimates which impact interpretation of PRS-based stratification; averaging across traits, only 0.8% (s.d. = 1.6%) of individuals with PRS point estimates in the top decile have corresponding 95% credible intervals fully contained in the top decile.”

Can only confidently [95% PPI int.] predict “you will be in top 10% of phenotype” for 0.8% of individuals (i.e. not 10%)



# Which PGS model is better? Statistical test for significance of difference in performance

---

ARTICLE

---

## Significance tests for $R^2$ of out-of-sample prediction using polygenic scores

Md. Moksedul Momin,<sup>1,2,3,4,\*</sup> Soohyun Lee,<sup>5</sup> Naomi R. Wray,<sup>6,7</sup> and S. Hong Lee<sup>1,2,4,\*</sup>

### Summary

The coefficient of determination ( $R^2$ ) is a well-established measure to indicate the predictive ability of polygenic scores (PGSs). However, the sampling variance of  $R^2$  is rarely considered so that 95% confidence intervals (CI) are not usually reported. Moreover, when comparisons are made between PGSs based on different discovery samples, the sampling covariance of  $R^2$  is required to test the difference between them. Here, we show how to estimate the variance and covariance of  $R^2$  values to assess the 95% CI and p value of the  $R^2$  difference. We apply this approach to real data calculating PGSs in 28,880 European participants derived from UK Biobank (UKBB) and Biobank Japan (BBJ) GWAS summary statistics for cholesterol and BMI. We quantify the significantly higher predictive ability of UKBB PGSs compared to BBJ PGSs (p value  $7.6e-31$  for cholesterol and  $1.4e-50$  for BMI). A joint model of UKBB and BBJ PGSs significantly improves the predictive ability, compared to a model of UKBB PGS only (p value  $3.5e-05$  for cholesterol and  $1.3e-28$  for BMI). We also show that the predictive ability of regulatory SNPs is significantly enriched over non-regulatory SNPs for cholesterol (p value  $8.9e-26$  for UKBB and  $3.8e-17$  for BBJ). We suggest that the proposed approach (available in R package `r2redux`) should be used to test the statistical significance of difference between pairs of PGSs, which may help to draw a correct conclusion about the comparative predictive ability of PGSs.

Statistical test for comparing PRS scores  
from different sources

r2redux: <https://github.com/mommy003/r2redux>

---

# PGS reporting standard (PGS-RS)

---

- PGS-RS to encourage PGS model sharing

Perspective | [Published: 10 March 2021](#)

## Improving reporting standards for polygenic scores in risk prediction studies

[Hannah Wand](#), [Samuel A. Lambert](#), [Cecelia Tamburro](#), [Michael A. Iacocca](#), [Jack W. O'Sullivan](#), [Catherine Sillari](#), [Iftikhar J. Kullo](#), [Robb Rowley](#), [Jacqueline S. Dron](#), [Deanna Brockman](#), [Eric Venner](#), [Mark I. McCarthy](#), [Antonis C. Antoniou](#), [Douglas F. Easton](#), [Robert A. Hegele](#), [Amit V. Khera](#), [Nilanjan Chatterjee](#), [Charles Kooperberg](#), [Karen Edwards](#), [Katherine Vlessis](#), [Kim Kinnear](#), [John N. Danesh](#), [Helen Parkinson](#), [Erin M. Ramos](#), ... [Genevieve L. Wojcik](#)  [+ Show authors](#)

[Nature](#) **591**, 211–219 (2021) | [Cite this article](#)

Can't just share PRS score between cohorts/studies. Need to also share metadata, correction factors, etc

- PGS equivalent (?) of the Minimum information about a microarray experiment (MIAME)
- Specify a wide range of recommendations for background, study population, risk model development and evaluation, limitations and clinical implications, and data availability

# PGS catalog – publicly available PGS weights and their (self-reported) evaluations



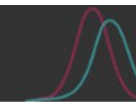
PGS Catalog

Home

Browse ▾

Downloads ▾

Documentation ▾



📅 Latest release: Feb. 8, 2023

## The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.



Examples: [breast cancer](#), [glaucoma](#), [BMI](#), [EFO\\_0001645](#)

### New tool!

We just released **pgsc\_calc**: a reproducible workflow to calculate both PGS Catalog and custom polygenic scores.

[> See more information](#)

Feedback

Can deposit models directly,  
then reuse directly

## Explore the Data

In the current PGS Catalog you can **browse** the scores and metadata through the following categories:

Polygenic Scores

⌘ 3,349

Traits

👤 584

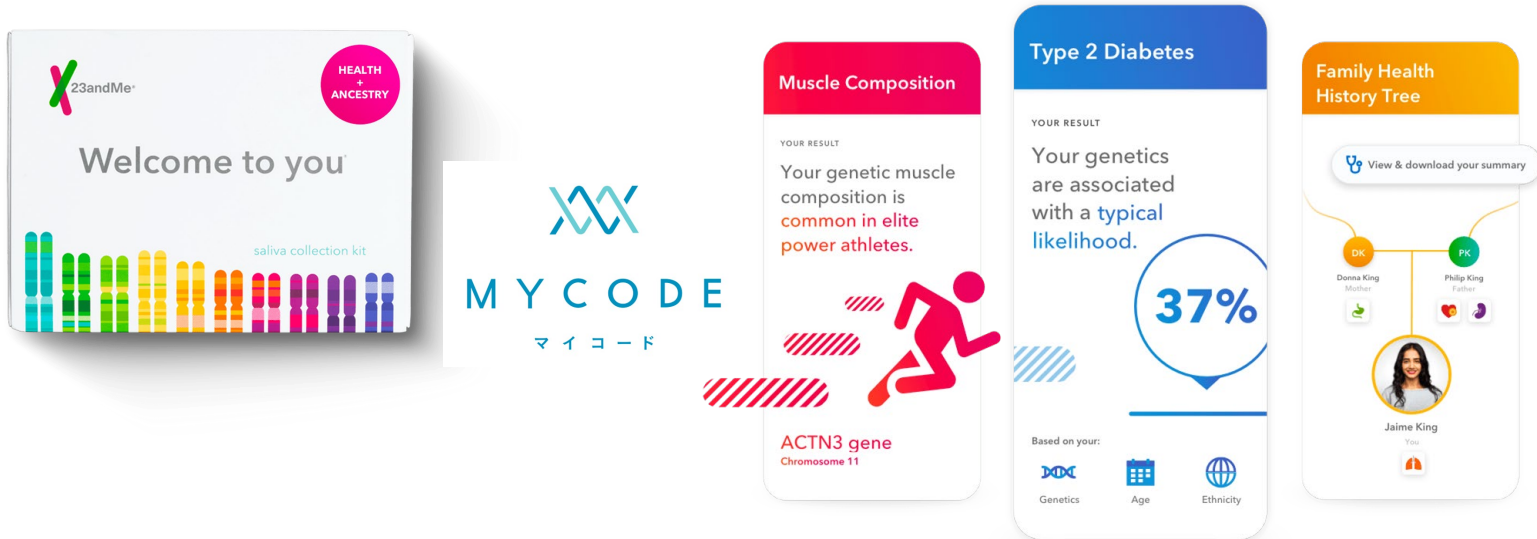
Publications

📖 426

<https://www.pgscatalog.org/>

# Bloom of D2C personal genomics companies

- Bloom of Direct-to-Consumer (D2C) personal genomics companies



- Considerations
  - Risk vs. benefits
  - Statistical significance vs. Clinical relevance
  - Ethics
  - Communications

## Good:

Power of information. Democratization  
More power to individual.

## But:

More dangers to misinterpret risk.  
Consequences to individuals.  
Treatments may come with risks.  
Doctors treat symptoms not risk.  
Benefits ↔ risk weighing...

## Solutions:

Need better warnings + general education.  
Regulatory supervision

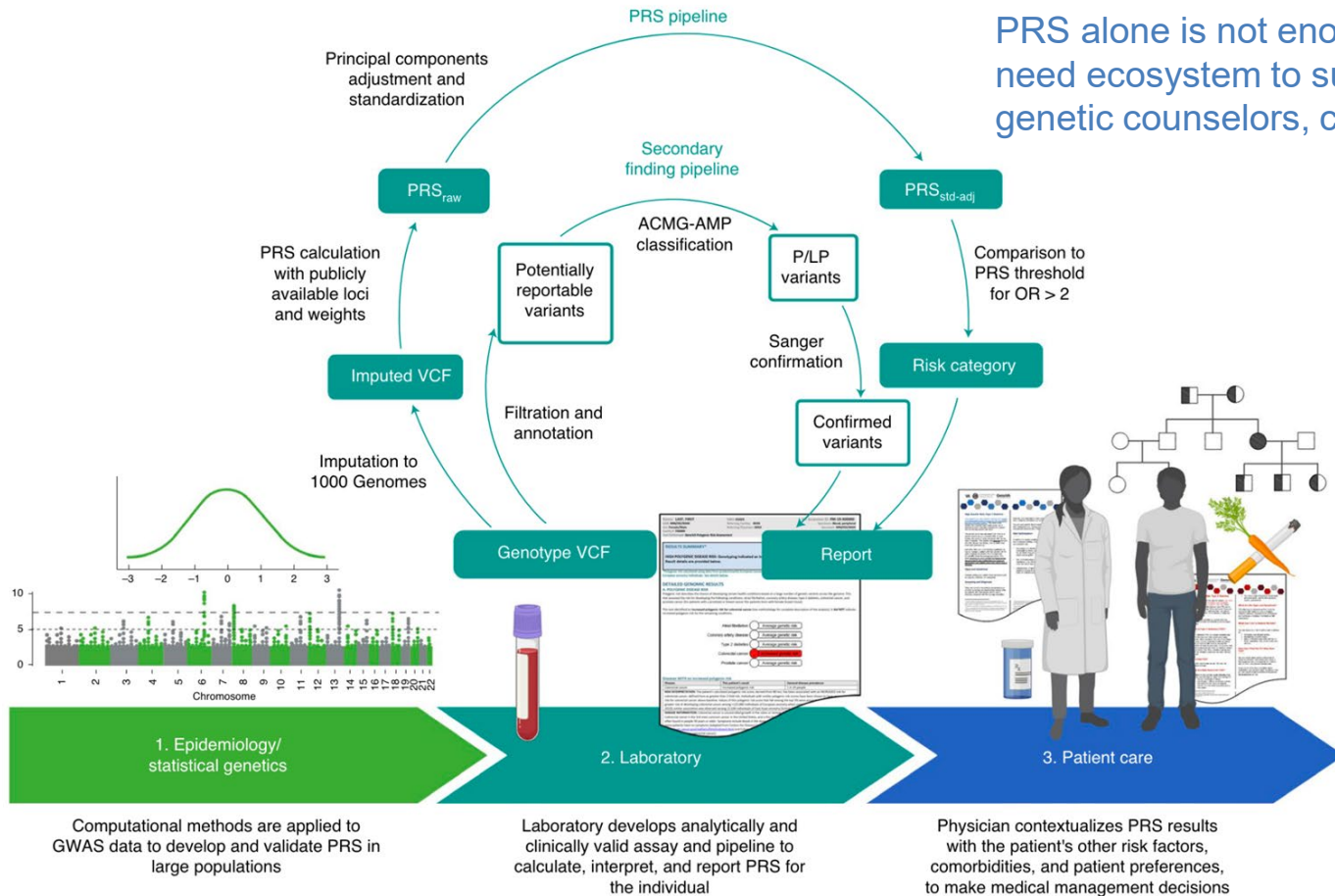
NB: This slide is not meant to endorse the service or products listed here.

<https://www.23andme.com/>

<https://mycode.jp>

# How do we bring back PGS results to clinic?

- Example: Veterans Affairs Genomic Medicine at Veterans Affairs (GenoVA) study (ongoing) develops PRS lab report and info packets







# How do we communicate PGS to patients?

## Perspectives of diverse Spanish- and English-speaking patients on the clinical use of polygenic risk scores



Sabrina A. Suckiel<sup>1,2</sup> , Giovanna T. Braganza<sup>1</sup>, Karla López Aguiñiga<sup>1</sup>,  
 Jacqueline A. Odgis<sup>1</sup>, Katherine E. Bonini<sup>1</sup>, Eimear E. Kenny<sup>1,2,3</sup>, Jada G. Hamilton<sup>4,5,6</sup>,  
 Noura S. Abul-Husn<sup>1,2,3,\*</sup> 

*“There was little concern among participants about the **limited predictive power of PRS for non-European** populations. **Barriers** to uptake of PRS testing and adoption of PRS-related recommendations included **socioeconomic factors, insurance status, race, ethnicity, language, and inadequate understanding of PRS**. Participants favored in-person PRS result disclosure by their physician”*

IN-PERSON				ELECTRONIC			
No specific reason				Patient portal is beneficial, time-efficient, accessible, familiar			
				Help support understanding		No specific reason	
				Email is time-efficient, accessible		Web-based application is time-efficient, accessible	
Read body language and tone of voice		Use preferred language		Avoid mail delays		TELEHEALTH	
		Confused by patient portal		Avoid tech challenges		MAIL	
				More secure		Read calmly at home	
				Equivalent to in-person			
				PHONE			
				Help support understanding			

**Fig.** Preferred methods for clinical PRS result disclosure and rationale

# Summary 4: Challenges and opportunities

---

- PGS model suffers from limited transferability
  - We lack GWAS data from diverse populations
  - Methodological innovations (weighted sum of PGSs)
- Remaining methodological challenges:
  - How to model the effects of population structure?
  - How to incorporate rare variants?
  - Uncertainties in individual-level PGS
- PGS model sharing and evaluation
  - Reporting standard & PGS catalog
- How to bring the results back to health care system?

# Acknowledgements

---

Kyoto-McGill International Joint Ph.D. Program in Genomic Medicine

## Kyoto University

Prof. Fumihiko Matsuda

Prof. Masao Nagasaki

Prof. Shuji Kawaguchi

Prof. Takahisa Kawaguchi



## McGill University

Prof. Mark Lathrop

Prof. Guillaume Bourque

