GWAS 9

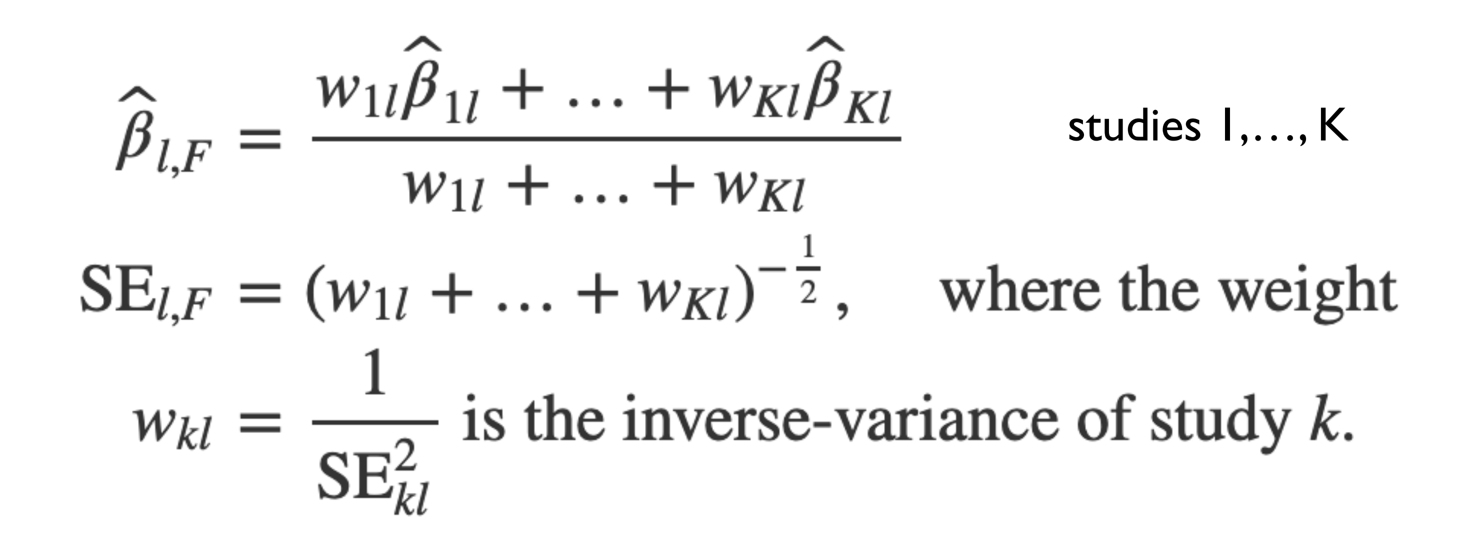
Matti Pirinen
University of Helsinki

20.4.2023

META-ANALYSIS

- Suppose that we have two independent estimates $\hat{x}_1 = 1.0$ and $\hat{x}_2 = 2.0$ of some unknown quantity *x*.
- Additionally, we are told that the "precision" of the first estimate is twice that of the second one.
- We combine the two estimates by weighting the first one twice as much as the second and hence our combined estimate is
  - $\hat{x} = (2\hat{x}_1$ + $\hat{x}_2)$ / (2 + 1) = (2.0 + 2.0) / 3 = 1.33
- This is the *fixed-effects* meta-analysis approach
  - "Precision" is 1/SE^2, that is, the inverse of the variance of the estimator

INVERSE VARIANCE WEIGHTED (IVW) FIXED-EFFECT (F) ESTIMATOR



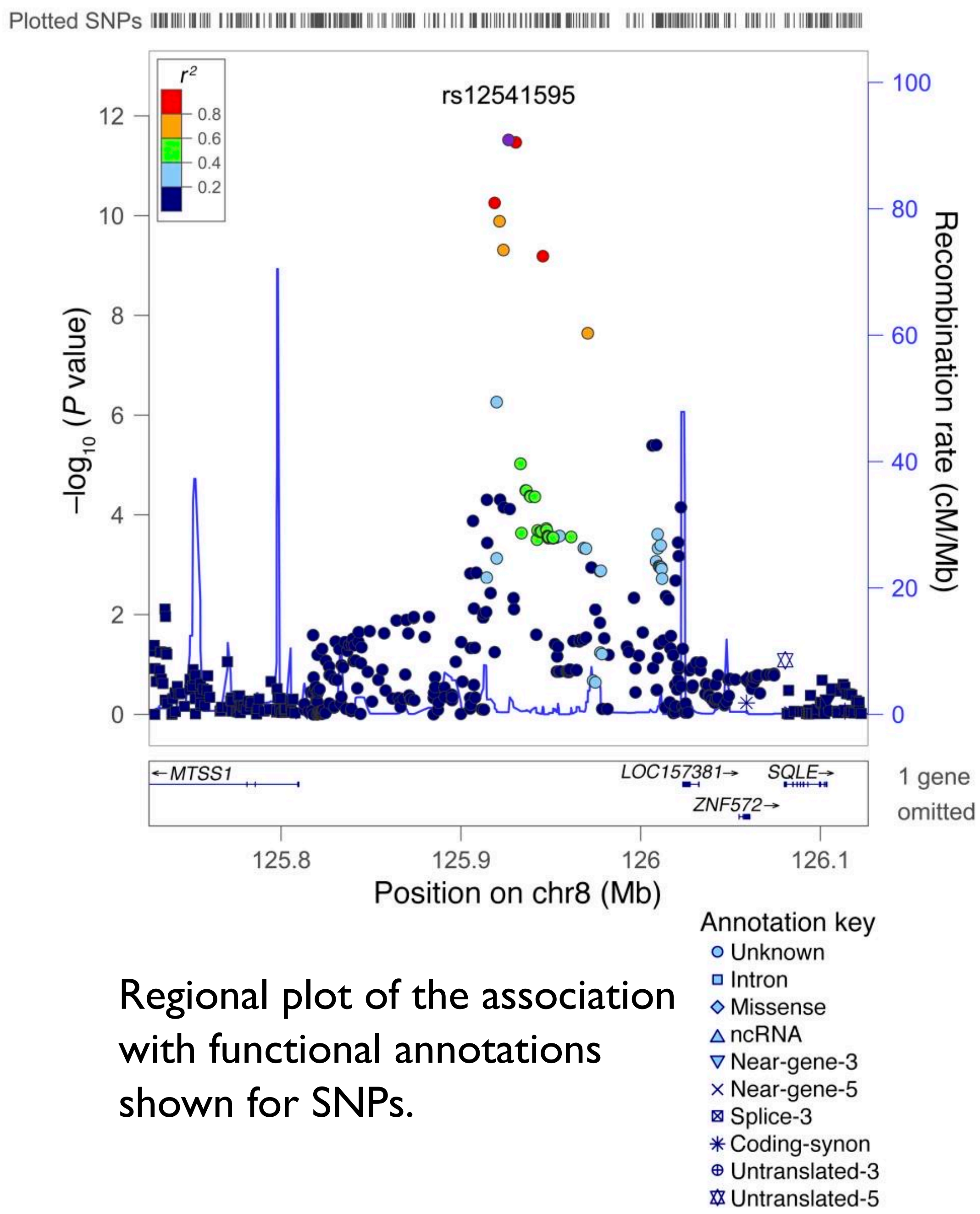
$$\widehat{\beta}_{l,F} = \frac{w_{1l}\widehat{\beta}_{1l} + \ldots + w_{Kl}\widehat{\beta}_{Kl}}{w_{1l} + \ldots + w_{Kl}}$$

studies 1,…, K

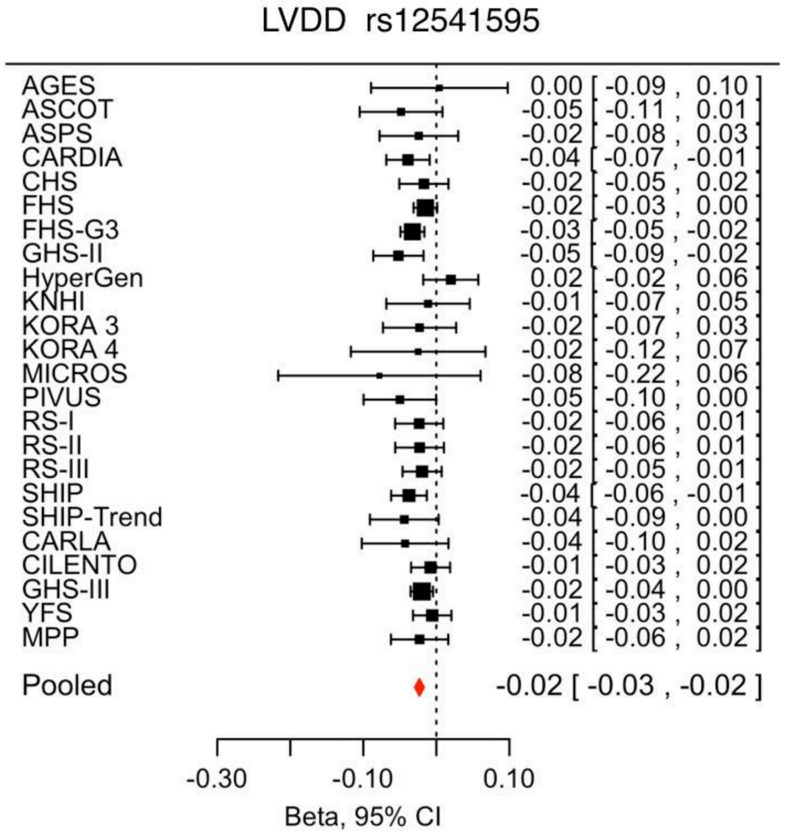
$$\mathrm{SE}_{l,F} = (w_{1l} + \ldots + w_{Kl})^{-\frac{1}{2}}, \quad \text{where the weight}$$

$$w_{kl} = \frac{1}{\mathrm{SE}_{kl}^2} \text{ is the inverse-variance of study } k.$$

- Each study is weighted by its precision (= inverse of the variance)
- Precision of the combined estimate is the sum of the precisions of the contributing estimates
- For binary outcomes, $\hat{\beta}$ is on the log-odds scale as in logistic regression output, **not** on the odds-ratio scale



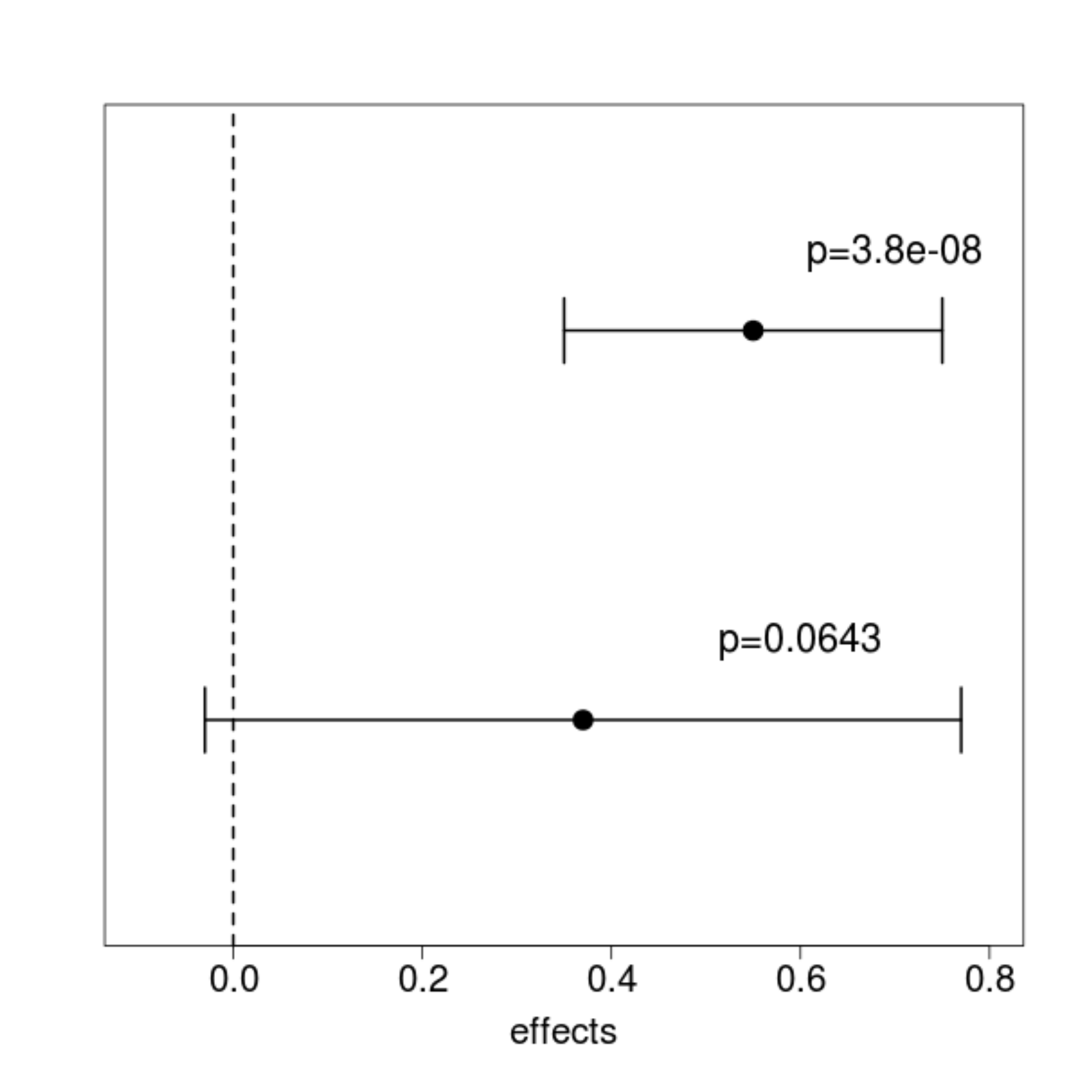
Regional plot of the association with functional annotations shown for SNPs.



Forest plot for the meta-analysis of the association between rs12541595 and left ventricular diastolic internal dimension (LVDD)

J Clin Invest. 2017;127(5):1798-1812

IS FIXED-EFFECTS ASSUMPTION REASONABLE?



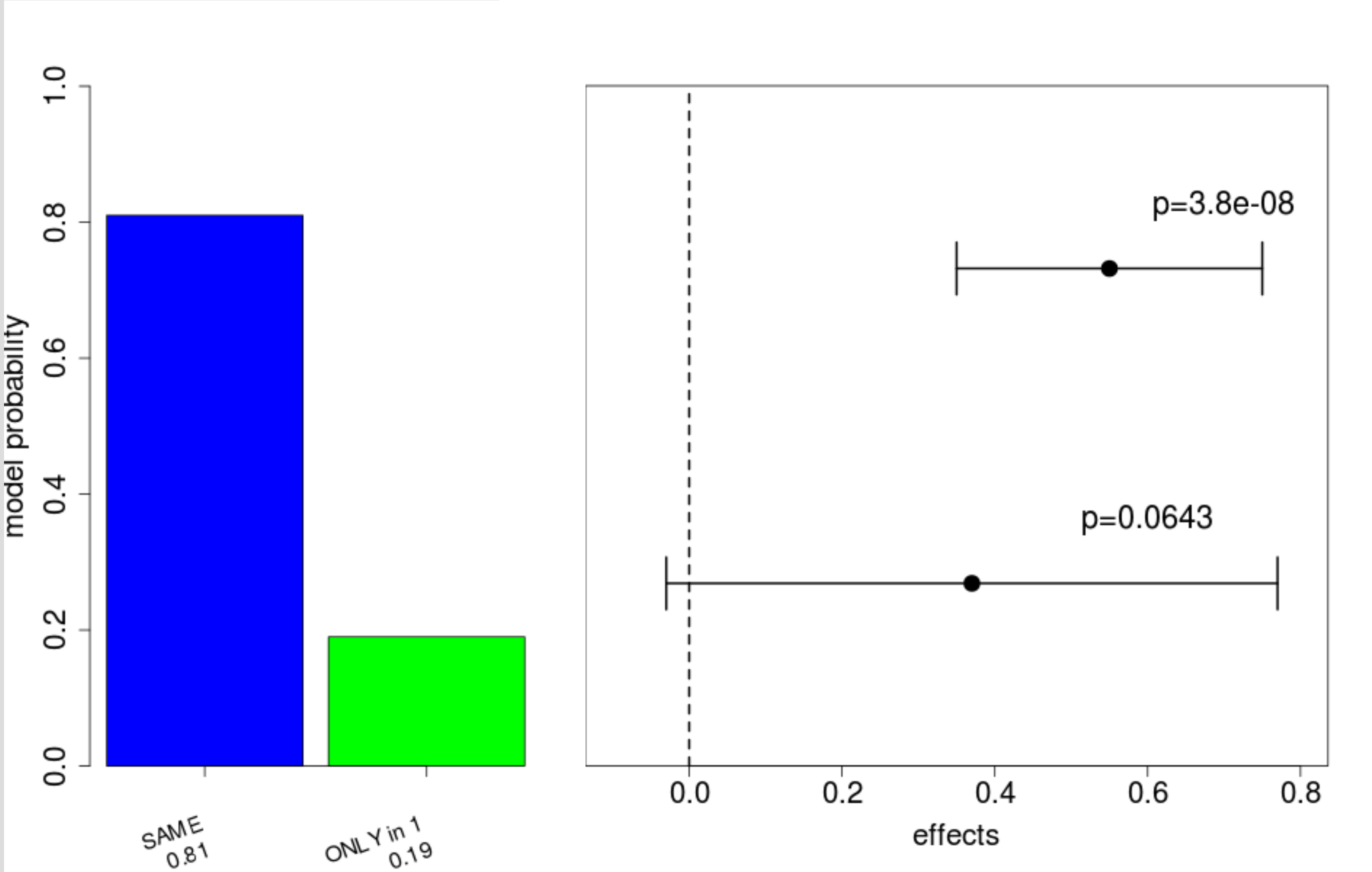
Suppose we have two estimates.

One is highly significant while the other is not.

We want to compare the same effect model with a model where the effect is present only in one of them.

How can we do that properly? These P-values alone cannot tell whether the effects are similar or different!

IS FIXED-EFFECTS ASSUMPTION REASONABLE?



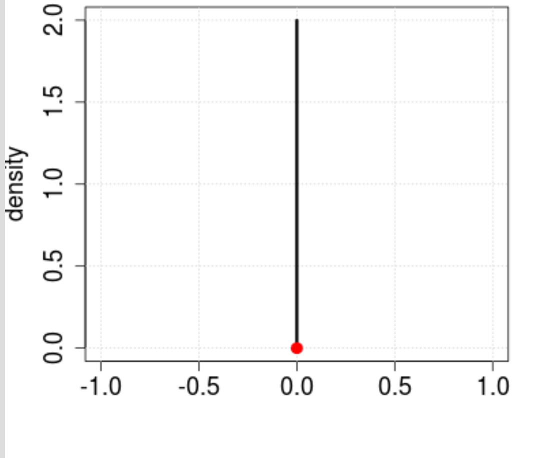
Suppose we have two estimates.

We write each of the possible explanations of the data in terms of a statistical model and compare how well each of the models describes the data by using a Bayesian model comparison framework.

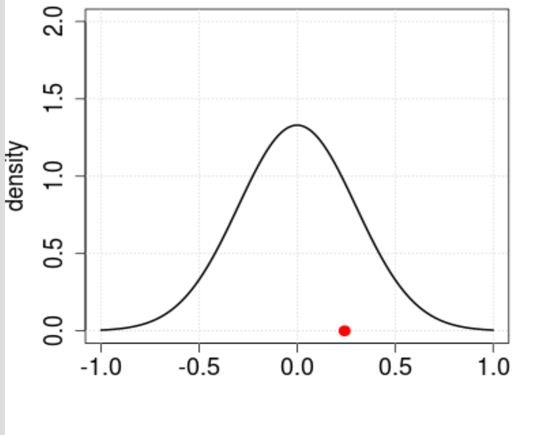
BUILDING MODELS FOR 2 EFFECTS

In GWAS 4 we compared model **E** and model **N** for one SNP.

N: (null model)



E: (effect model)

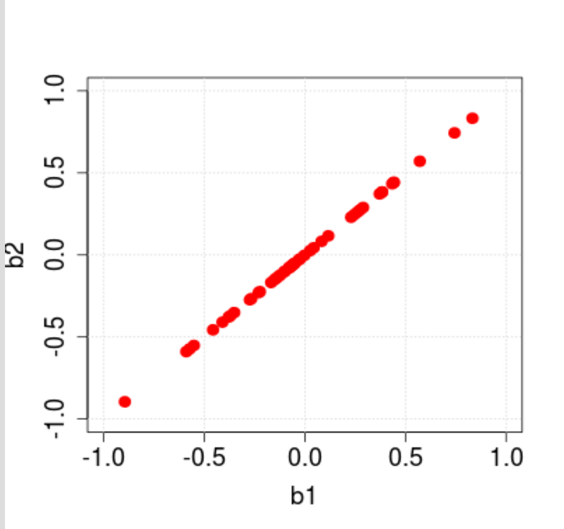


Here we have two SNPs and build joint models for them.

Example data points from each model

SAME EFFECT:

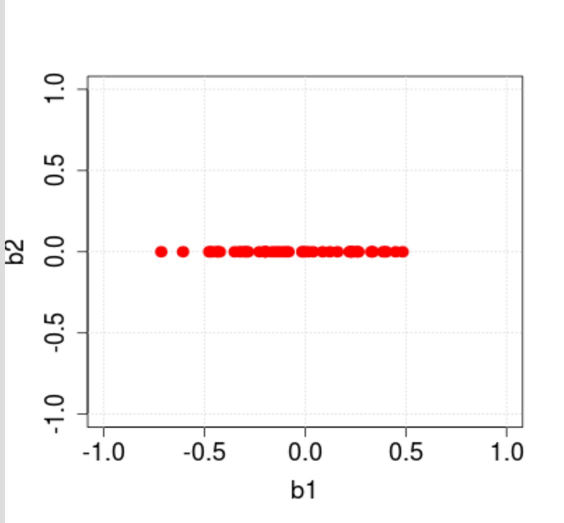
- Pick value $x \sim E$
- Set $\beta_1 = \beta_2 = x$



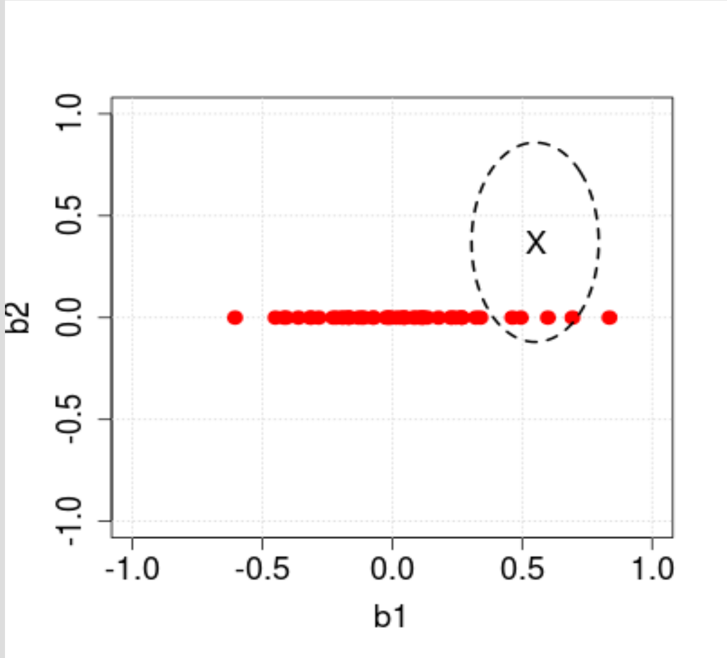
EFFECT IN ONLY 1:

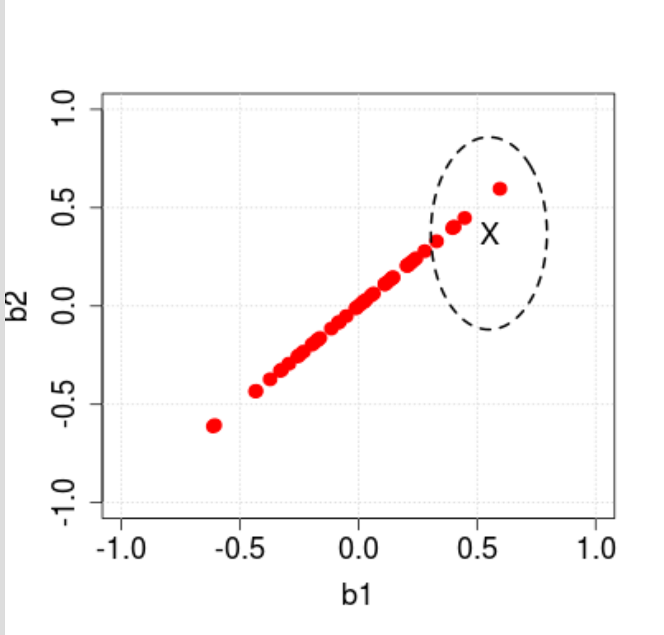
- Pick $\beta_1 \sim E$
- Pick $\beta_2 \sim N$

i.e. $\beta_2 = 0$



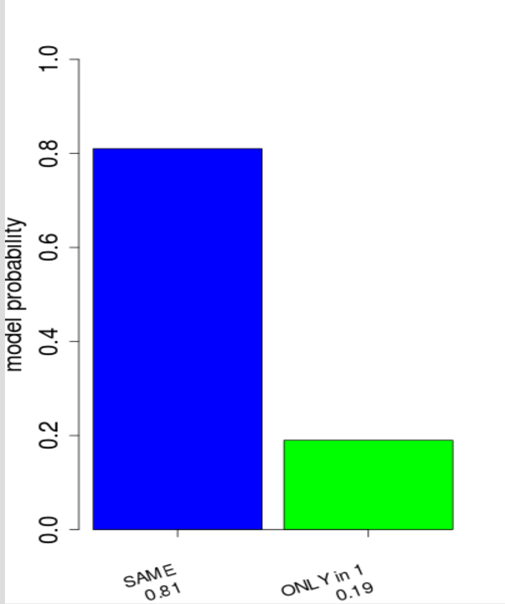
HOW WELL THE MODELS EXPLAIN THE DATA?





In our example the estimates were similar but SE of $\hat{\beta}_2$ was much larger.

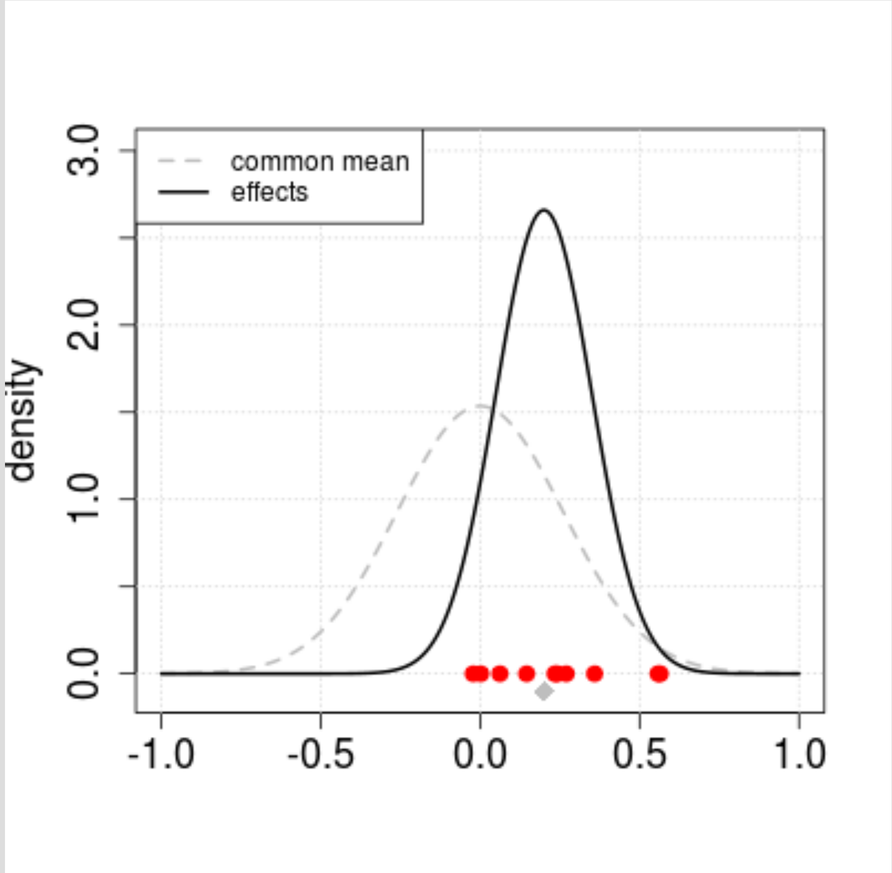
Same effect model is a better explanation here.



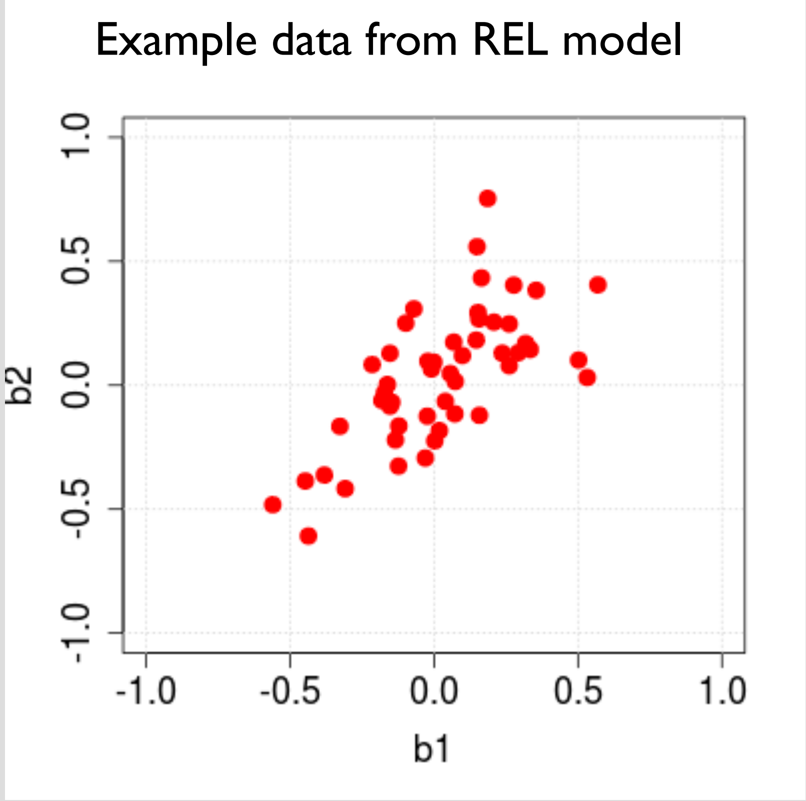
Bayes factor compares two models

$$\mathrm{BF}_{S:1} = \frac{\Pr(\mathrm{Data} \mid \text{ Model ”Same”})}{\Pr(\mathrm{Data} \mid \text{ Model ”Only in 1”})}$$

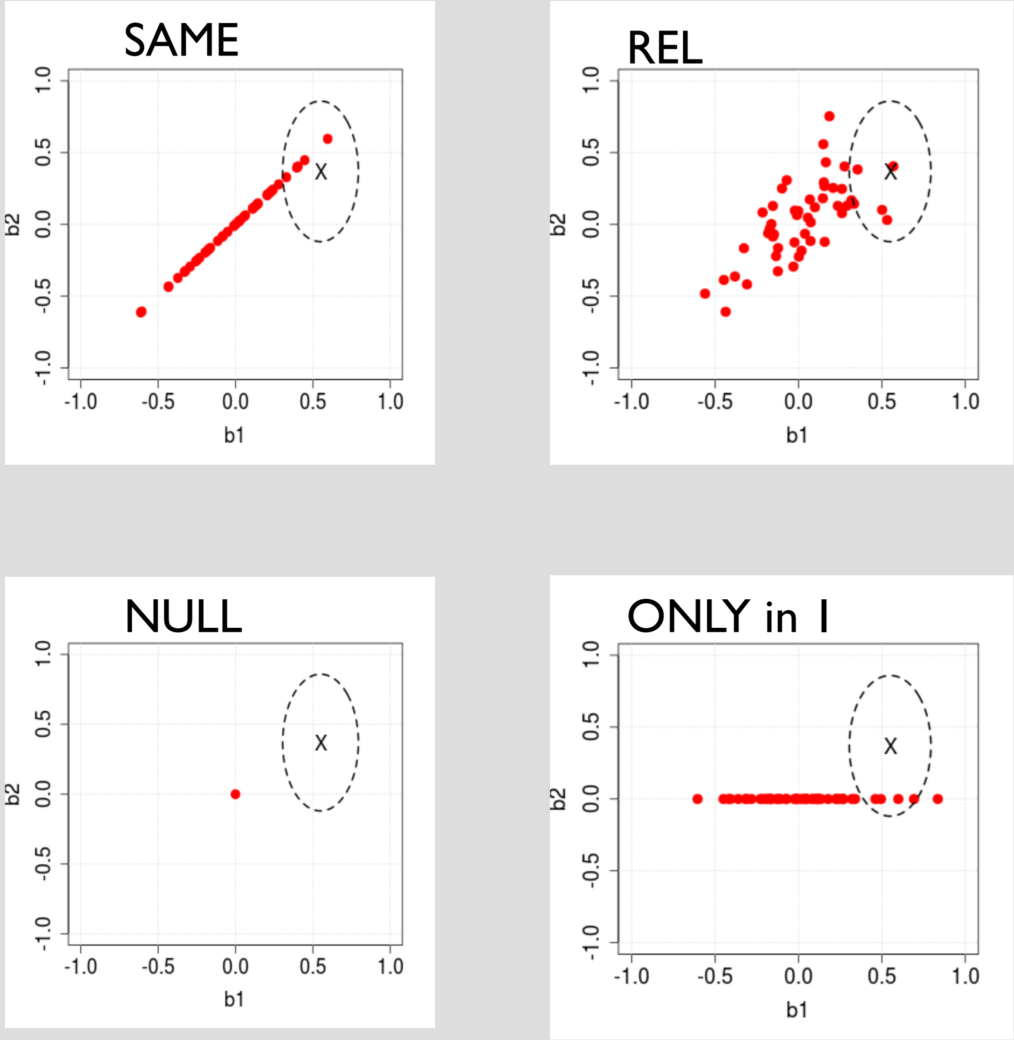
RELATED EFFECTS MODEL

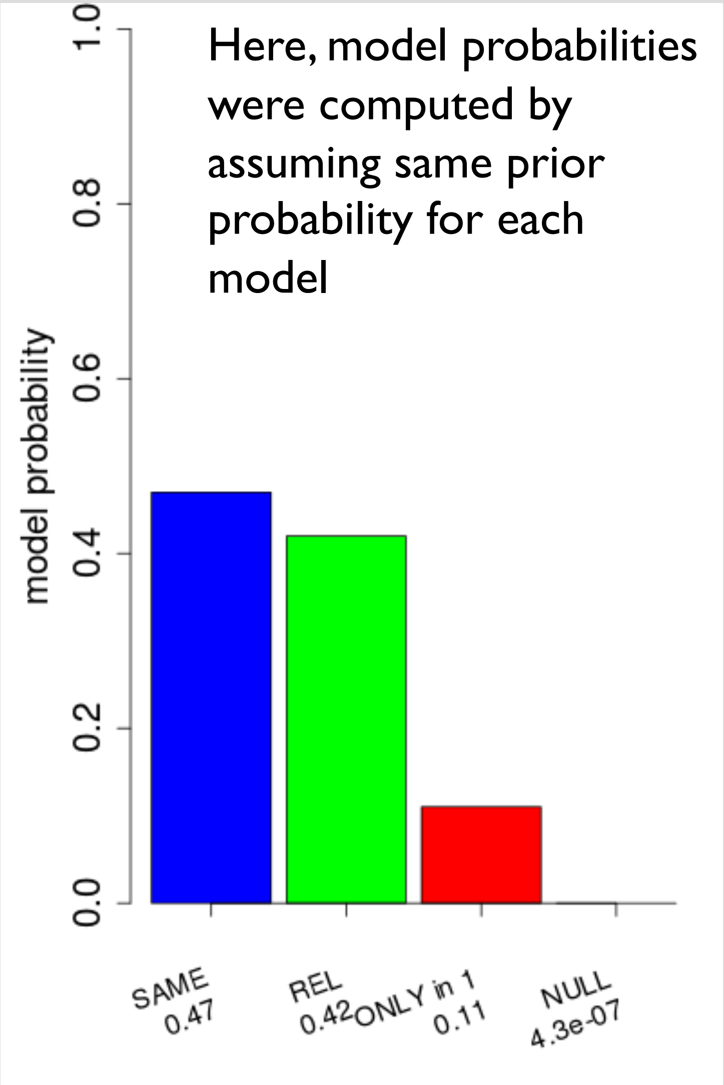


Model "REL"
- Pick $\mu \sim N(0, s^2)$
- Given μ, pick each $\beta_i \sim N(\mu, t^2)$


Example data from REL model

MODEL COMPARISON

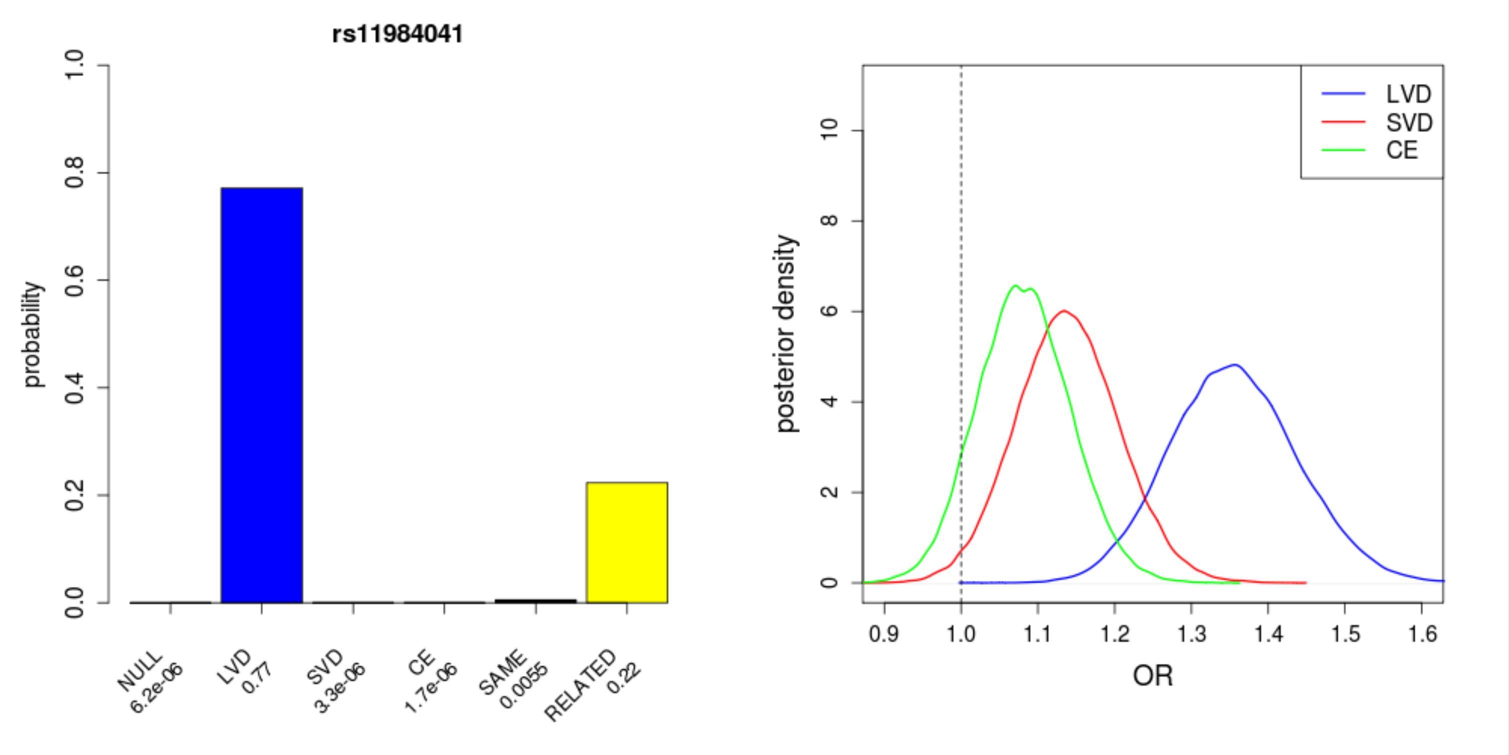




Here, model probabilities were computed by assuming same prior probability for each model

ISCHEMIC STROKE AND *HDAC9* SNP

Type	Cases	OR	P-value
LVD	844	1.42 (1.28-1.57)	2e-11
SVD	580	1.13 (1.00-1.28)	0.06
CE	790	1.10 (0.98-1.23)	0.12

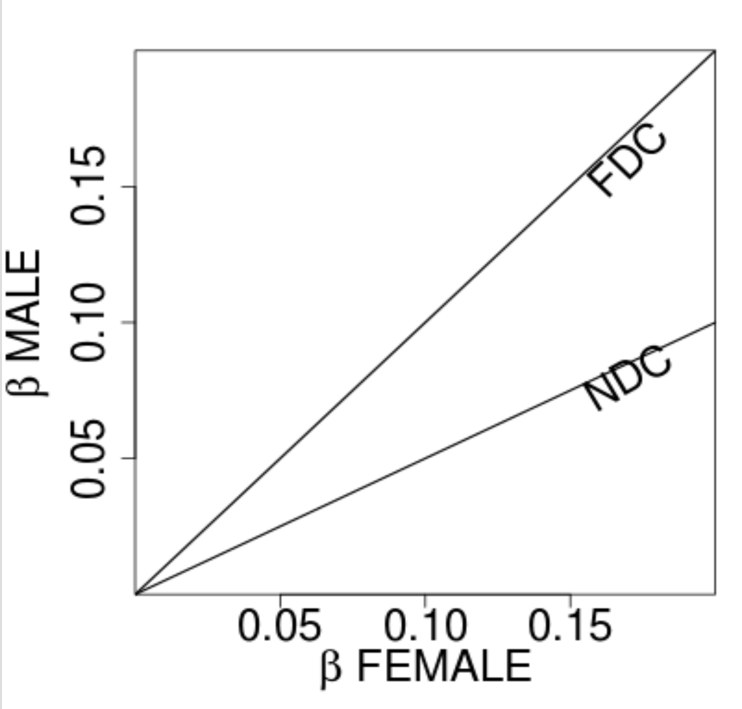


LVD = large vessel disease
SVD = small vessel disease
CE = cardioembolic stroke

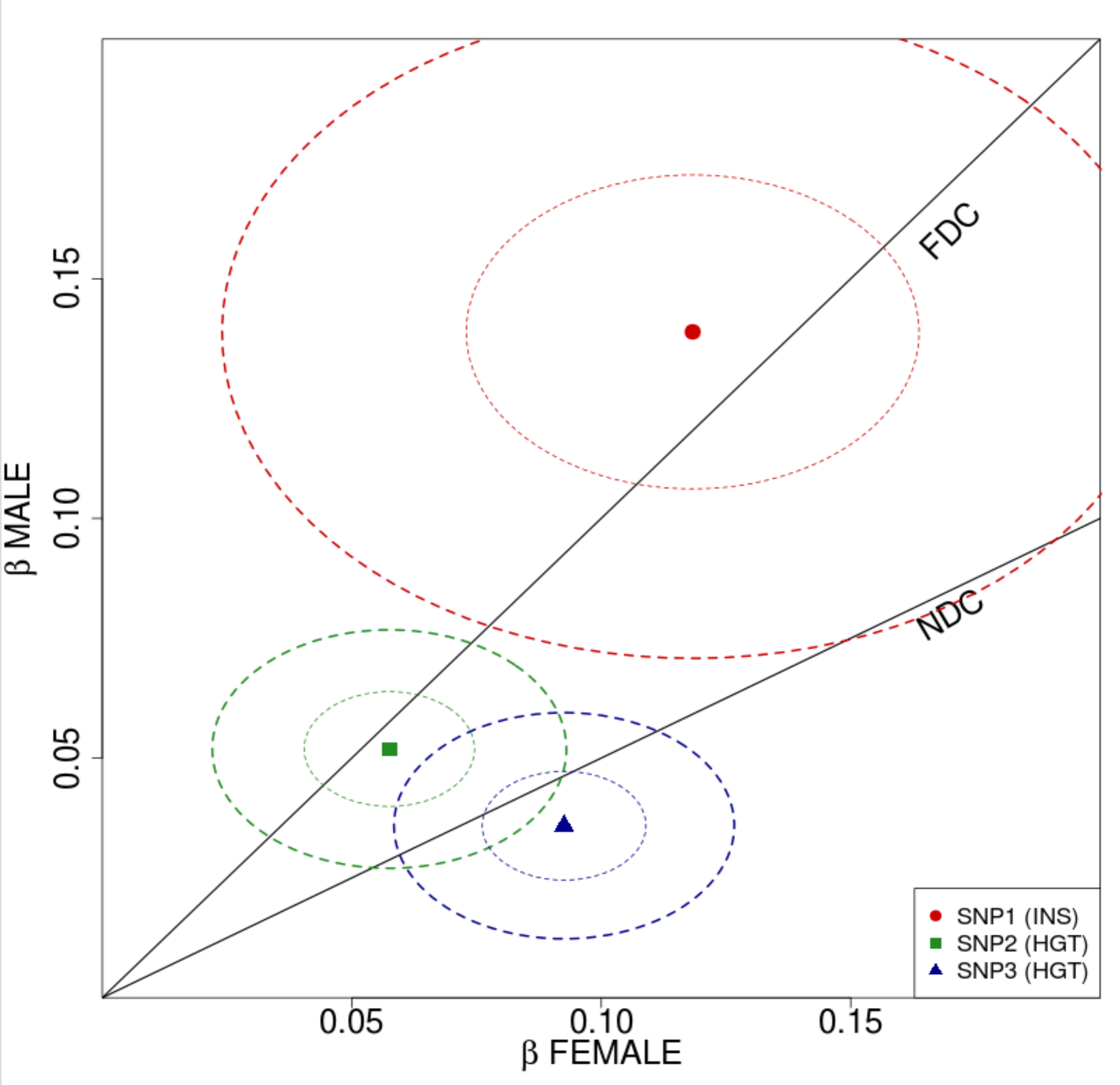
Bellenguez et al. 2012 Nat Gen

CHR X DOSAGE COMPENSATION

- One of the female's X chrs in each cell is inactivated
  - To balance difference in chr X number between the sexes (dosage compensation)
  - Inactivation is not complete, 15%-25% of genes escape from it to some degree
- Code female genotypes as 0,1 and 2 and male genotypes as 0 and 2
  - If there is full dosage compensation (FDC) i.e. complete X inactivation, then effect size in males and females is equal
  - If there is no dosage compensation (NDC) i.e. no X inactivation, then the effect size in females is twice the effect size in males

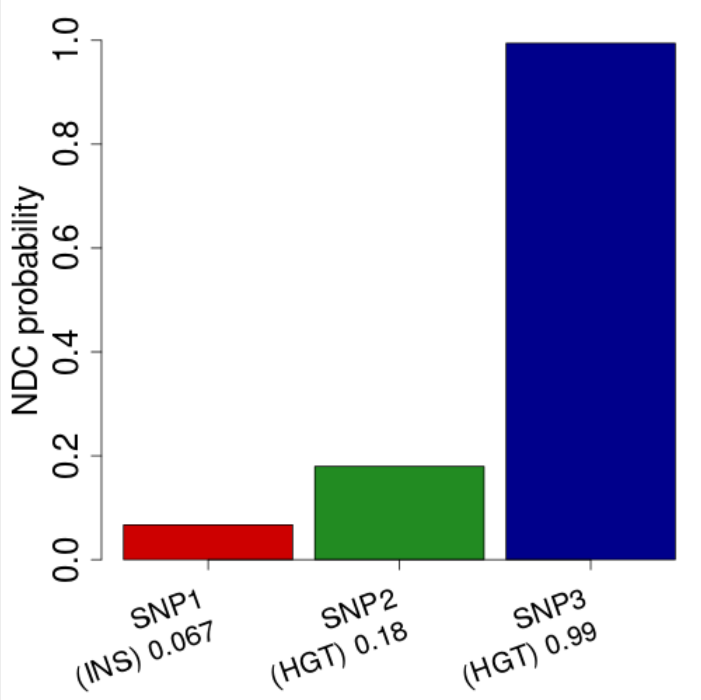


Tukiainen et al. 2014 PLoS Gen



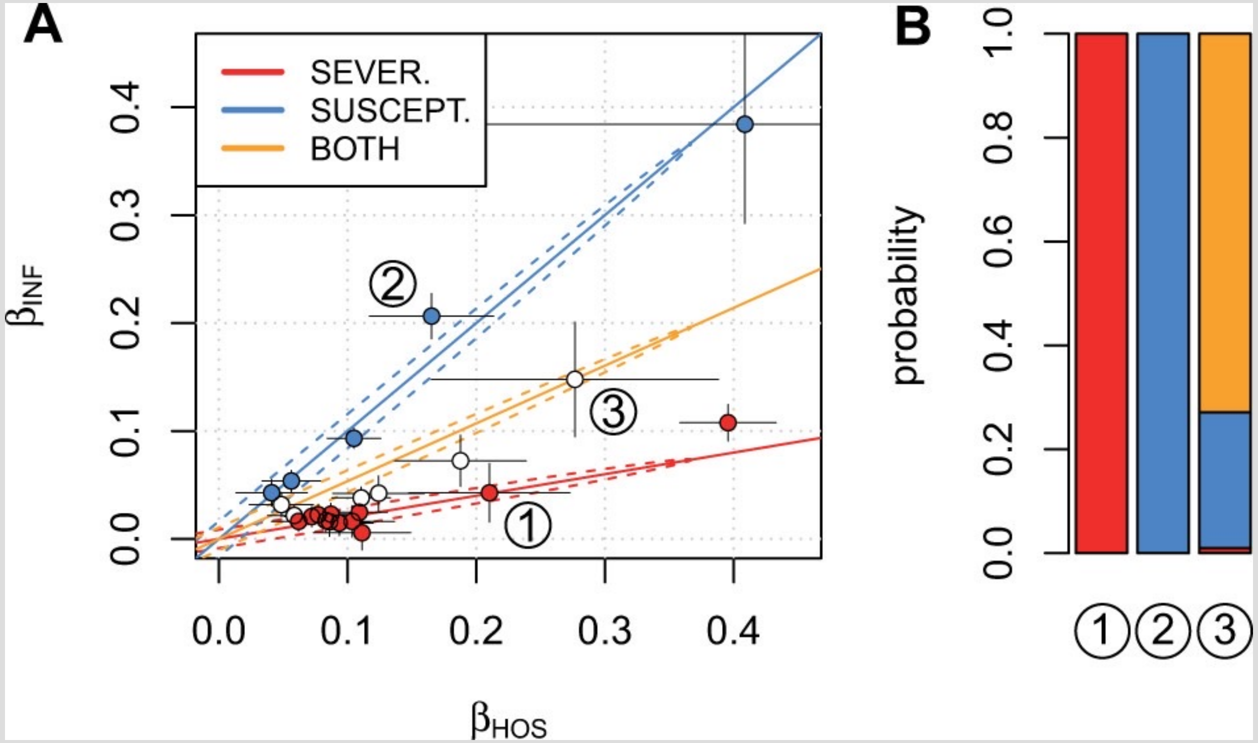
We have 3 chr X associations with Insulin levels or with height.

One of them (in *ITM2A* gene) seems to escape dosage compensation while the other two seem to follow FDC.



COVID-19 HOST GENETICS INITIATIVE

COVID19-HGI Nature 2022



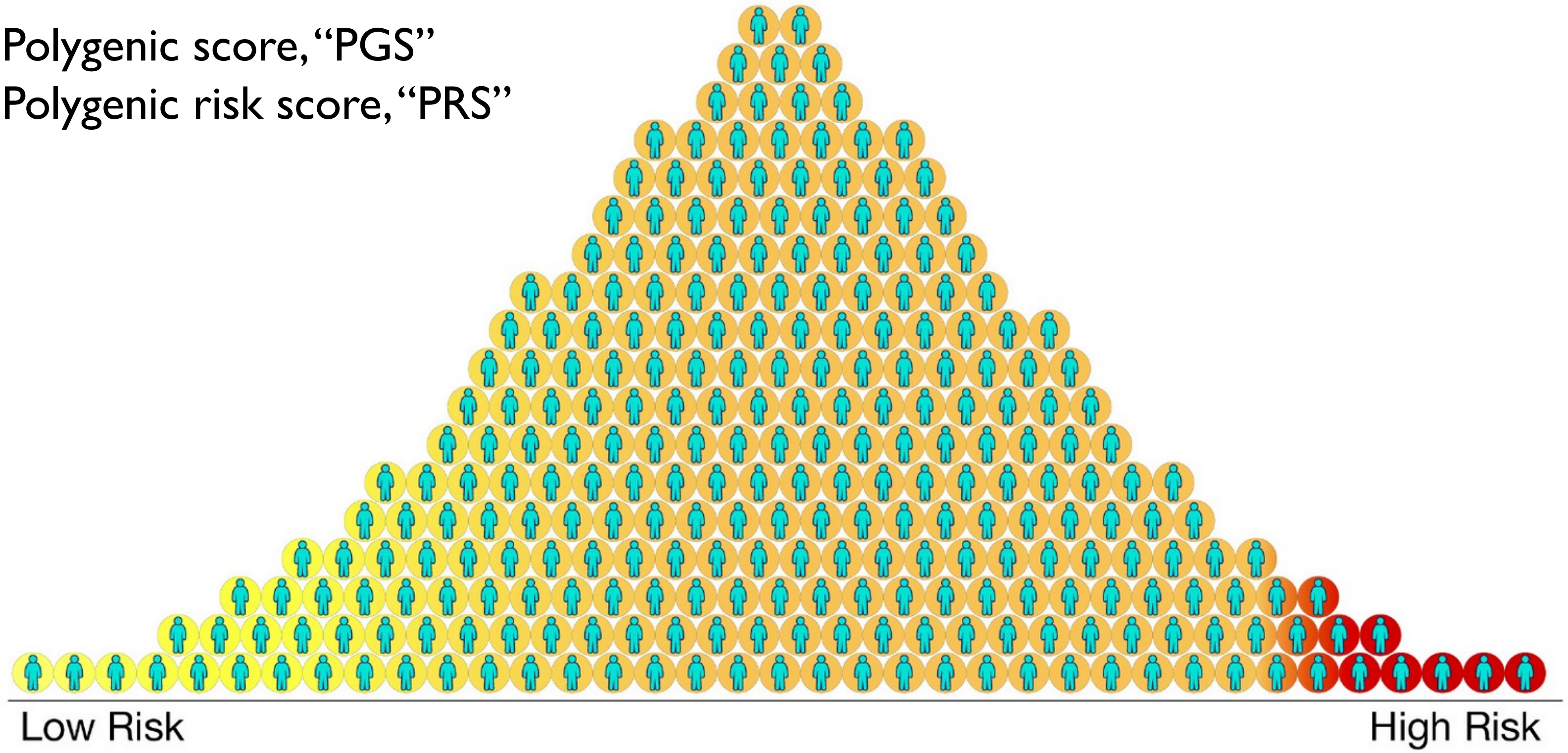
Question:
Which variants affect susceptibility to infection and which severity of the disease?
Basis for inference are differences between effect sizes from infection GWAS and hospitalization GWAS.
Figure defines line models for susceptibility and severity variants, and variants that affect both.

Figure: Pirinen 2023 Bioinformatics
GitHub mjpirinen/linemodels

COVID-19 HGI effect sizes from hospitalization (HOS) GWAS and infection (INF) GWAS for 23 variants with 95% confidence intervals. Three line models with 95% regions are shown by coloured lines. Variants with posterior probability >95% in one of the models are coloured according to the corresponding model. Three variants are labelled and posterior distributions of their assignment probabilities are shown in panel B.

POLYGENIC SCORES

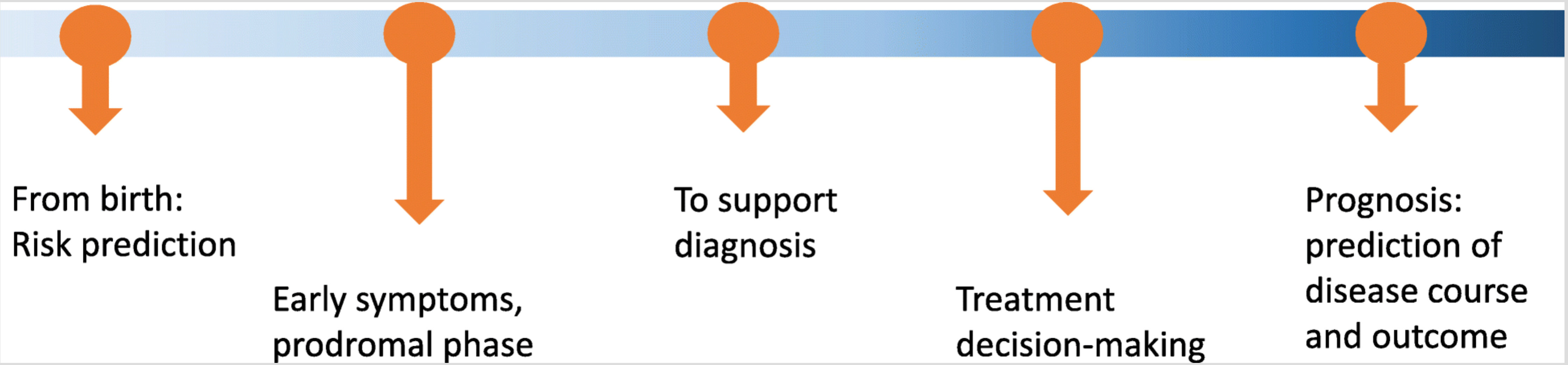
Polygenic score, "PGS"
Polygenic risk score, "PRS"



Use GWAS results to predict external individuals' risk for a disease from his/her genotypes.

Figure: NIH

(FUTURE) USES OF GENETIC SCORES

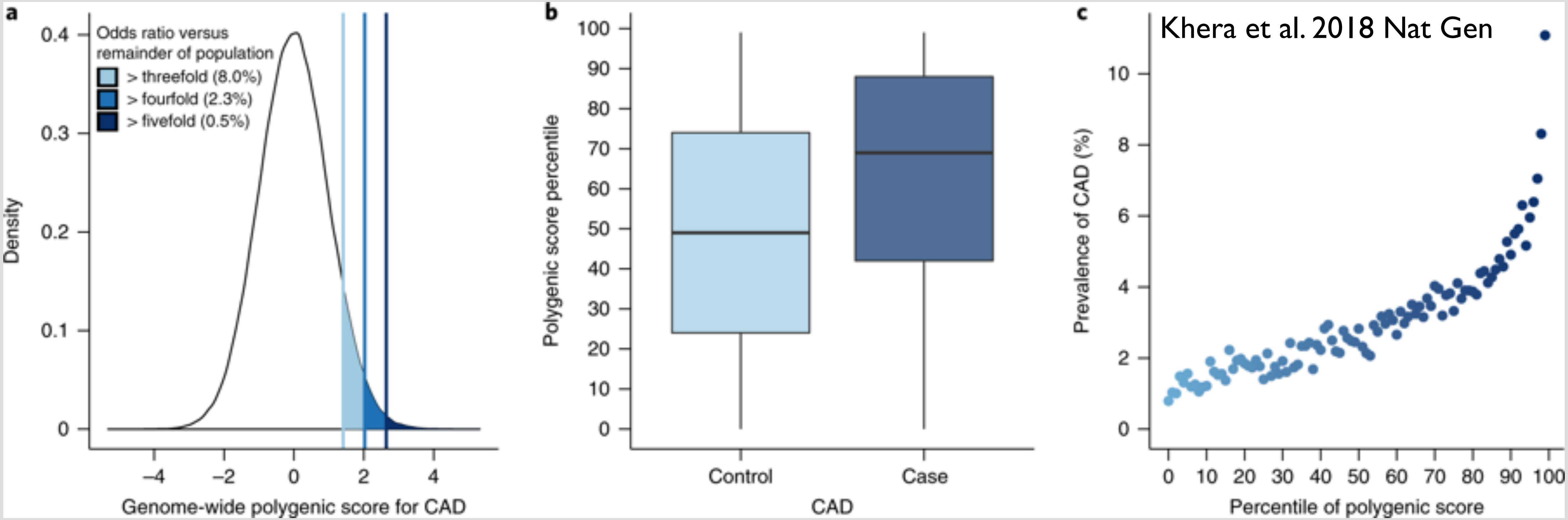


Help in prevention

- lifestyle change
- screening programmes

How best to treat this person?

Lewis & Vassos, Genome Medicine **12:** 44 (2020)



Khera et al. 2018 Nat Gen

a. Distribution of PGS_{CAD} in the UK Biobank testing dataset (*n* = 288,978). The *x* axis represents PGS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation.

Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, PGS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping.

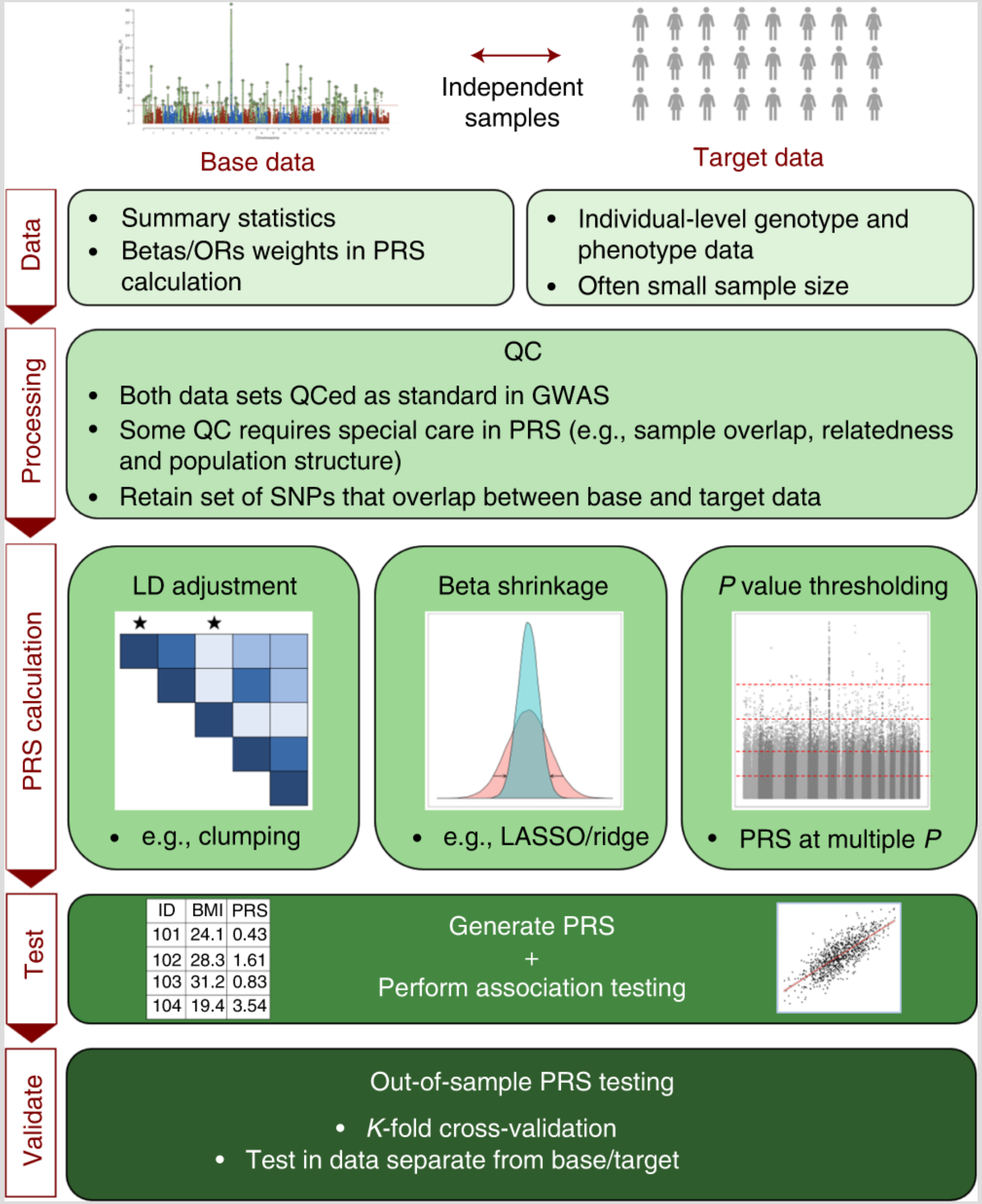
c, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the PGS_{CAD}.

GENERATING POLYGENIC SCORES

- Take allelic effect estimates ($\hat{\beta}_k$) from GWAS
  - Ideally causal effects estimated by multiple regression but often marginal effects used
- Take target individual’s genotypes (g_{il}) at the loci $\mathrm{l} = 1, \ldots L$
- Compute PRS for individual *i* as sum

$$PRS_i = \sum_{l=1}^{\#Loci} g_{il}\beta_k$$

Choi et al. 2020 Nat Protocols



- Summary statistics
- Betas/ORs weights in PRS calculation

- Individual-level genotype and phenotype data
- Often small sample size

QC

- Both data sets QCed as standard in GWAS
- Some QC requires special care in PRS (e.g., sample overlap, relatedness and population structure)
- Retain set of SNPs that overlap between base and target data

LD adjustment

- e.g., clumping

Beta shrinkage

- e.g., LASSO/ridge

P value thresholding

- PRS at multiple *P*

ID	BMI	PRS
101	24.1	0.43
102	28.3	1.61
103	31.2	0.83
104	19.4	3.54

Generate PRS
+
Perform association testing

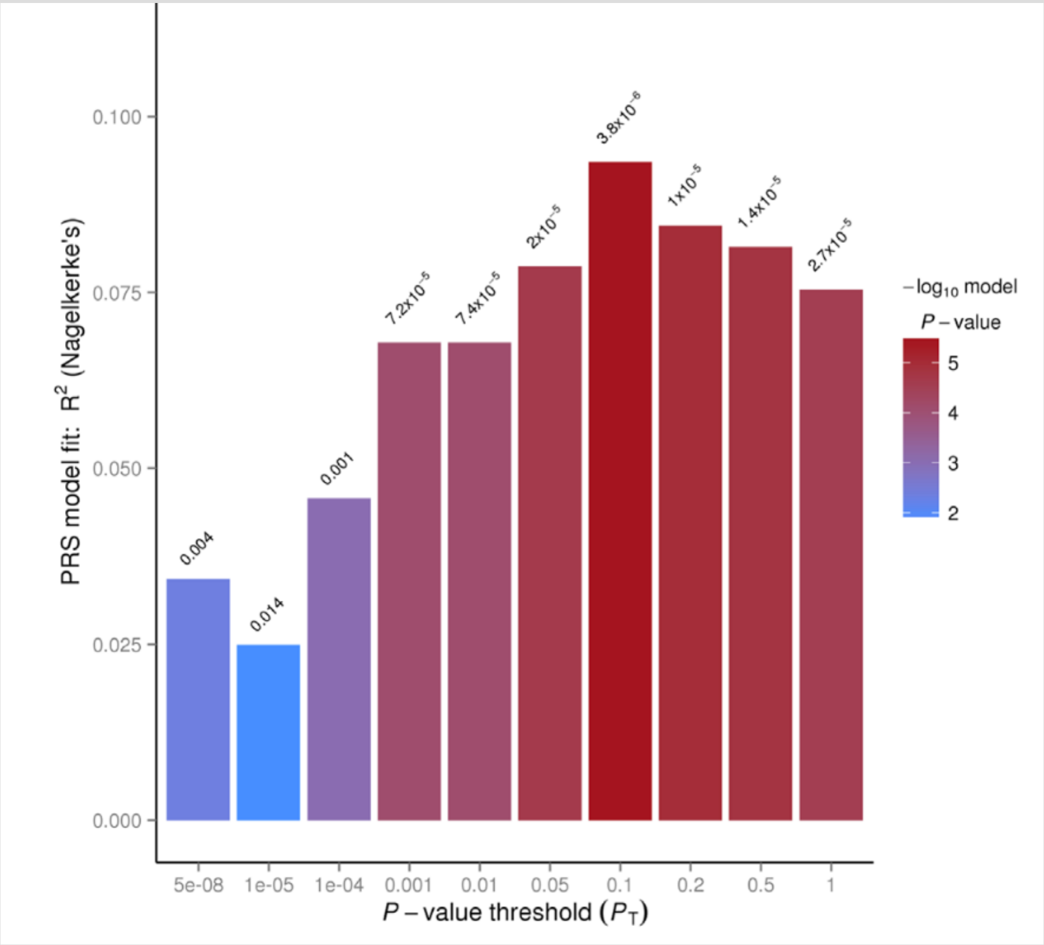
Out-of-sample PRS testing

- *K*-fold cross-validation
- Test in data separate from base/target

STANDARD PRS METHOD: CLUMPING & THRESHOLDING

- Consider only SNPs with GWAS P-value < P_{thr} , where P_{thr} is a threshold
- From two SNPs that are in LD > r^2 , choose the one with a smaller GWAS P-value
  - This forms "clumps" of "significant" SNPs in LD with each other and only picks the most "significant" ANP as the only representative of the clump
  - A light version of conditional analysis where no joint regression is used but r^2 value alone determines whether two SNPs have "independent signals"
- Use marginal allelic effect estimates in PRS calculation
- Tune parameters P_{thr} and r^2 in a validation set to optimize performance

CHOOSING THRESHOLDS



Vassos et al. Biological Psychiatry, 2017; 81:470–477

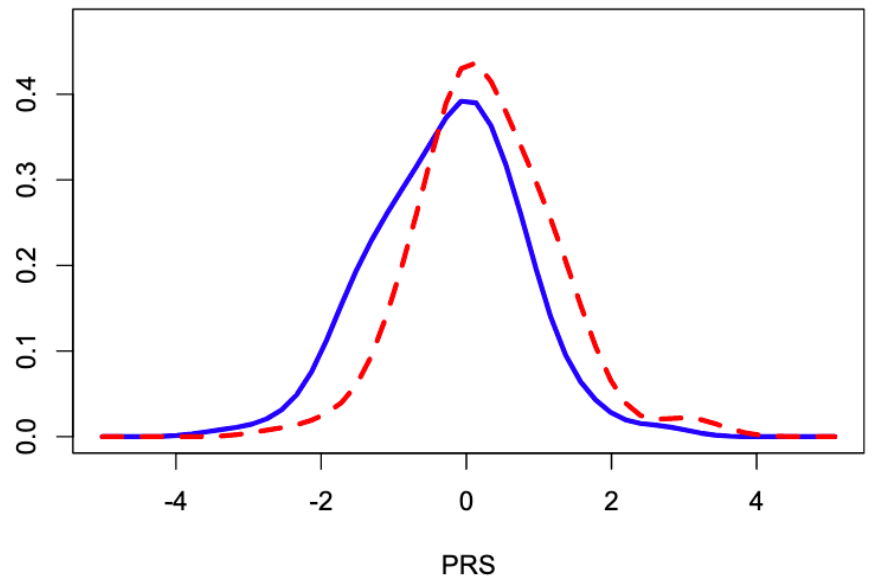


Figure 2. Density distribution of polygenic risk score (PRS) in European first-episode psychosis case and control subjects. PRS represents the standardized residuals of PRS after adjustment for the 10 principal components. Blue line indicates control subjects; red line, case subjects.

Goal: Predicting psychosis cases by schizophrenia PRS.

Left: Optimal PRS uses P_{thr} = 0.1.

r^2 threshold was fixed to 0.1 (not tuned).

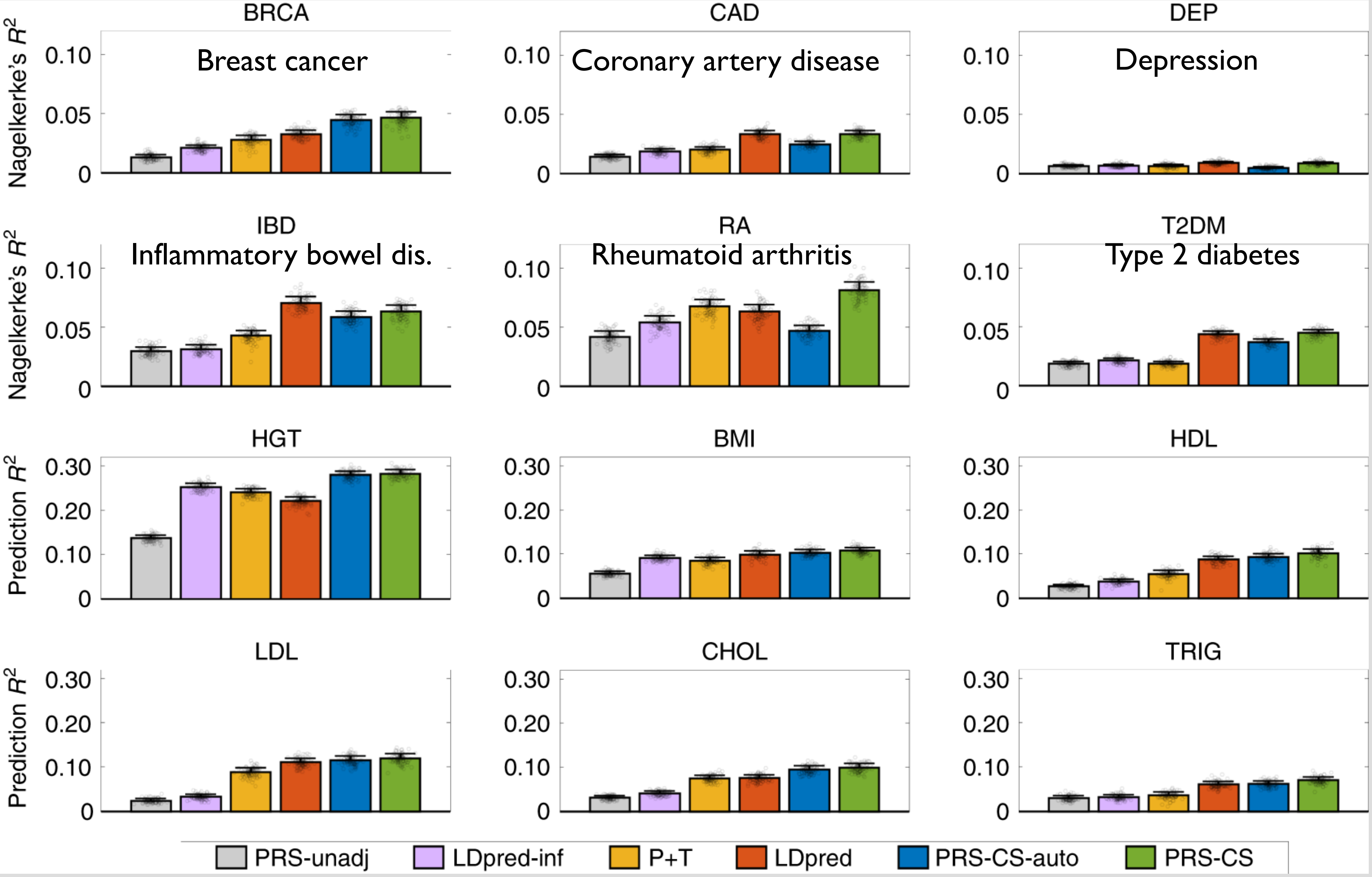
Computed using PRSice software.

LDPRED (VILHJALMSSON ET AL. AJHG 97:576-592)

- Assume prior $\lambda_l \sim \begin{cases} N\left(0,\ \frac{h^2}{p\theta}\right), \text{with prob. } \theta \\ 0, \text{with prob. } 1-\theta \end{cases}$,

  where h^2 is heritability and p is #SNPs

- Given marginal GWAS effects $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_l)$ and SEs, LDpred computes posterior expectation of the causal effects $E(\boldsymbol{\lambda} \mid \widehat{\boldsymbol{\beta}}, \boldsymbol{R}, h^2, \theta)$, where $\boldsymbol{R}$ is the LD matrix.
  - In practice, LD-matrix is considered only within a certain window
  - h^2 could be estimated externally using LMM or LDSC
  - Grid of θ values are evaluated and the best performing model is chosen
- These estimated causal effects are used as weights in PRS computation



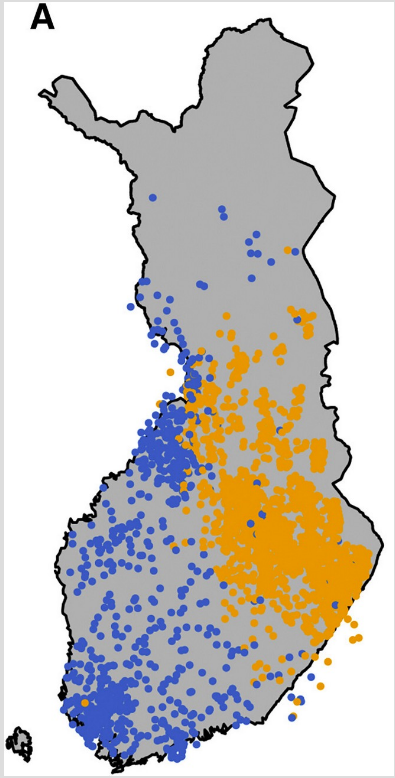
Ge et al. Nat Comm 2019

Inf = infinitesimal model P+T= pruning & thresholding PRS-CS = Another Bayesian method

BIASES

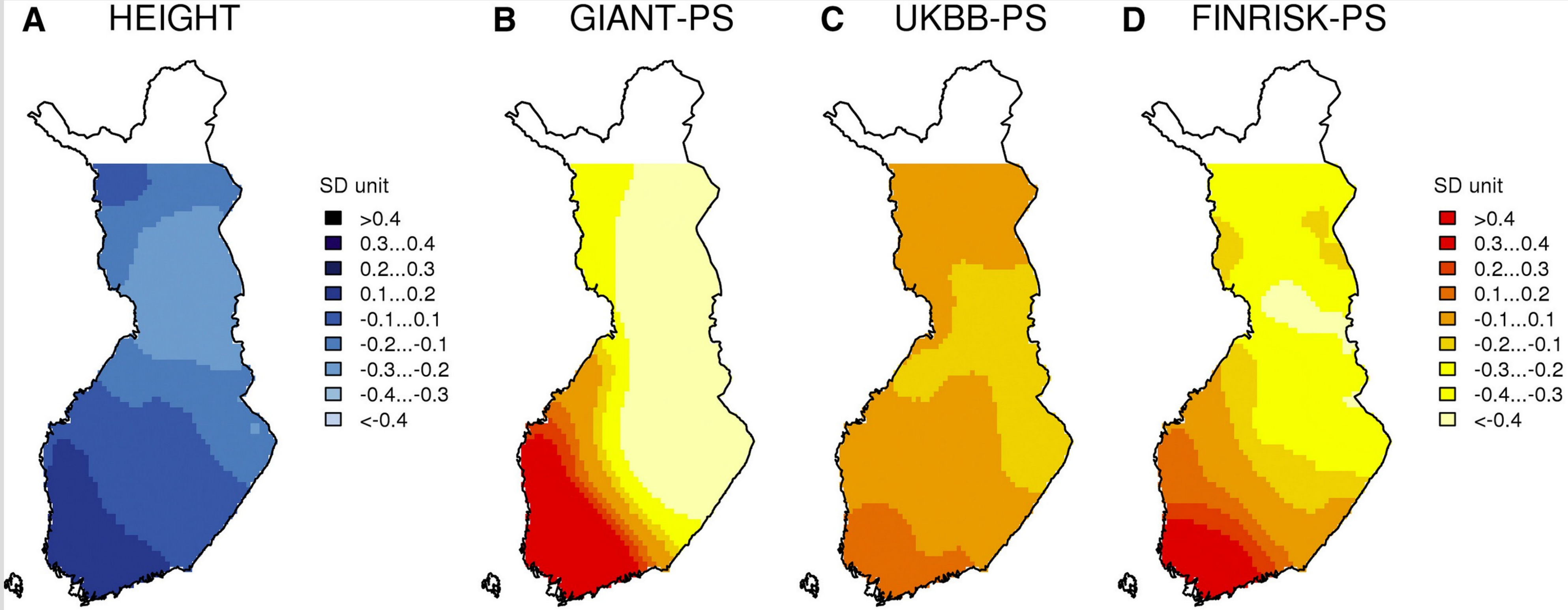
- If PGS is used to predict phenotypes of individuals who were included in the base GWAS, the prediction will be dramatically over optimistic
  - Make sure there is no overlap between GWAS and target sample
- Even if there is no overlap, relatedness and population structure can cause biases
- PGS based on European ancestry GWAS do not work equally well in other ancestries

PREDICTING HEIGHT IN FINLAND



Main pop. structure

Kerminen et al. 2019
AJHG



Distribution of
height
(age, sex adjusted)

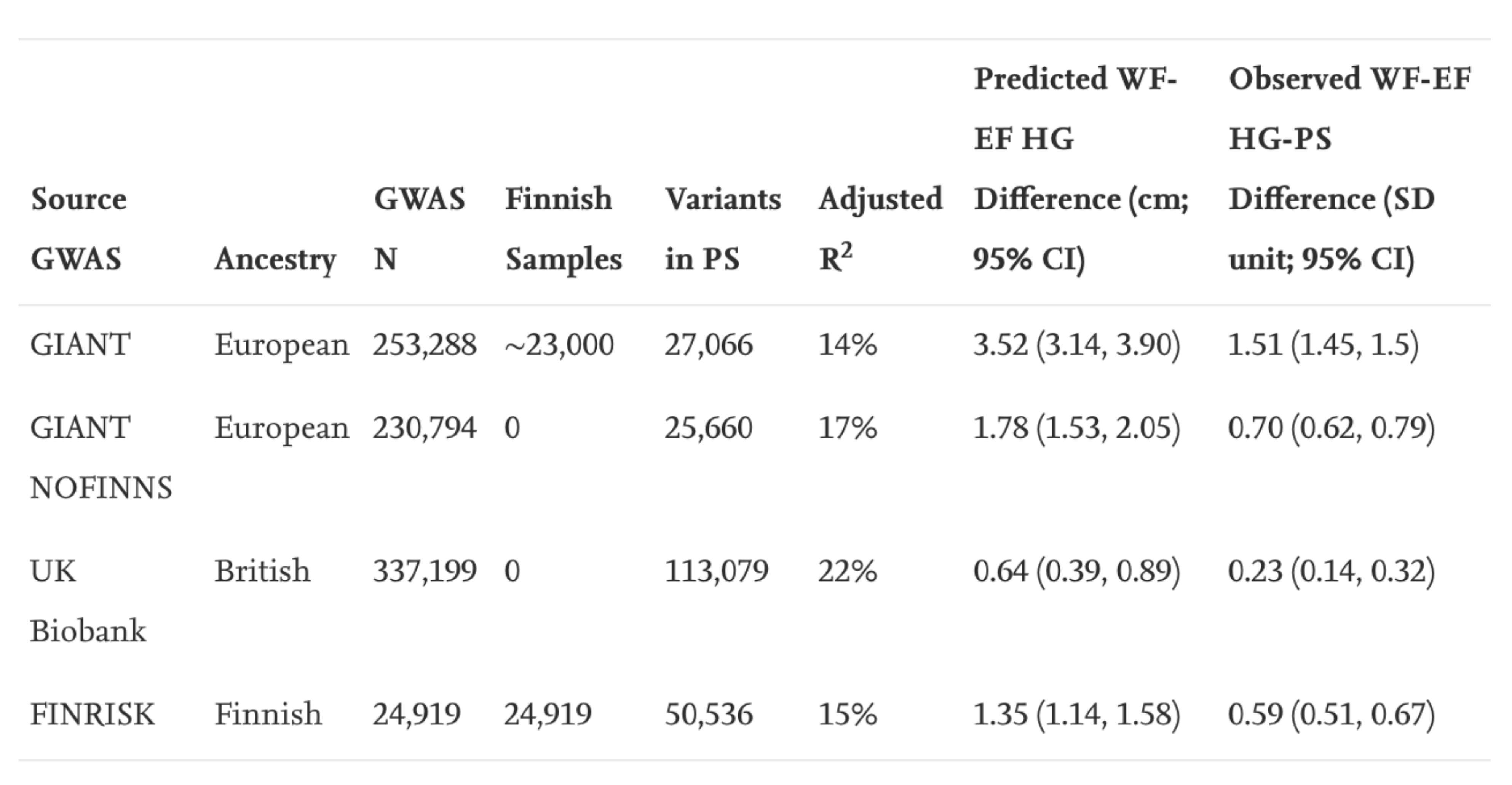
GIANT GWAS
N = 250,000
Includes Finns

UKBB GWAS
N = 337,000
No Finns

FINRISK GWAS
N = 24,000
All Finns

COMPARING PREDICTIONS

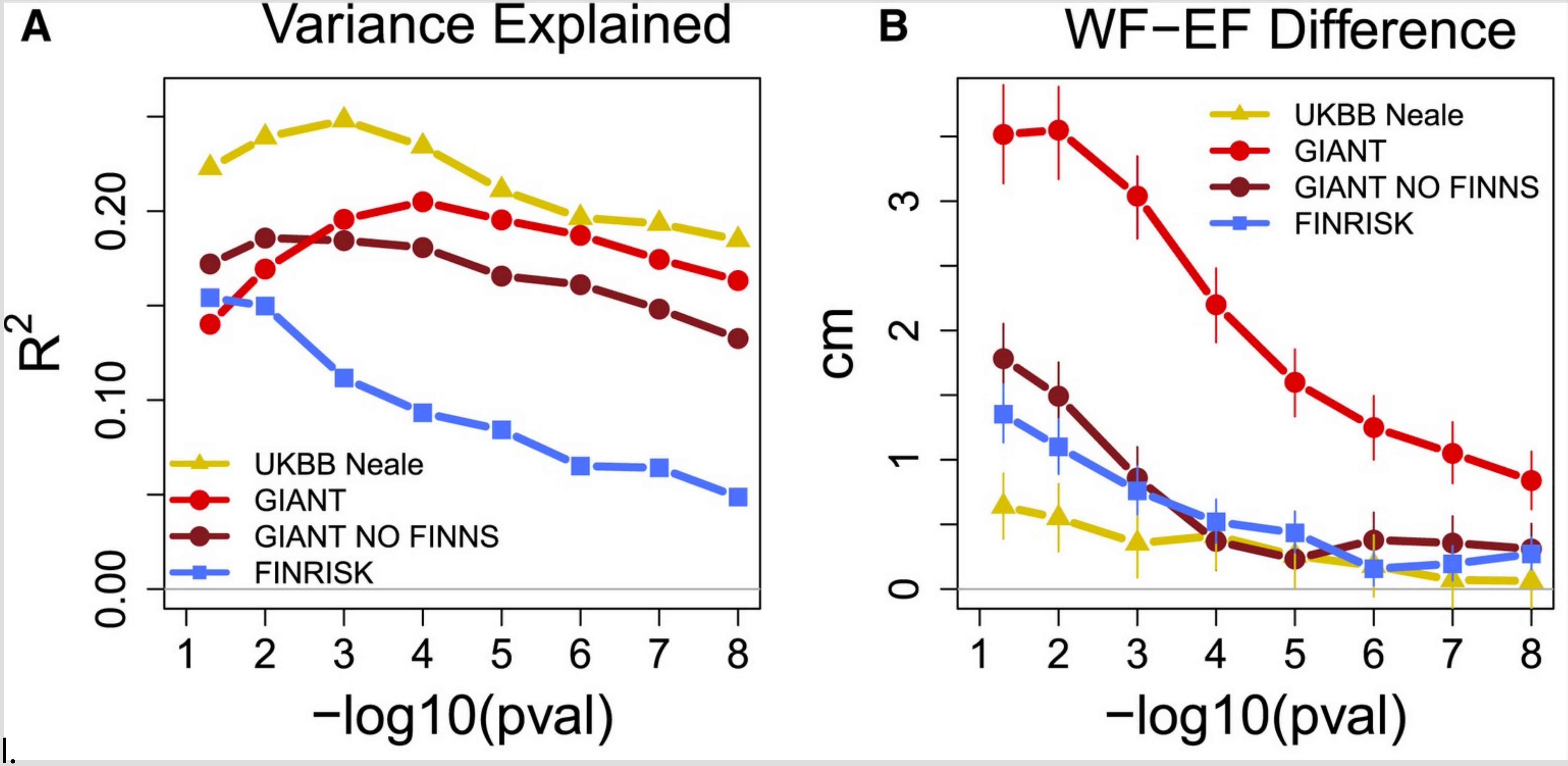
Kerminen et al. 2019
AJHG



Source GWAS	Ancestry	GWAS N	Finnish Samples	Variants in PS	Adjusted R^2	Predicted WF-EF HG Difference (cm; 95% CI)	Observed WF-EF HG-PS Difference (SD unit; 95% CI)
GIANT	European	253,288	~23,000	27,066	14%	3.52 (3.14, 3.90)	1.51 (1.45, 1.5)
GIANT NOFINNS	European	230,794	0	25,660	17%	1.78 (1.53, 2.05)	0.70 (0.62, 0.79)
UK Biobank	British	337,199	0	113,079	22%	0.64 (0.39, 0.89)	0.23 (0.14, 0.32)
FINRISK	Finnish	24,919	24,919	50,536	15%	1.35 (1.14, 1.58)	0.59 (0.51, 0.67)

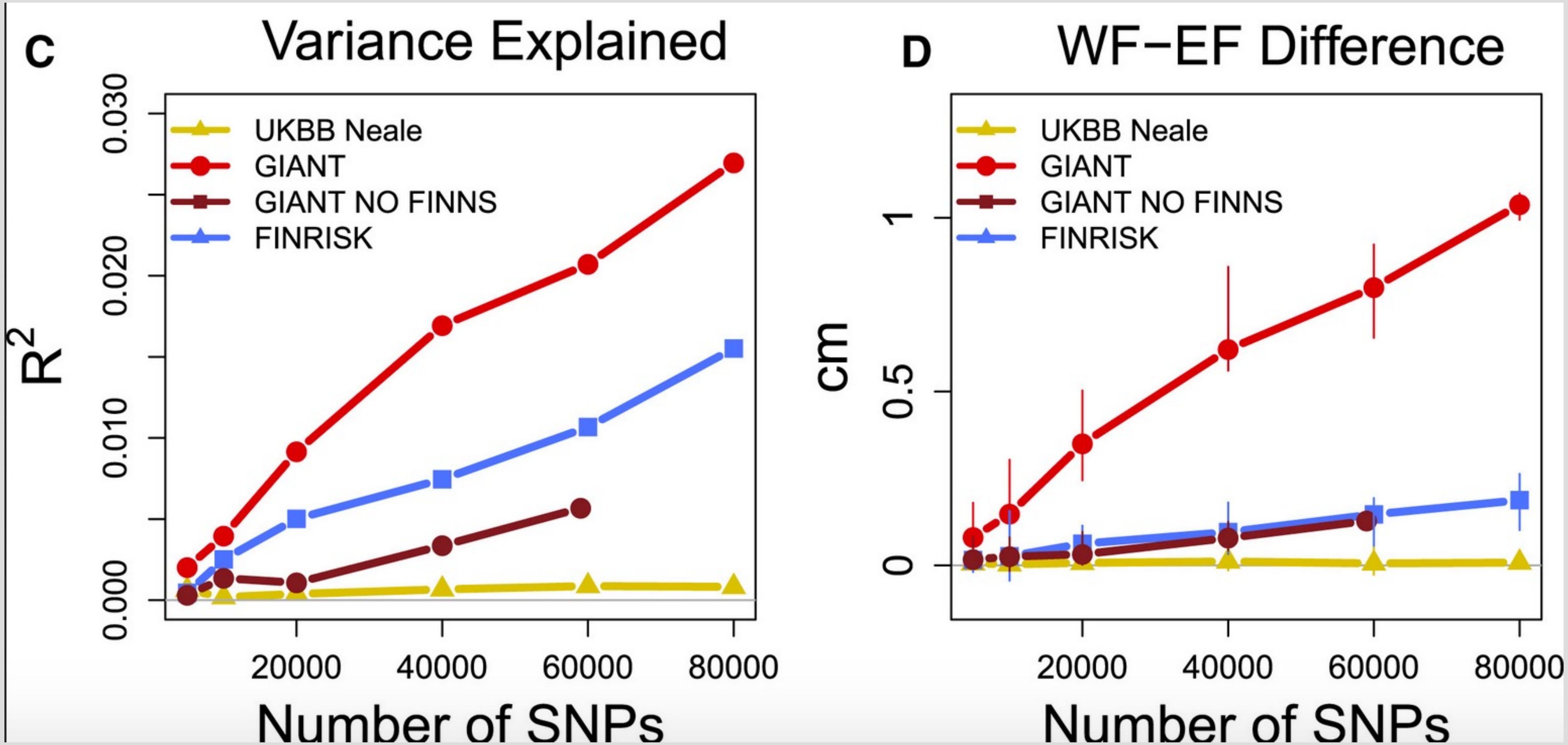
True East-West height difference is 1.6 cm, and these PGS should only predict < 1/3 of it.

PERFORMANCE DEPENDS ON P_{THR}



Kerminen et al.
2019 AJHG

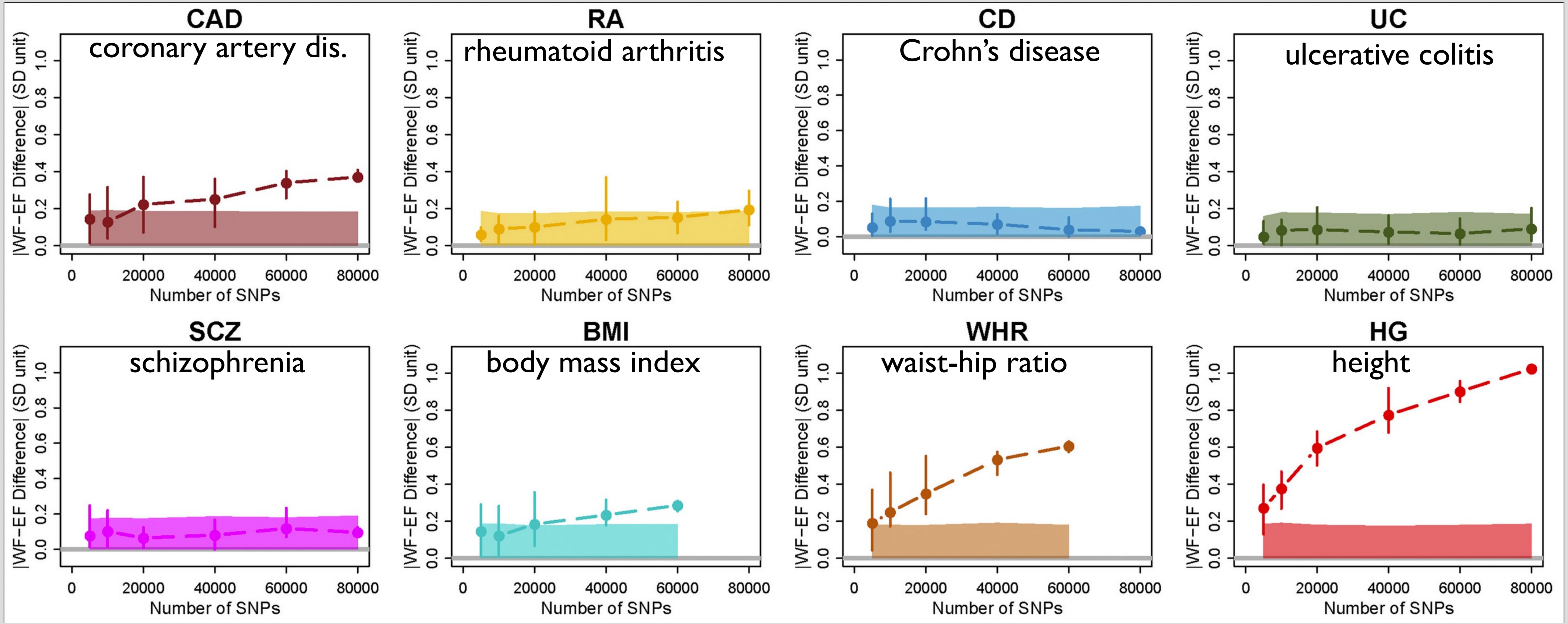
"RANDOM" PGS



Kerminen et al. 2019
AJHG

To test the suspiciously large East-West differences in predicted genetic height,
Include only SNPs that have P > 0.5, and should not be associated with height.

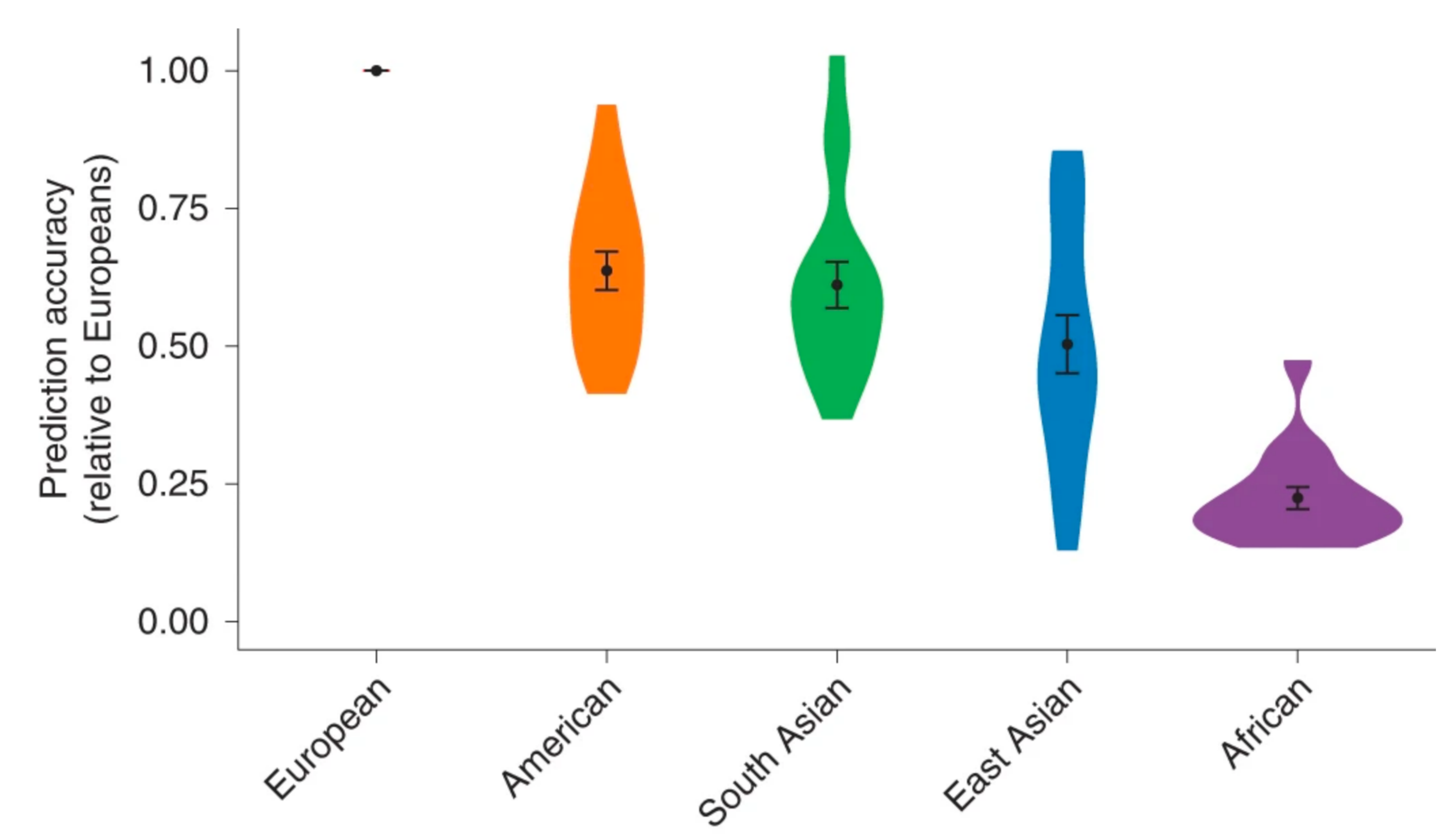
RANDOM SCORES



Kerminen et al. 2019 AJHG

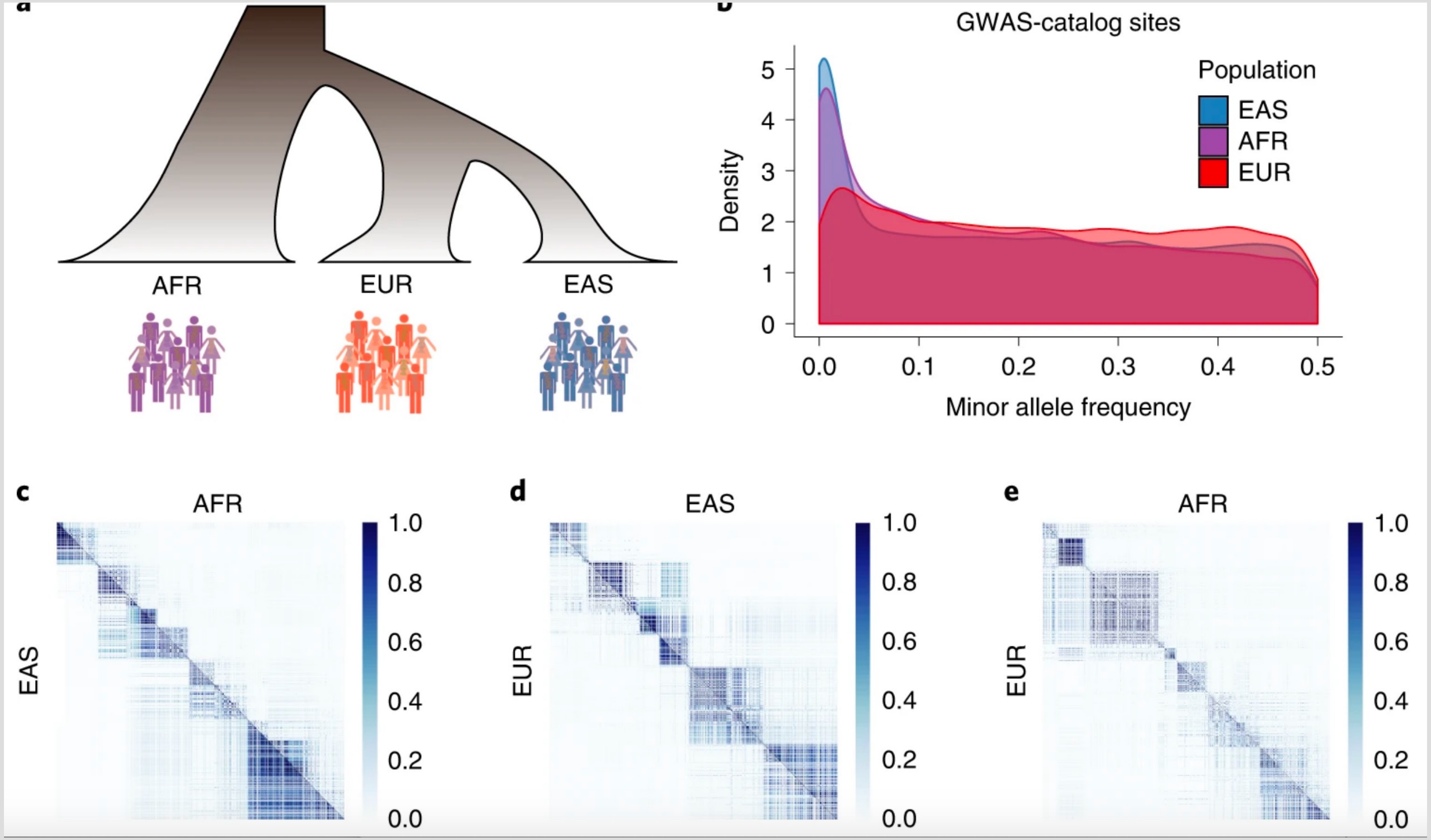
Possible problems present in CAD, BMI, WHR and Height.

LACK OF TRANSFERRABILITY BTW POPULATIONS



Martin et al. 2019 Nat Gen: **Clinical use of current polygenic risk scores may exacerbate health disparities**

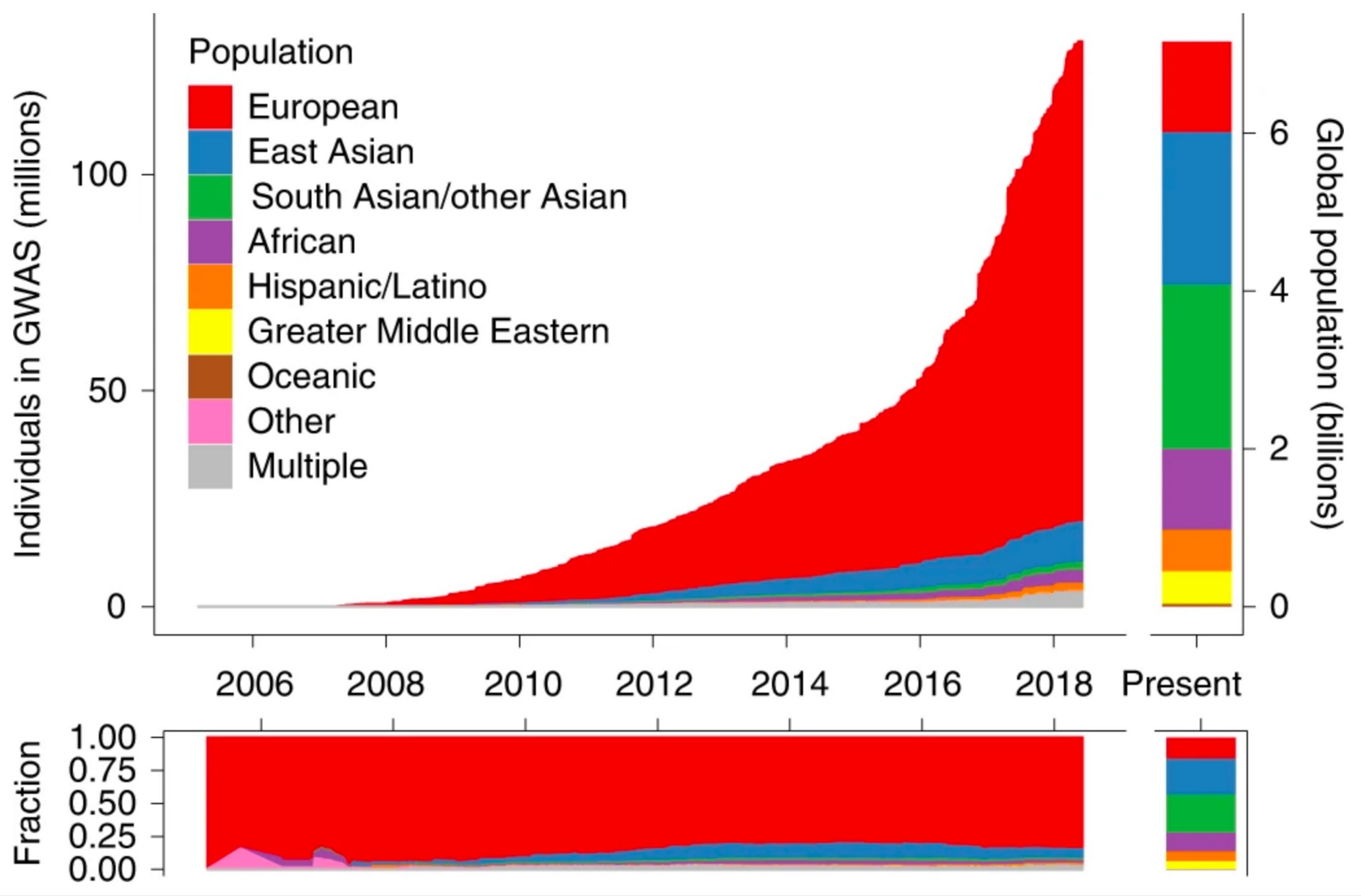
SOME CAUSES FOR DISPARITIES



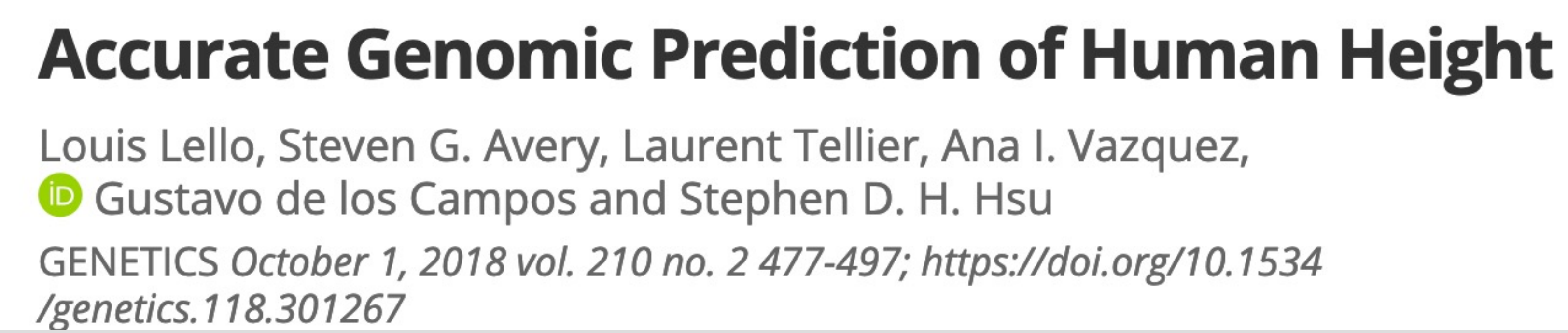
AFR, continental African;
EUR, European;
EAS, East Asian.
a, Relationships among populations.
b, Allele frequency distributions of variants from the GWAS catalog.
c–**e**, Color axis shows LD scale (r^2) for the indicated LD comparisons between pairs of populations; Illustrating variable LD patterns across populations.

Martin et al. 2019n Nat Gen

DIVERSITY CURRENTLY LACKING IN GWAS DATA



Martin et al. 2019
Nat Gen



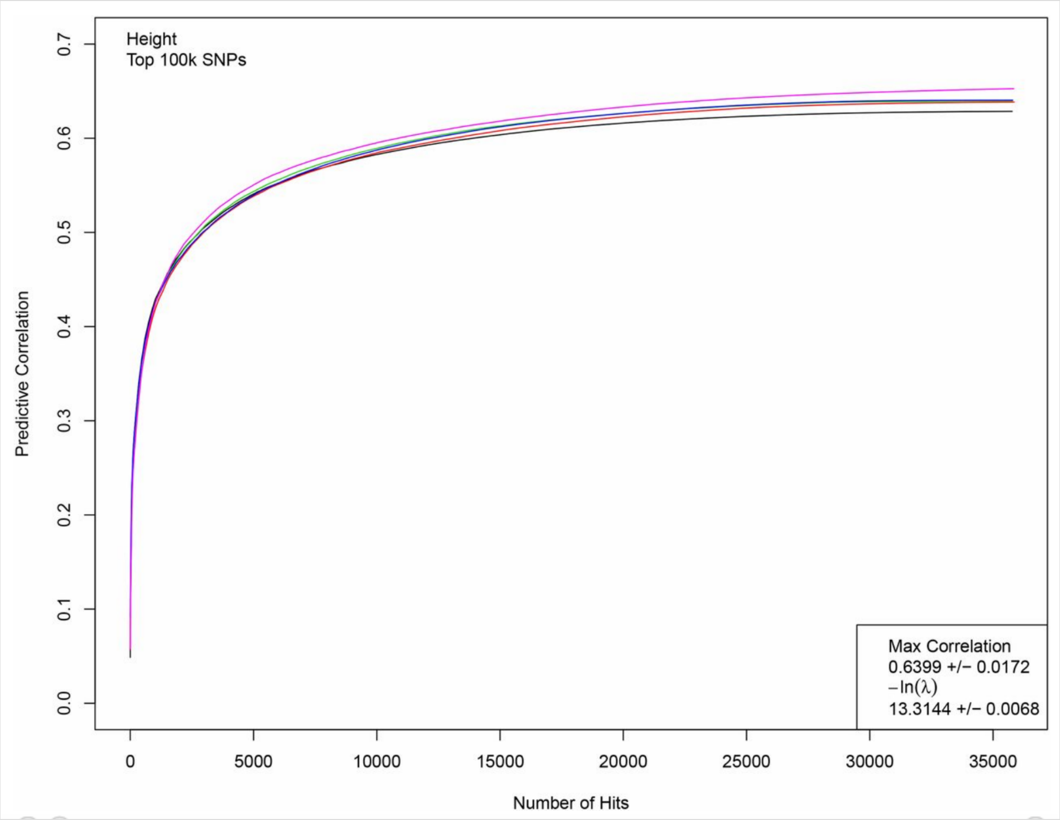
Accurate Genomic Prediction of Human Height

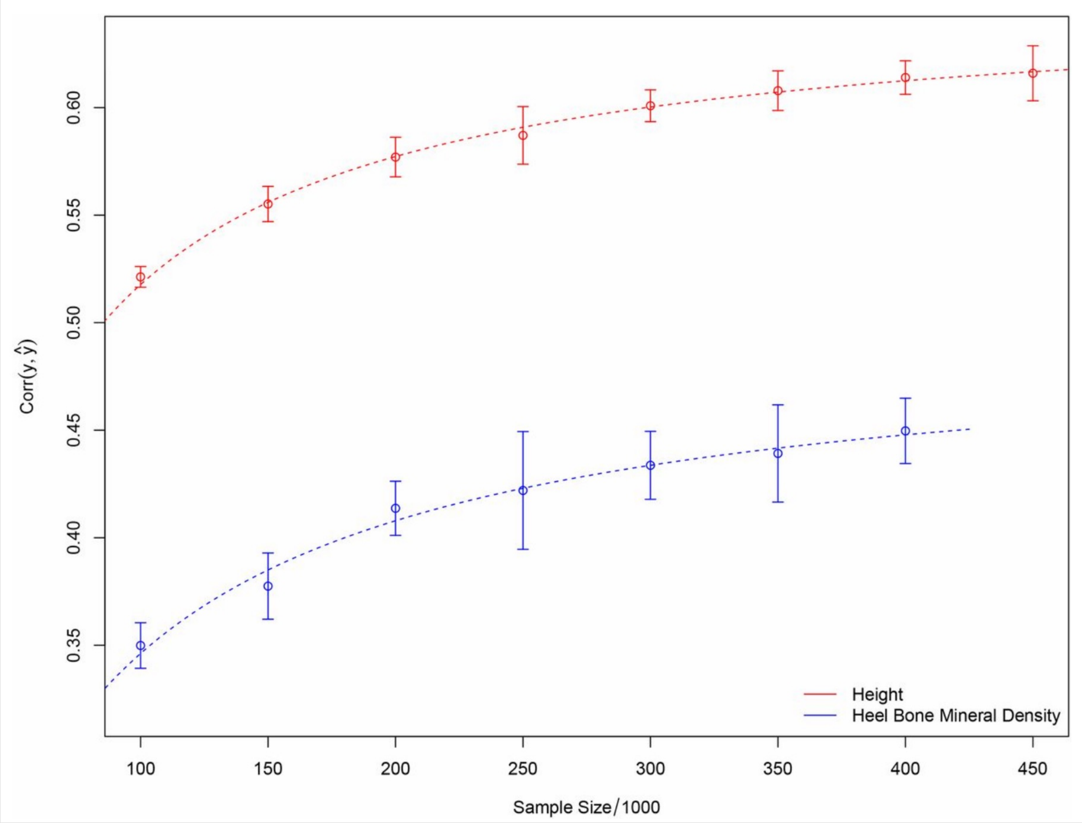
Louis Lello, Steven G. Avery, Laurent Tellier, Ana I. Vazquez,
Gustavo de los Campos and Stephen D. H. Hsu

GENETICS *October 1, 2018 vol. 210 no. 2 477-497; https://doi.org/10.1534/genetics.118.301267*

- Start with 650,000 genetic variants and 420,000 individuals with height measurements
- Use LASSO method for building the predictive model
- A first screening based on standard univariate regression on the training set to reduce the set of candidate predictors from 645,589 to the top p = 50k and 100k by statistical significance
- Age and sex were regressed out from the outcome variable (=height) and predictors and outcome were standardized

PGS DEVELOPMENT

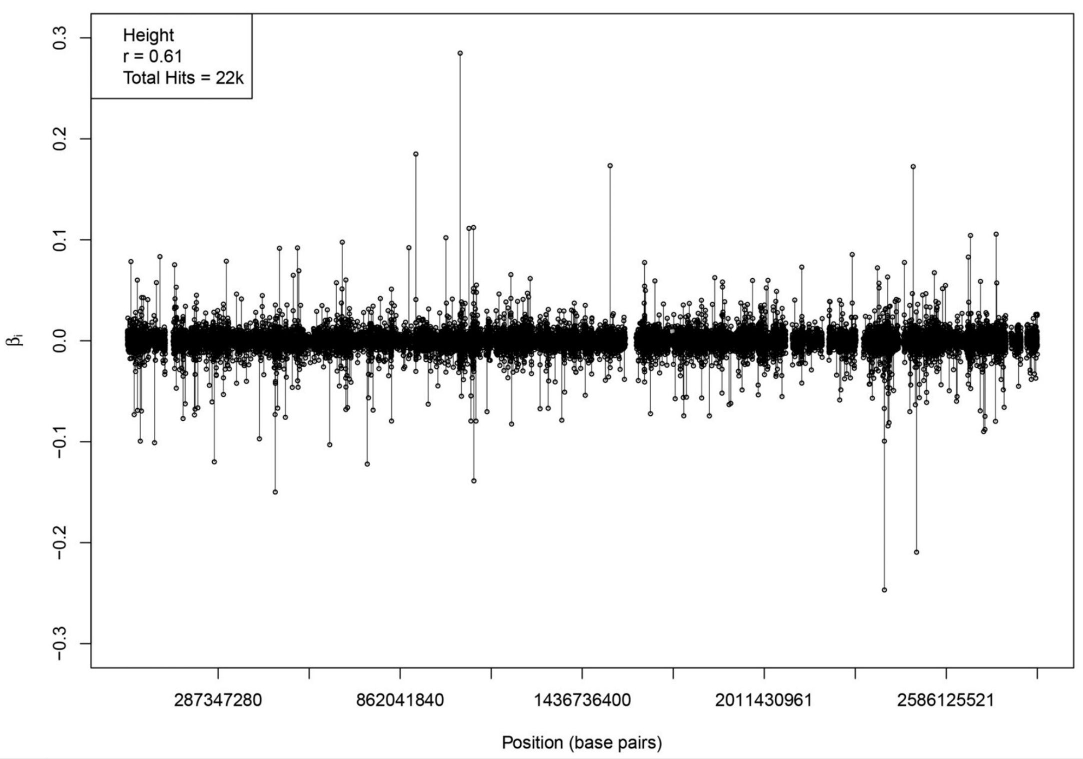


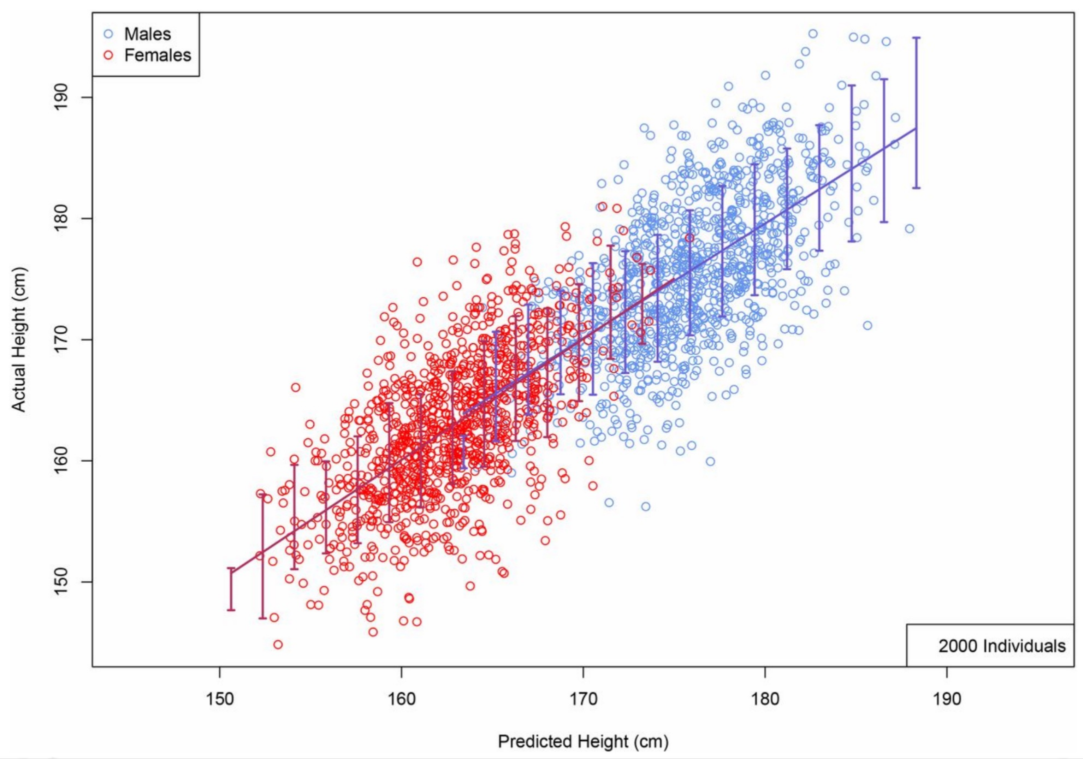


How many non-zero coefficients in the model?

How large GWAS was used?

FINAL MODEL



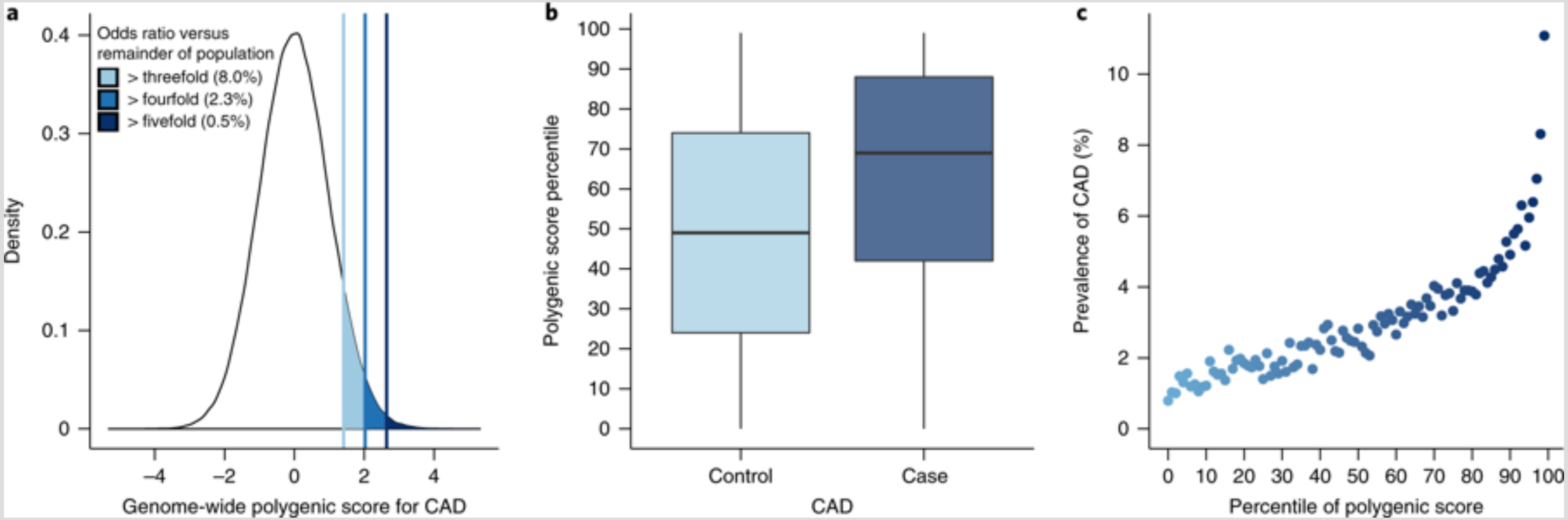


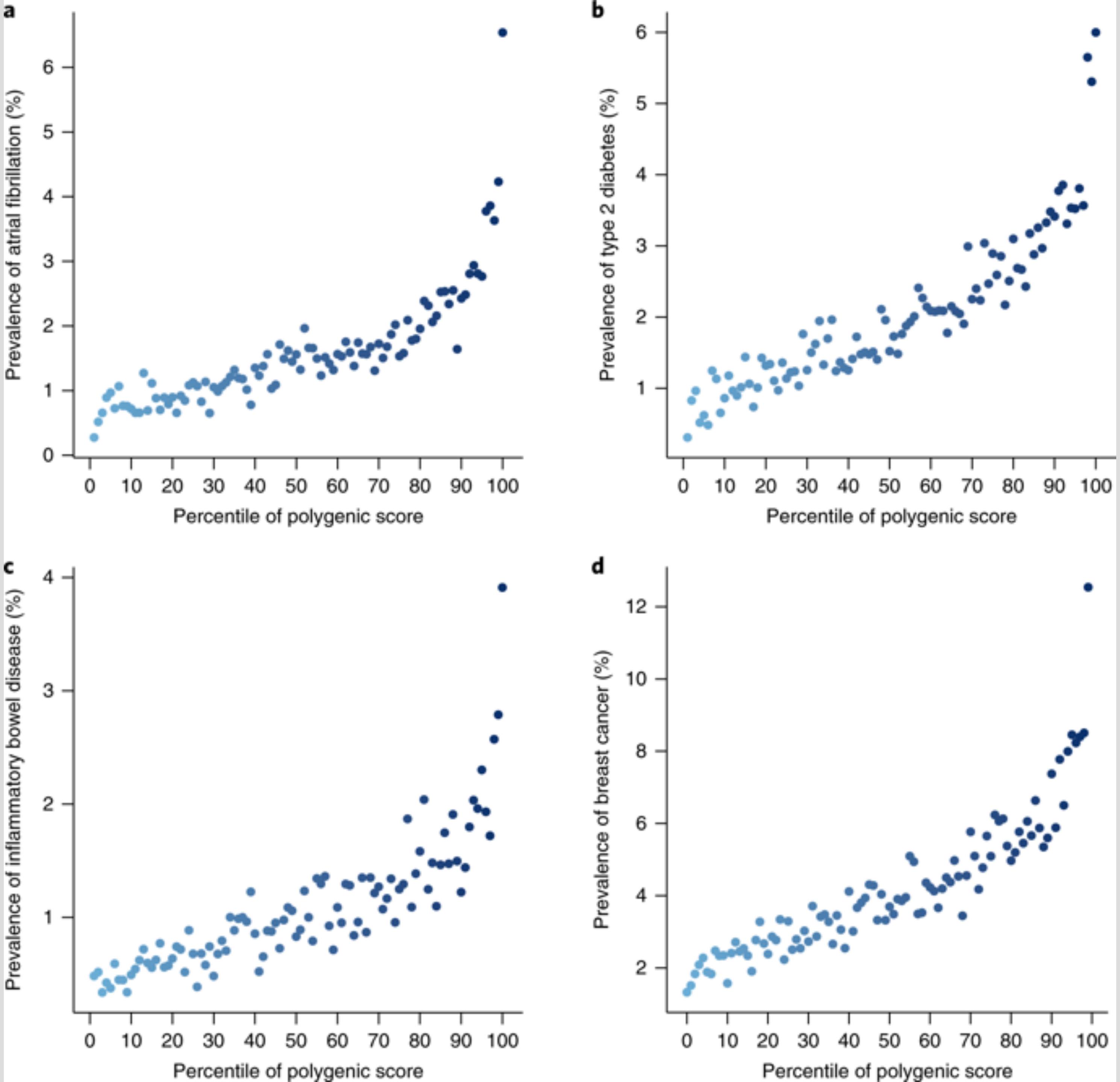
Uses 22,000 non-zero coefficients for SNPs across genome

Achieves r = 0.58, i.e., R^2 = 0.34 in UKB test data set and r = 0.54, i.e., R^2 = 0.29 in ARIC data that are independent of UKB.

KHERA ET AL. 2018 NAT GEN

- Allele effects from CARDIoGramplusC4D GWAS (n=60,000 cases/ 120,000 controls)
- Target individuals from the UK Biobank
- Identifies 8% of population with 3-fold risk compared to rest
  - Severe hypercholesterolemia mutations have similar risk but are <0.5% in population





PGS AND PREVALENCE

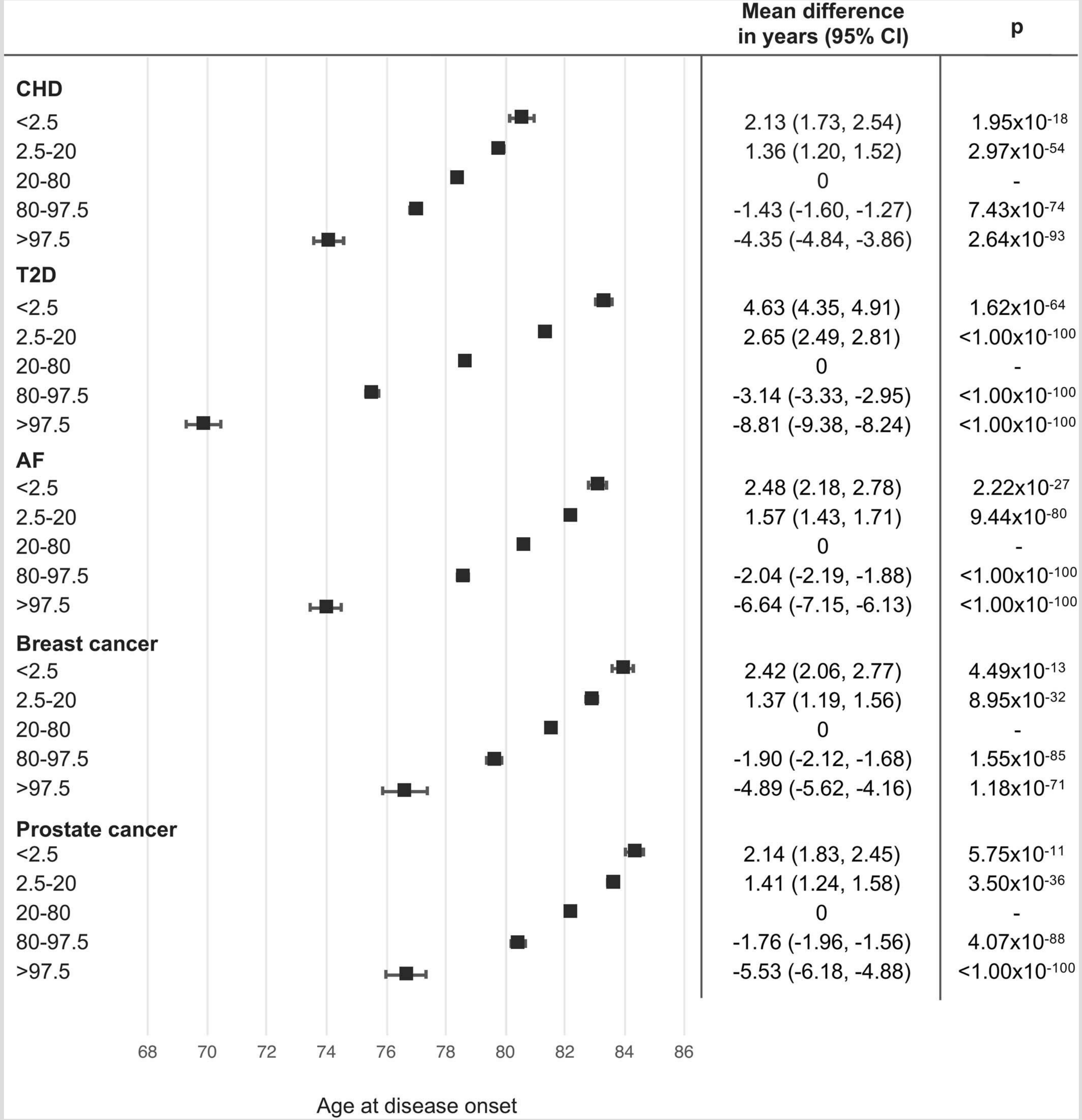
100 groups of the testing dataset were derived according to the percentile of the disease-specific PGS. **a**–**d**, Prevalence of disease displayed for the risk of
atrial fibrillation (**a**),
type 2 diabetes (**b**),
inflammatory bowel disease (**c**),
and breast cancer (**d**)
according to the PGS percentile.

Khera et al. 2018 Nat Gen

PRS AND DISEASE ONSET

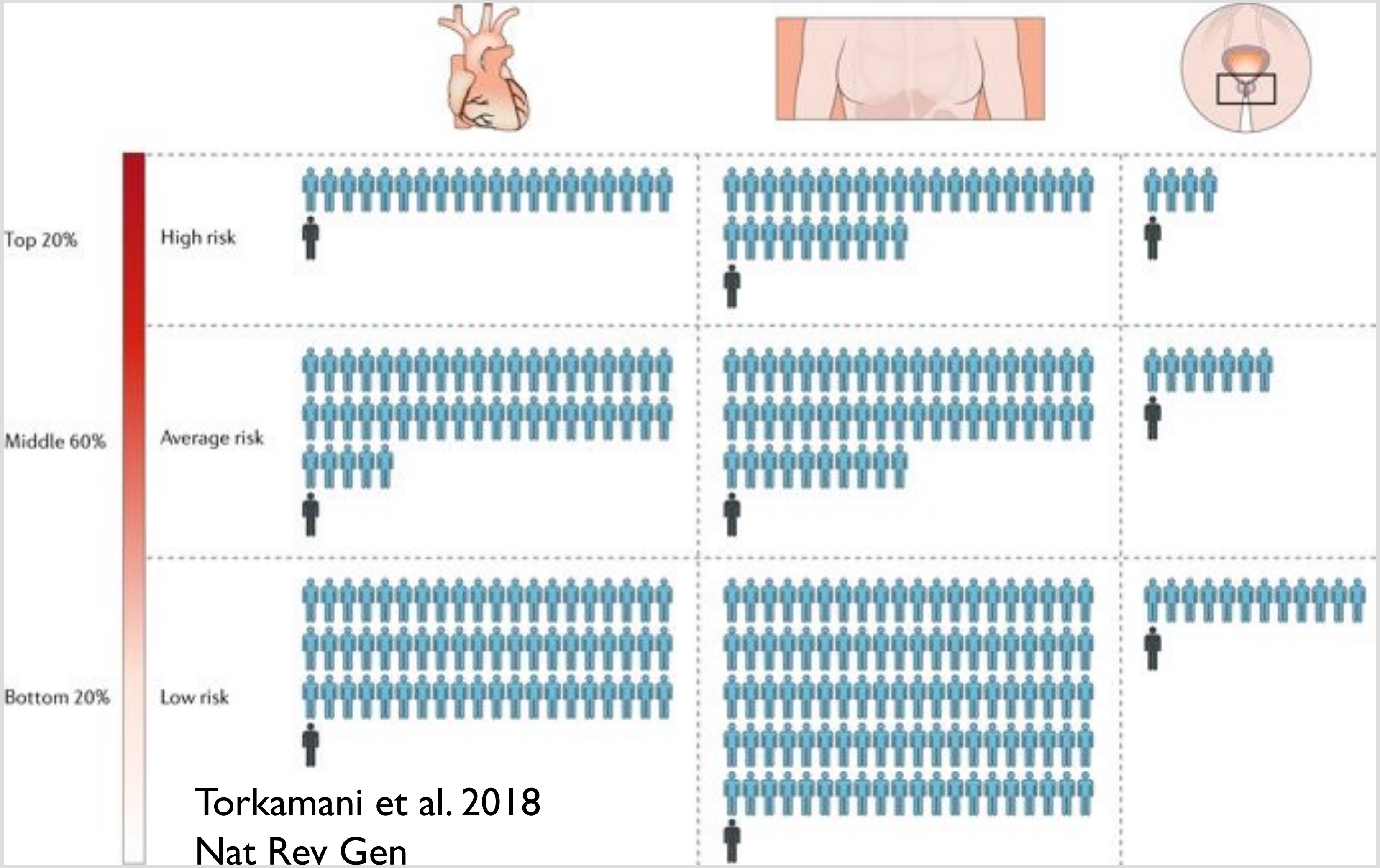
- FinnGen data (N=135,000)
- PRS could inform screening practices for cancers and other diseases where prevention is possible

Mars et al. 2020 Nat Medicine



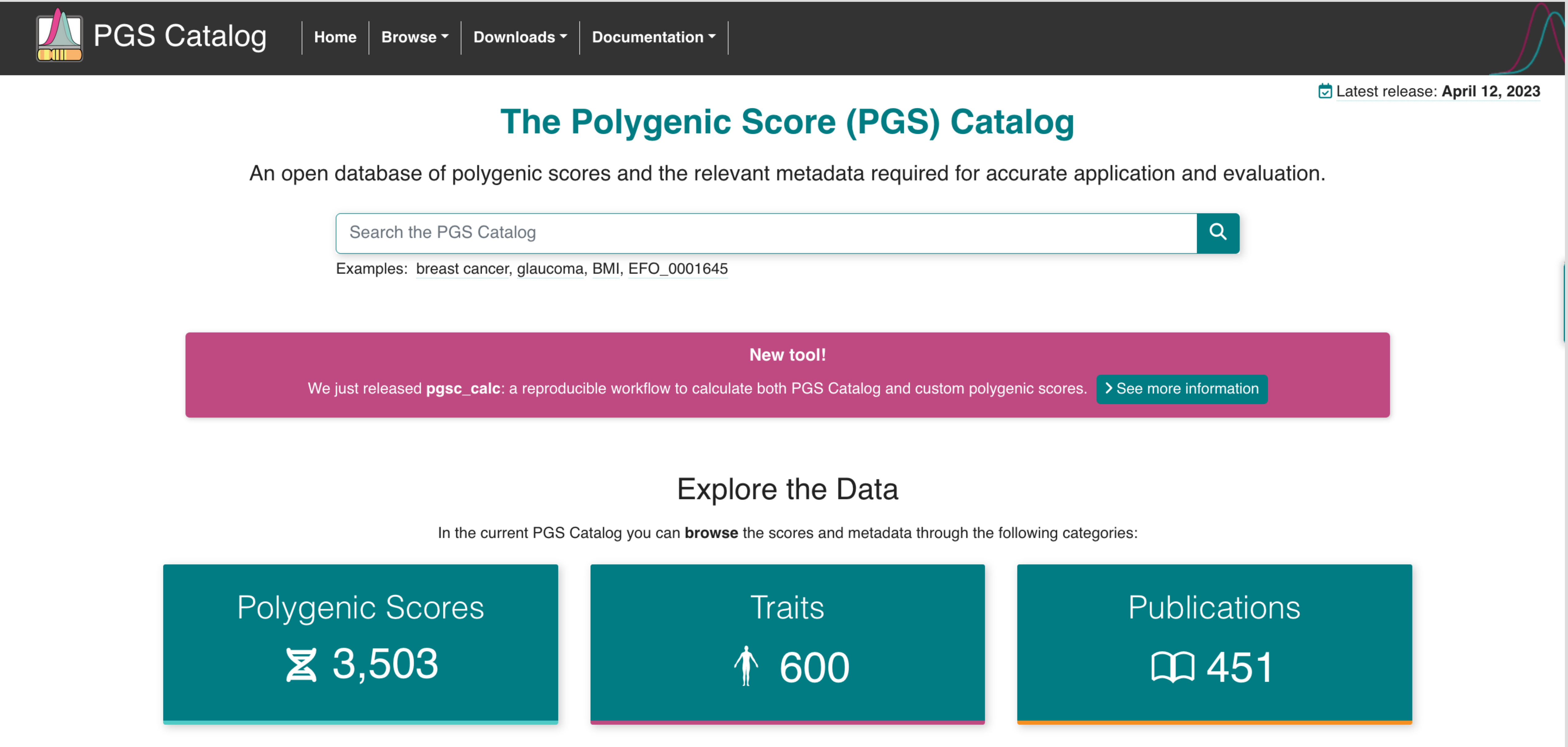
	Mean difference in years (95% CI)	p
CHD		
<2.5	2.13 (1.73, 2.54)	1.95x10^{-18}
2.5-20	1.36 (1.20, 1.52)	2.97x10^{-54}
20-80	0	-
80-97.5	-1.43 (-1.60, -1.27)	7.43x10^{-74}
>97.5	-4.35 (-4.84, -3.86)	2.64x10^{-93}
T2D		
<2.5	4.63 (4.35, 4.91)	1.62x10^{-64}
2.5-20	2.65 (2.49, 2.81)	<1.00x10^{-100}
20-80	0	-
80-97.5	-3.14 (-3.33, -2.95)	<1.00x10^{-100}
>97.5	-8.81 (-9.38, -8.24)	<1.00x10^{-100}
AF		
<2.5	2.48 (2.18, 2.78)	2.22x10^{-27}
2.5-20	1.57 (1.43, 1.71)	9.44x10^{-80}
20-80	0	-
80-97.5	-2.04 (-2.19, -1.88)	<1.00x10^{-100}
>97.5	-6.64 (-7.15, -6.13)	<1.00x10^{-100}
Breast cancer		
<2.5	2.42 (2.06, 2.77)	4.49x10^{-13}
2.5-20	1.37 (1.19, 1.56)	8.95x10^{-32}
20-80	0	-
80-97.5	-1.90 (-2.12, -1.68)	1.55x10^{-85}
>97.5	-4.89 (-5.62, -4.16)	1.18x10^{-71}
Prostate cancer		
<2.5	2.14 (1.83, 2.45)	5.75x10^{-11}
2.5-20	1.41 (1.24, 1.58)	3.50x10^{-36}
20-80	0	-
80-97.5	-1.76 (-1.96, -1.56)	4.07x10^{-88}
>97.5	-5.53 (-6.18, -4.88)	<1.00x10^{-100}

UTILITY OF PRS IN CLINICAL DECISIONS



Torkamani et al. 2018
Nat Rev Gen

The number of individuals treated or screened relative to the number of individuals receiving a benefit from the intervention is broken down by polygenic risk score (PRS) tier (top 20%,
from the 20% to the 80%
and bottom 20% of genetic risk).
Coronary artery disease
(left — number needed to treat with statins to prevent a heart attack
Breast cancer (middle — number of women screened to detect incident breast cancer)
Prostate cancer (right — positive predictive value of prostate-specific antigen (PSA) testing). Blue are healthy, black are unhealthy individuals.



Latest release: **April 12, 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.

Explore the Data

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

Polygenic Scores	Traits	Publications
3,503	600	451

Provides the SNP weights of thousands of published PGSes in a standardized format