# GWAS 9

Matti Pirinen
University of Helsinki
2.12.2020

# 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 "precisions" 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 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 each estimator

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

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

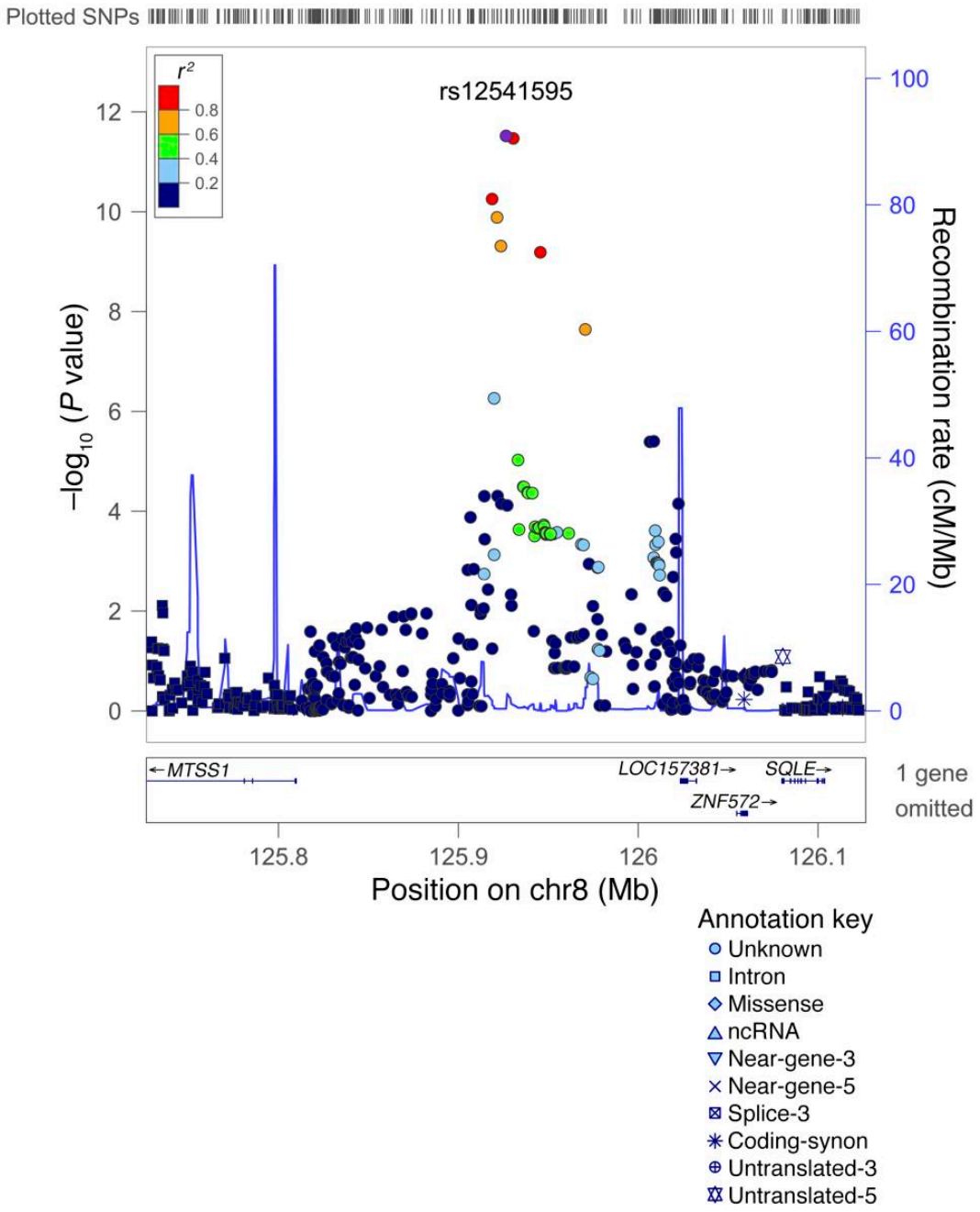
studies 1,…, K

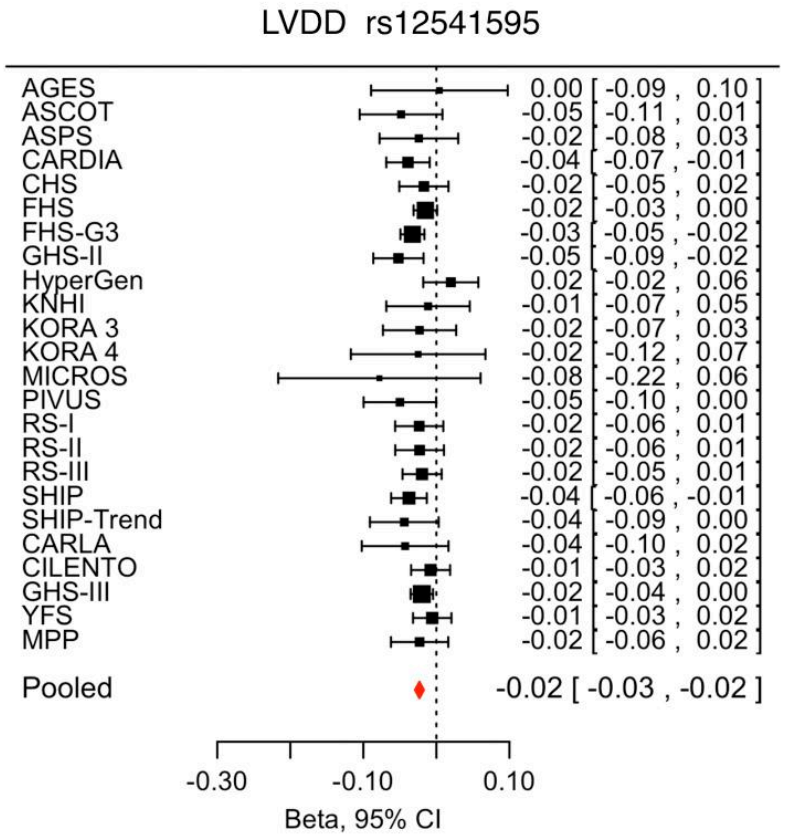
$$\mathrm{SE}_{l,F} = (w_{1l} + \ldots + w_{Kl})^{-\frac{1}{2}},$$

where the weight

$$w_{kl} = \frac{1}{\mathrm{SE}_{kl}^2}$$ 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)



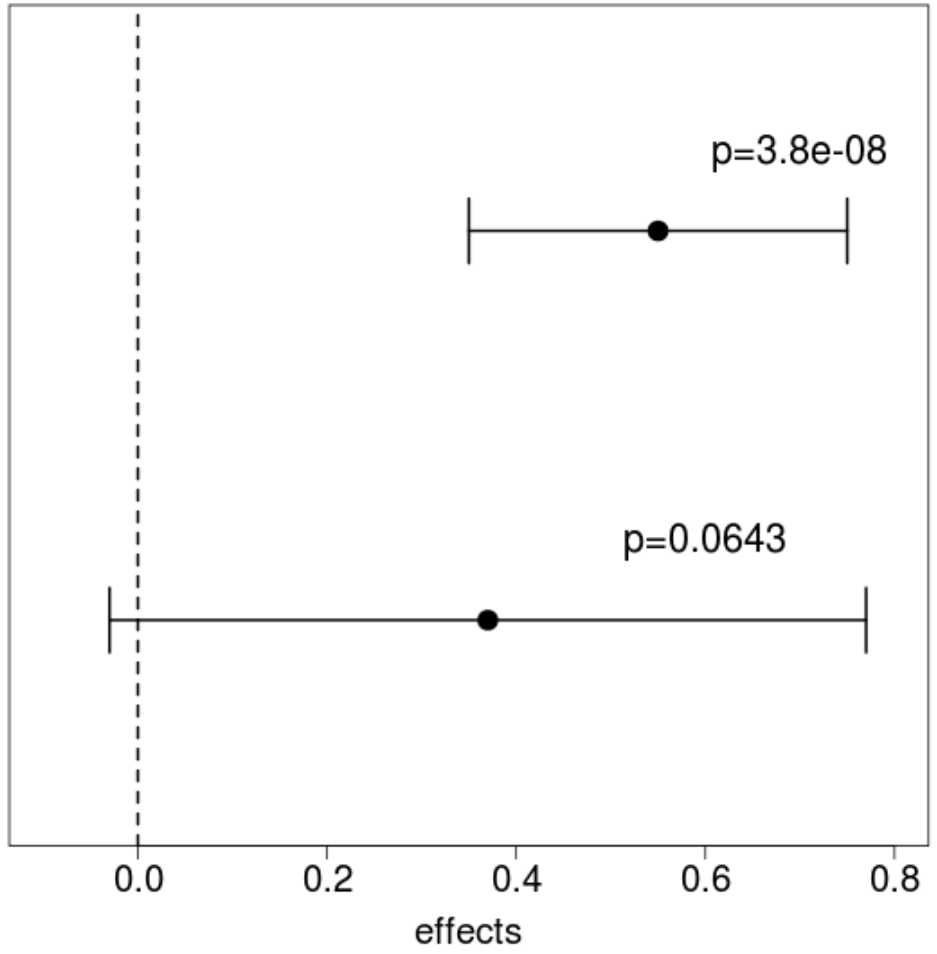




Forest plot for the meta-analysis of the association between rs12541595 and LVDD, with the corresponding regional plot including functional annotation.

LVDD = left ventricular diastolic internal dimension

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

# IS FIXED-EFFECTS ASSUMPTION REASONABLE?



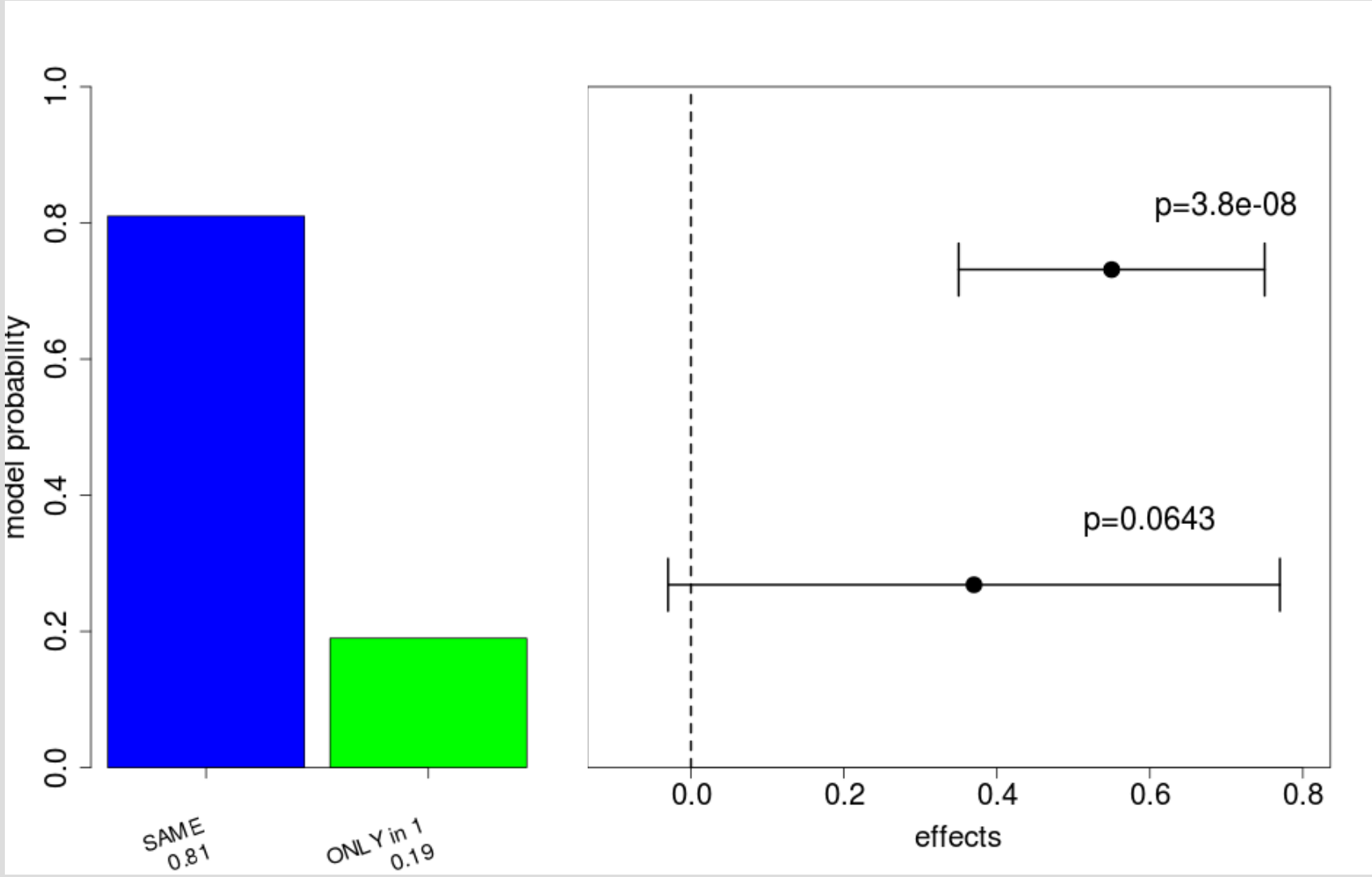

Suppose we have two estimates.

(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 same!)

# IS FIXED-EFFECTS ASSUMPTION REASONABLE?




Suppose we have two estimates.

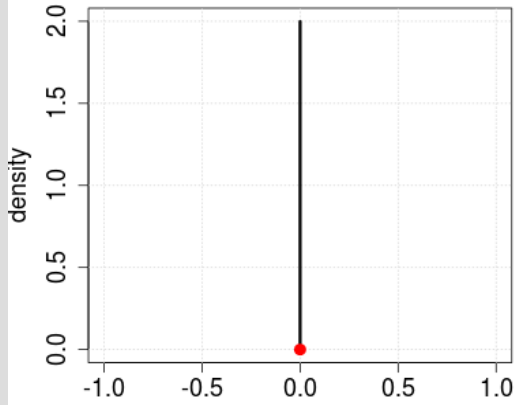
We write each of the possible explanations of the data in terms of a statistical model and compare how well each of them describes the data by using the Bayesian model comparison framework.

# BUILDING MODELS FOR 2 EFFECTS

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

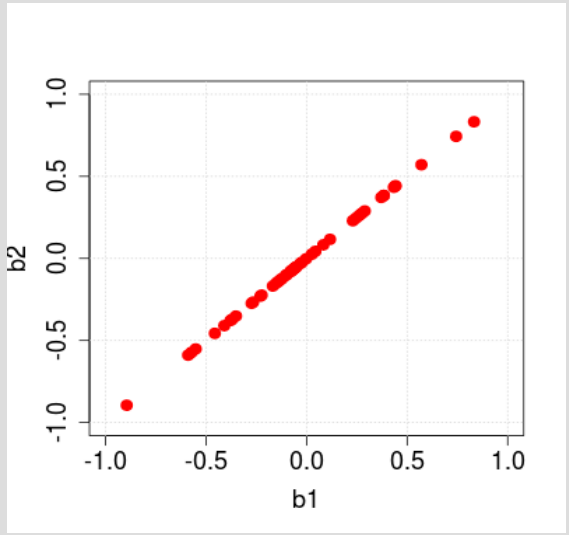
Here we have two SNPs and build joint models for them

N: (null model)

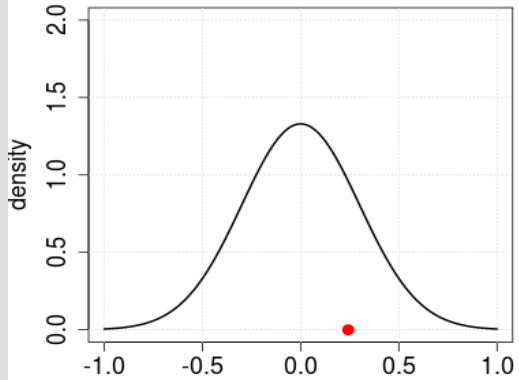



SAME EFFECT:
- Pick value x ~ E
- Set $b_1 = b_2 = x$

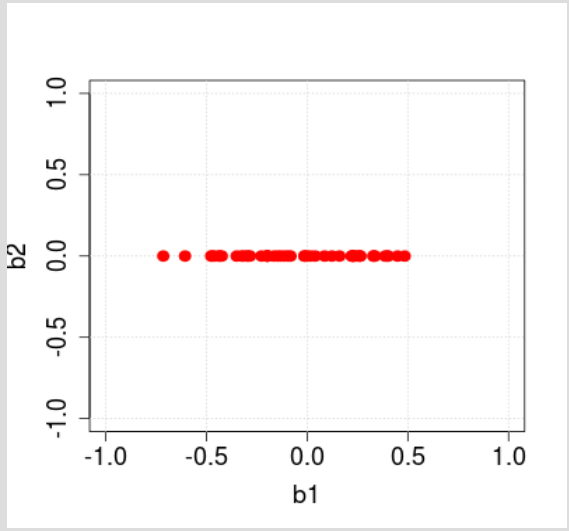
Example data from each model



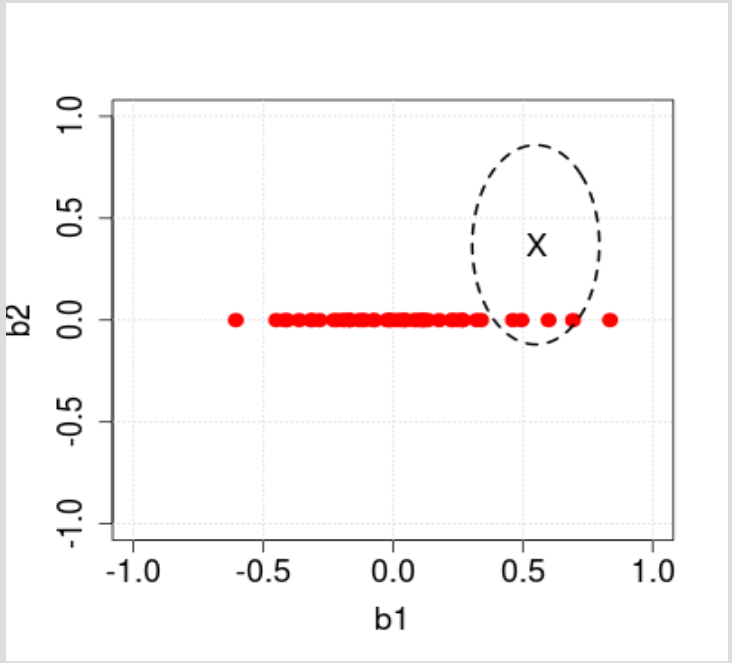

E: (effect model)

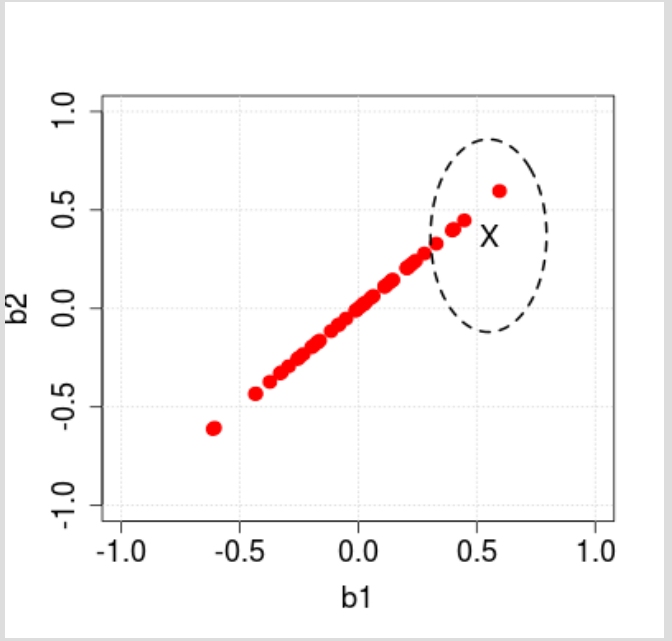



EFFECT IN ONLY 1:
- Pick $b_1$ ~ E
- Pick $b_2$ ~ N
(i.e. $b_2 = 0$)



# HOW WELL THE MODELS EXPLAIN THE DATA?

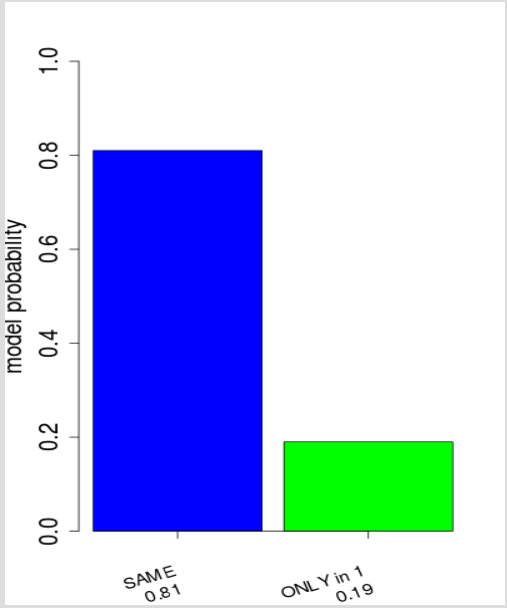


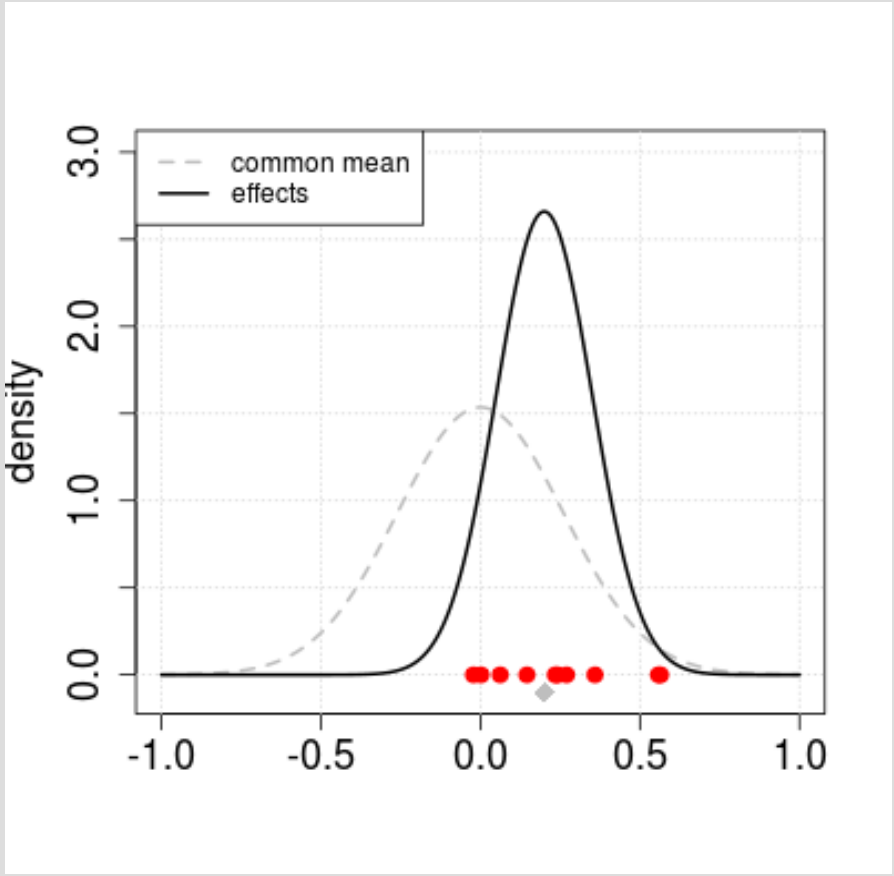

In our example the estimates were similar but SE of $b_2$ was much larger.

Same effect model is a better explanation here.

$$\frac{P(\text{obs data} \mid \text{Model 1})}{P(\text{obs data} \mid \text{Model 2})}$$



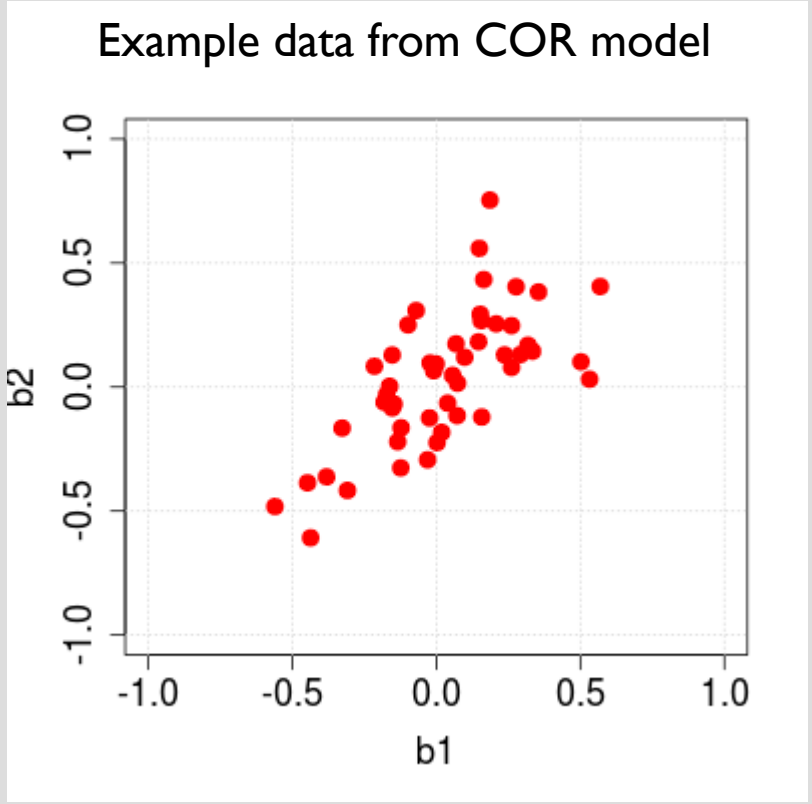
# CORRELATED EFFECTS MODEL



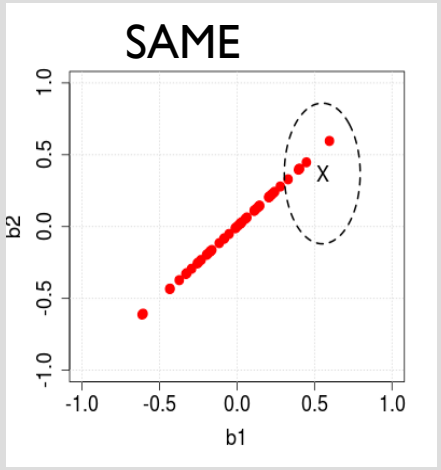

Model ”COR” a.k.a ”REL”

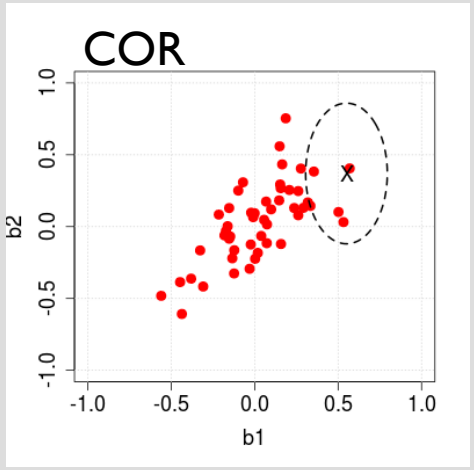
- Pick $\mu \sim N(0, s^2)$
- Given μ, pick each $b_i \sim N(\mu, t^2)$

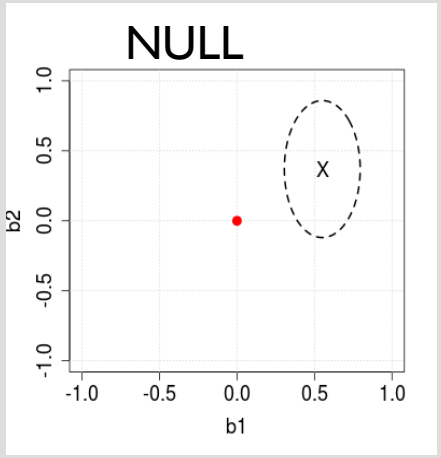
Example data from COR model

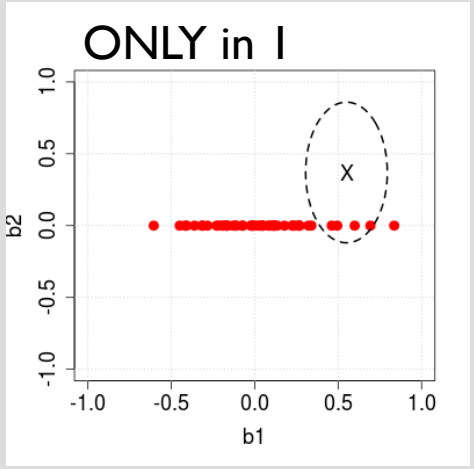


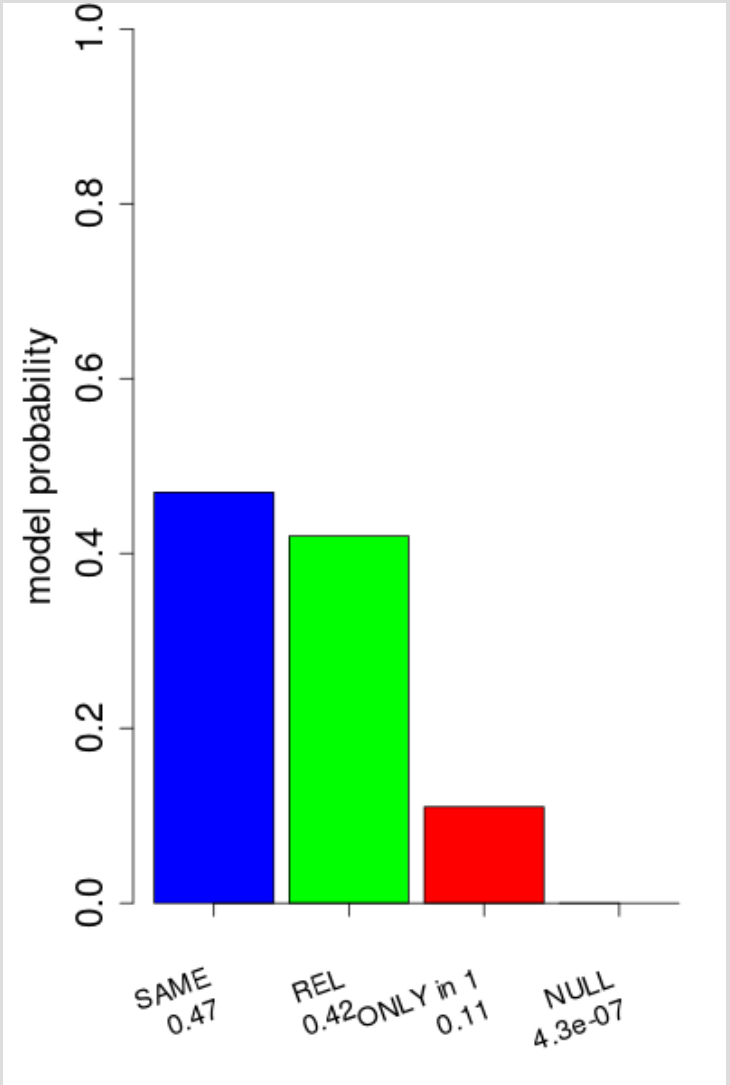
# MODEL COMPARISON



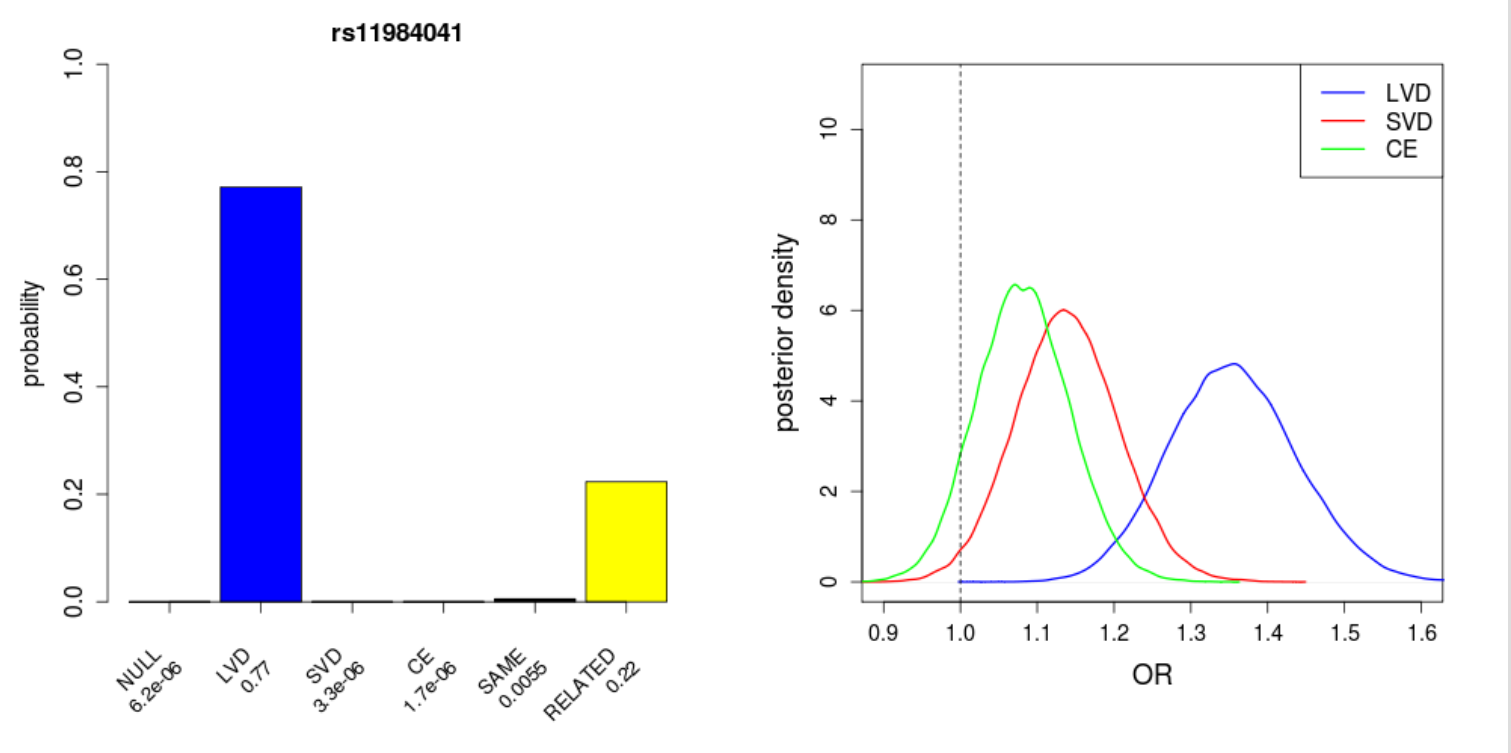



# ISCHEMIC STROKE AND *HDAC9* SNP

| 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 |

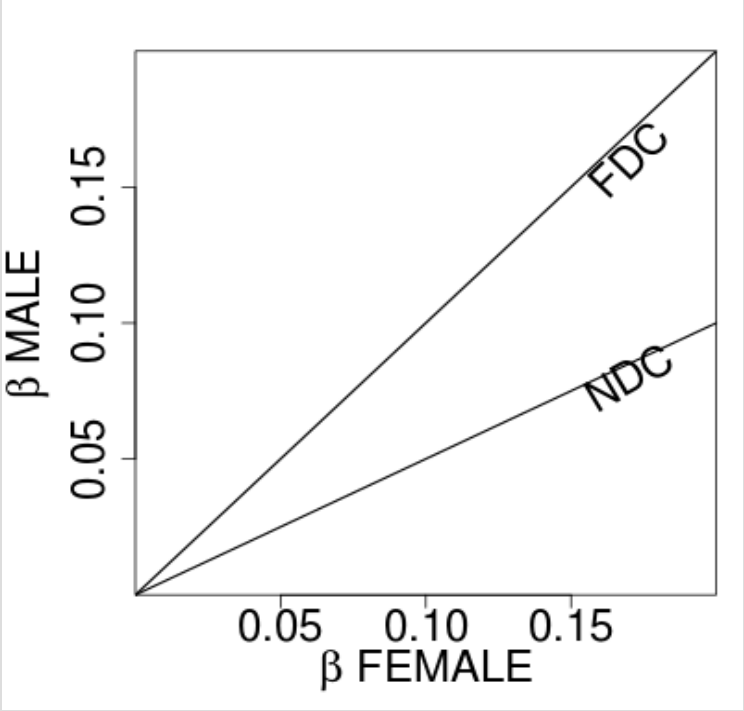



LVD = large vessel disease
SVD = small vessel disease
CE = cardioembolic stroke

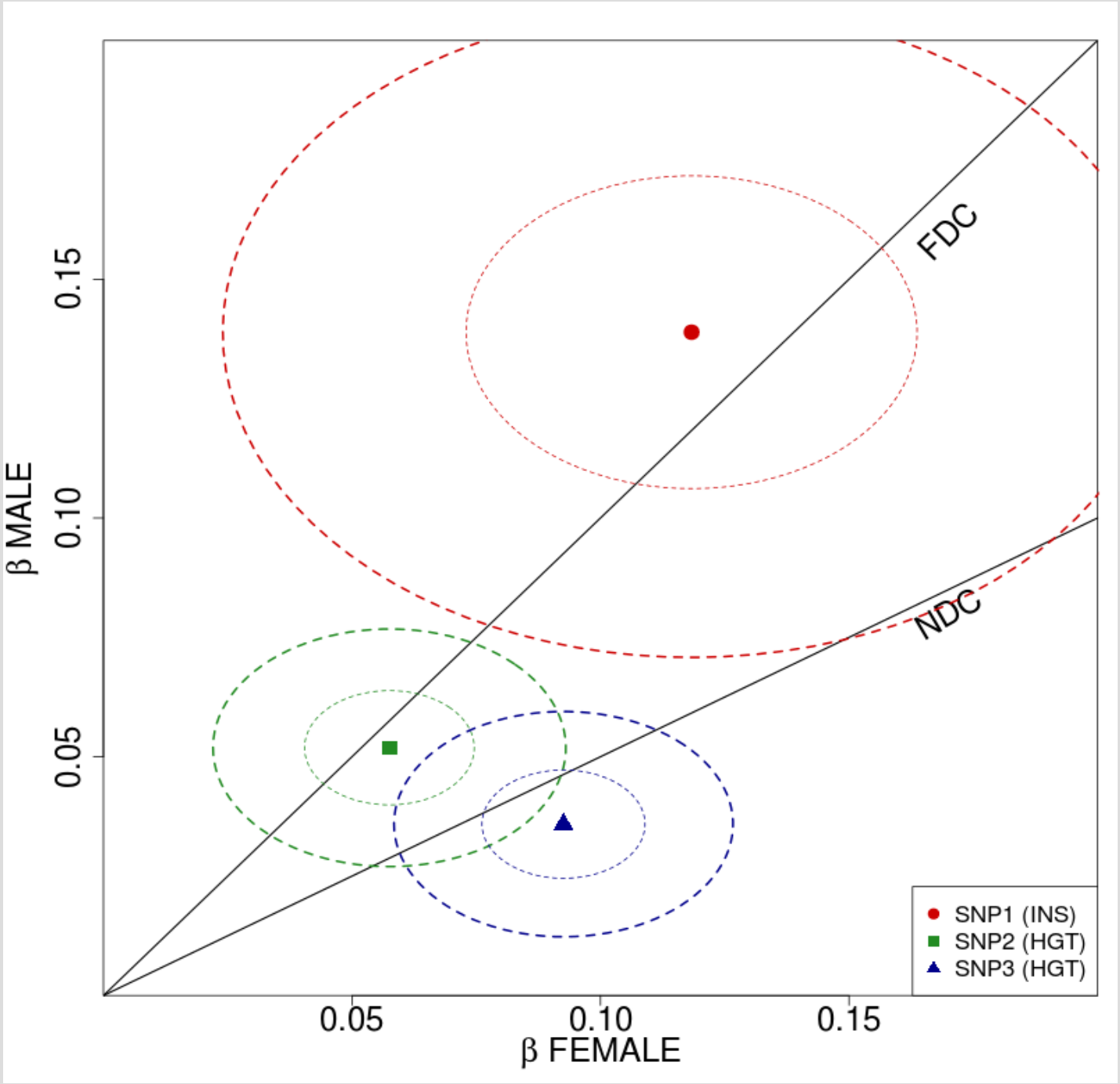
Bellenguez et al. 2012 Nat Gen

# CHR X DOSAGE COMPENSATION

- One of the female's X chrs in each cell is inactivated
  - Balances difference in chr X number between the sexes (dosage compensation)
  - Inactivation is not complete, 15%-25% of genes escape from it
- 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

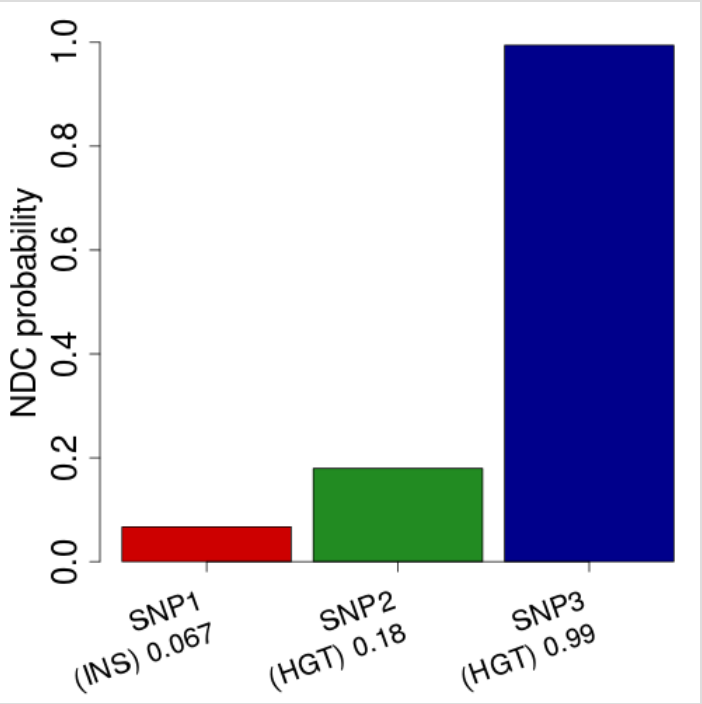


Tukiainen et al. 2014 PLoS Gen

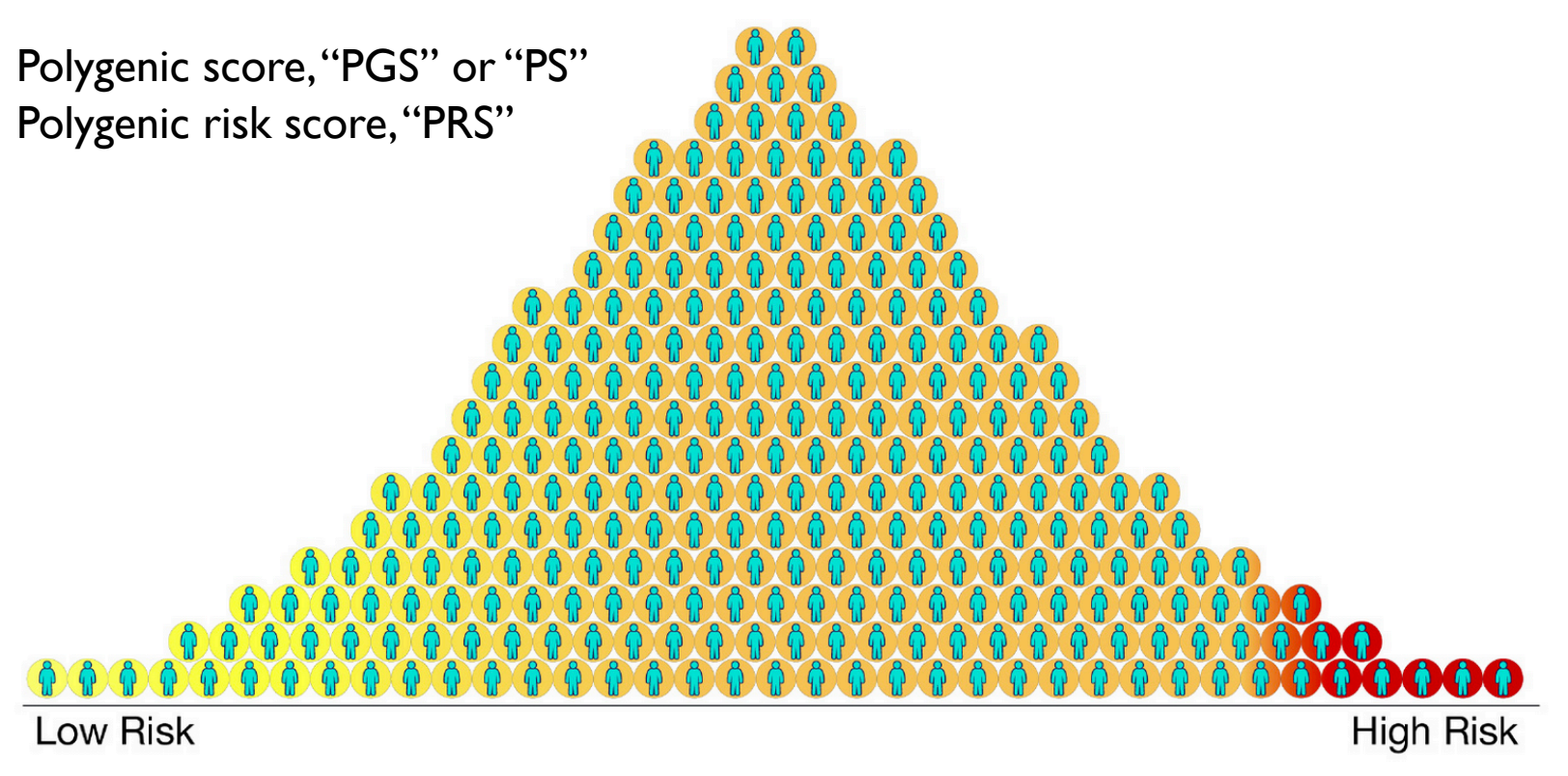



We have 3 chr X associations with Insulin levels or with height.

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



# POLYGENIC SCORES




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

Figure: NIH

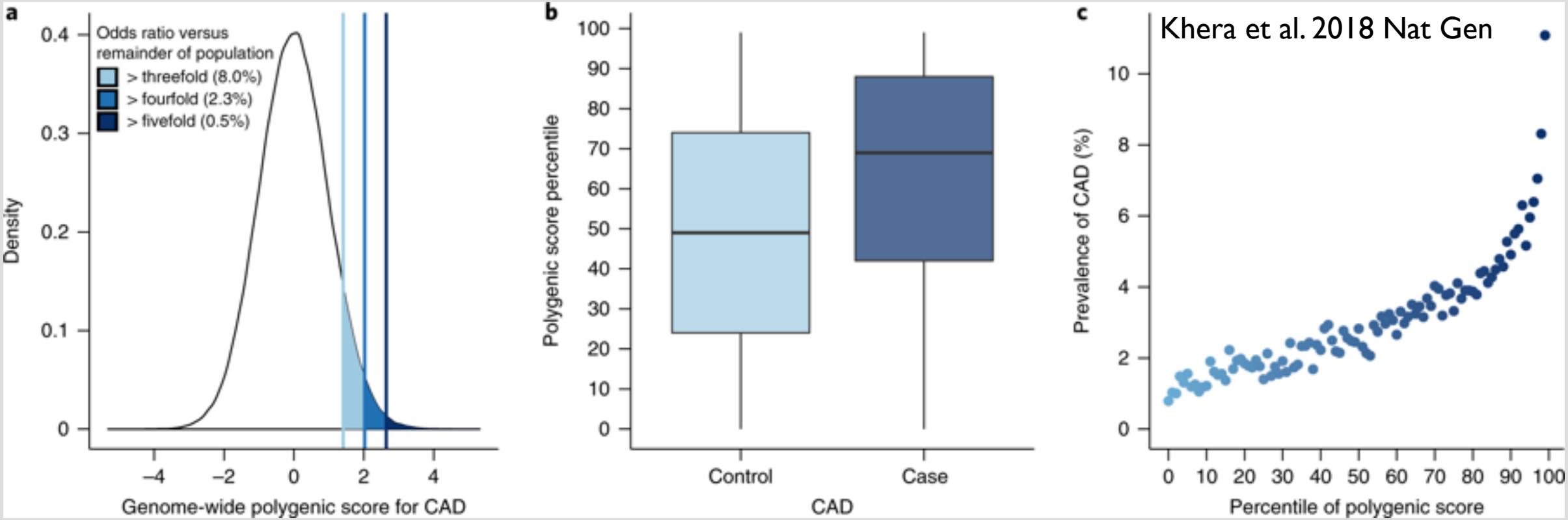
# (FUTURE) USES OF GENETIC SCORES


Help in prevention

- lifestyle change
- Screening program

How best to treat this person?

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

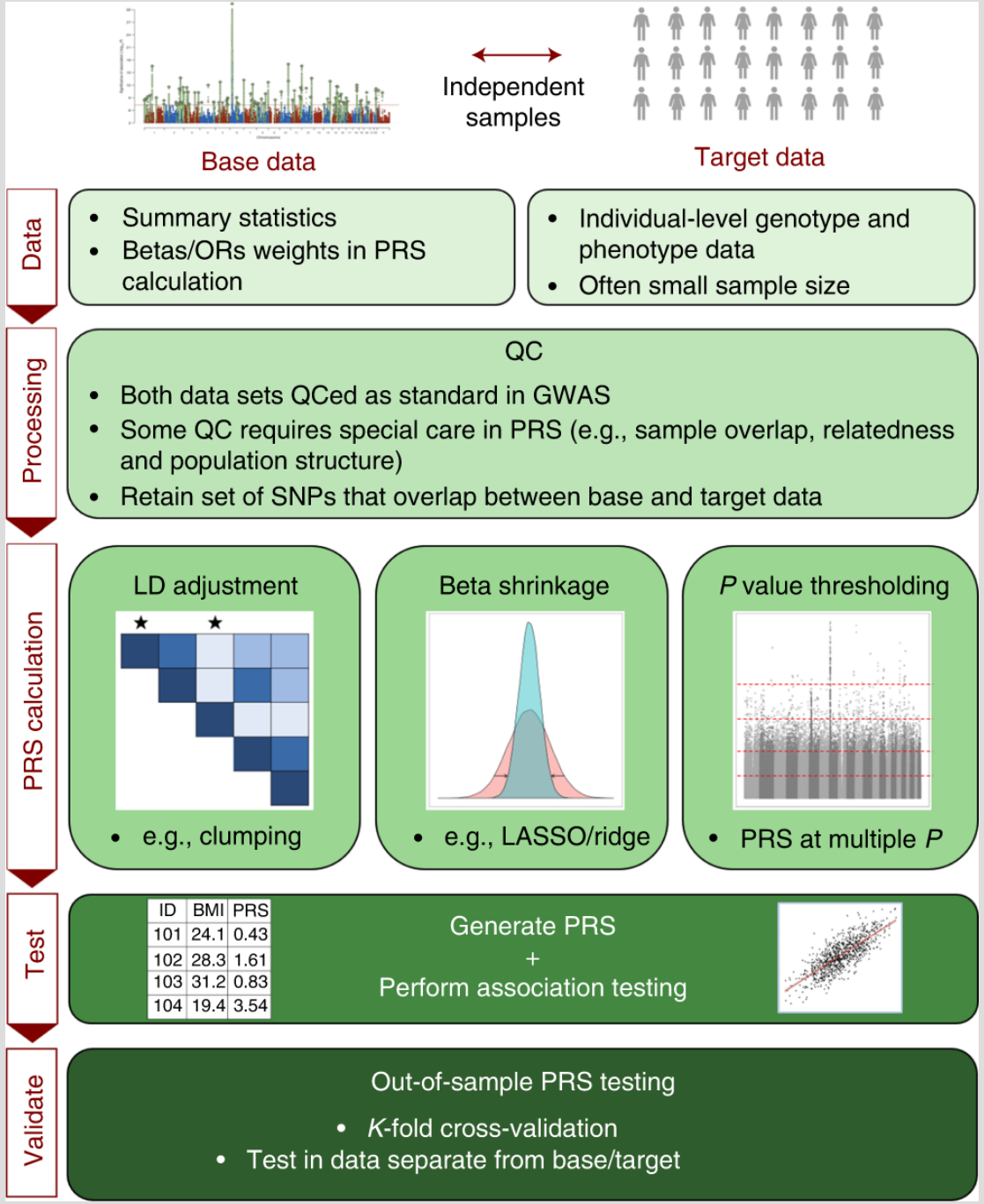



**a.** Distribution of $PGS_{CAD}$ in the UK Biobank testing dataset (*n* = 288,978). The *x* axis represents $PGS_{CAD}$, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation.
Shading reflects the proportion of the population with three-, four-, and fivefold increased 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 ( $\beta_k$) from GWAS
  - Not necessarily marginal effects
- Take target individual's genotypes ($g$) at the loci
- Compute *PRS* for individual *i* as sum

$$PRS_i = \sum_{k=1}^{\#Loci} g_{ik}\beta_k$$

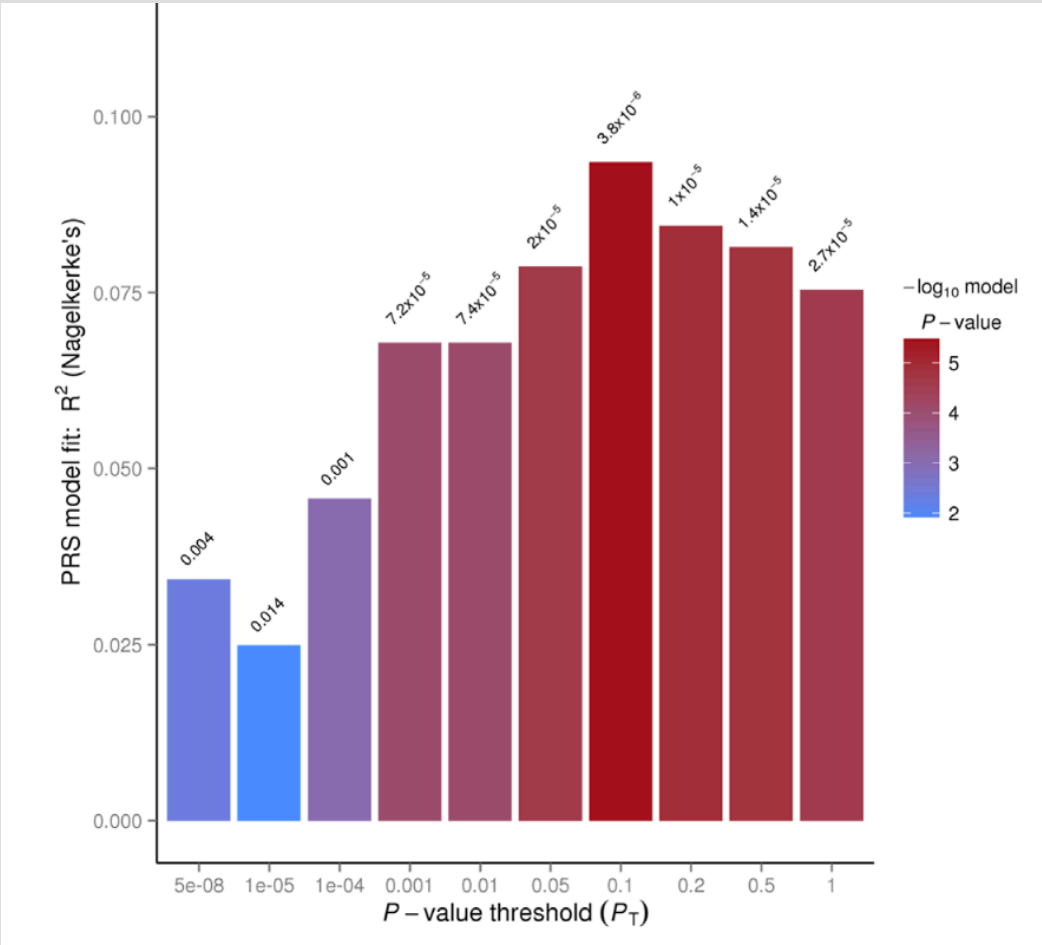



Choi et al. 2020 Nat Protocols

# 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 smaller GWAS P-value
  - This forms "clumps" of "significant" SNPs in LD with each other and only picks the most "significant" 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

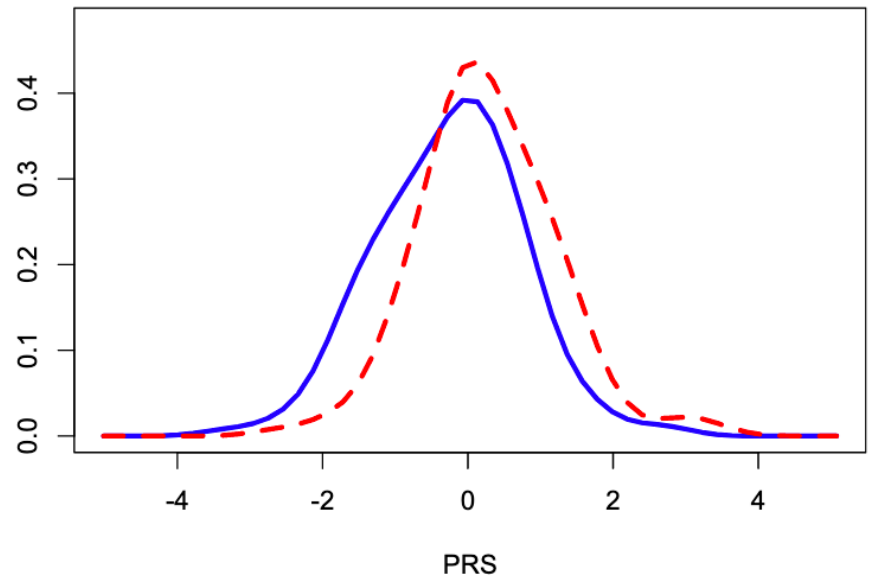
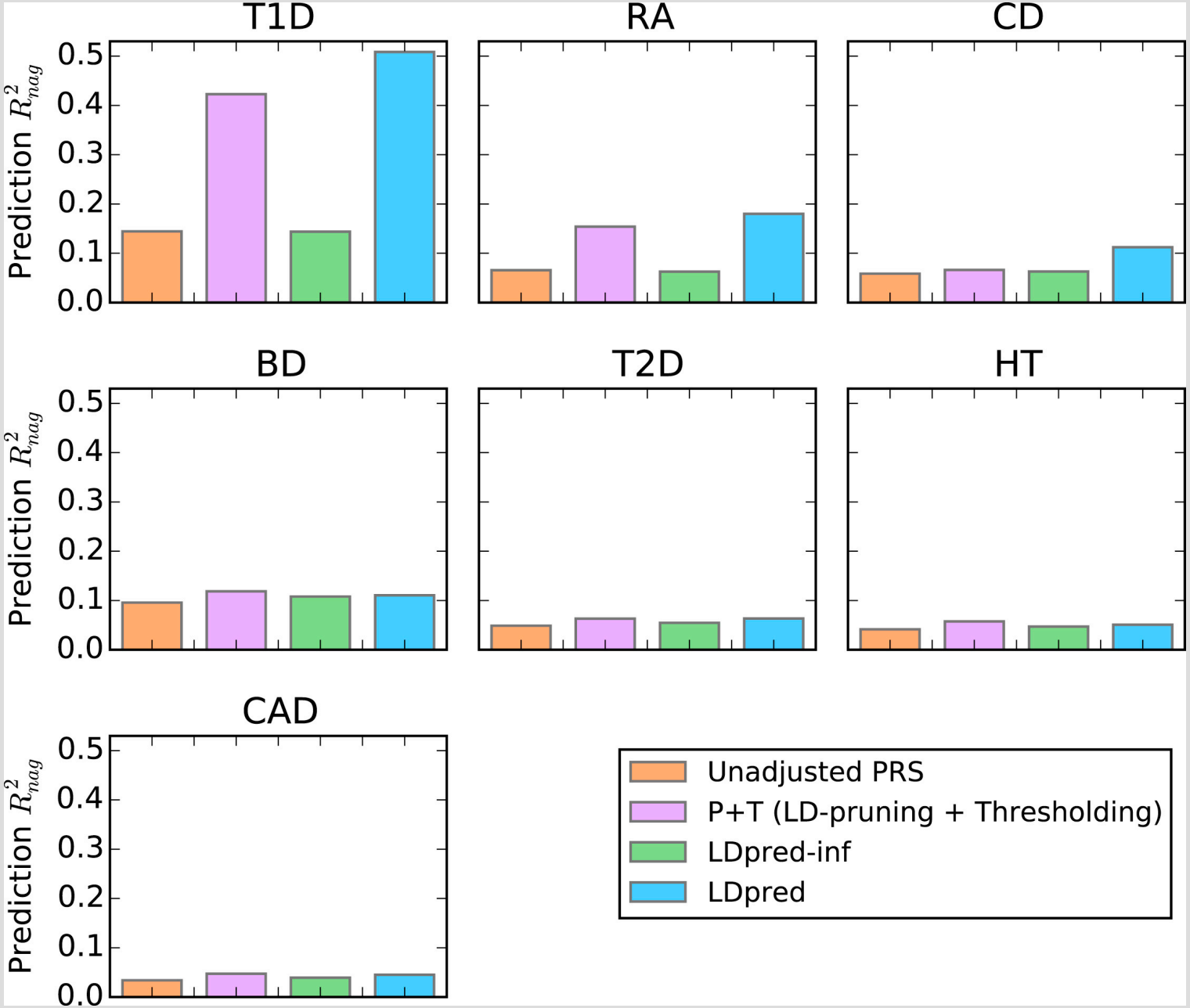



**Figure 2.** Density distribution of polygenic risk score (PRS) in European first-episode psychosis case and control subjects. PRS represents the standardized residuals of PRS after adjustment for the 10 principal components. Blue line indicates control subjects; red line, case subjects.

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 $\lambda_k \sim \begin{cases} N\left(0, \frac{h^2}{p\theta}\right), \text{with prob. } \theta \\ 0, \text{with prob. } 1-\theta \end{cases}$, where $h^2$ is heritability and $p$ is #SNPs
- Given marginal GWAS effects $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_k)$ and SEs, LDpred computes posterior expectation of the causal effects $E(\boldsymbol{\lambda} | \widehat{\boldsymbol{\beta}}, \boldsymbol{R}, h^2, \theta)$, 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 $\theta$ values are evaluated and best performing model is chosen
- These estimated causal effects are used as weights in PRS computation




T1D = Type 1 diabetes
RA = Rheumatoid arthritis
CD = Crohn's disease
BD = Bipolar disorder
T2D = Type 2 diabetes
HT = hypertension
CAD = Coronary artery disease

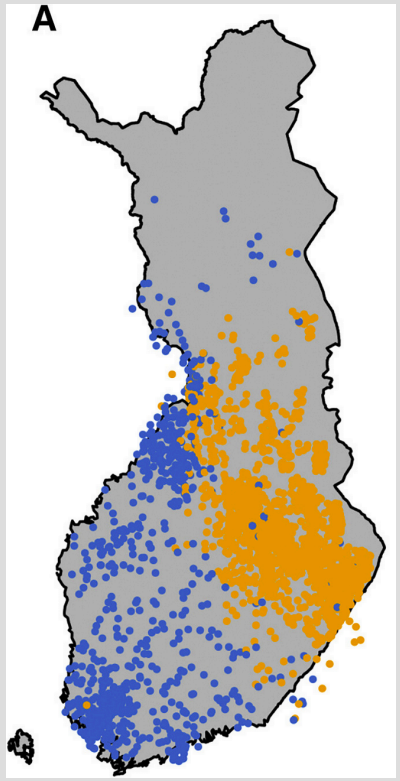
LDpred-inf assumes that every SNP has a non-zero effect
"infinitesimal" model = polygenic model

(Vilhjalmsson et al. AJHG 97:576-592)

# BIASES

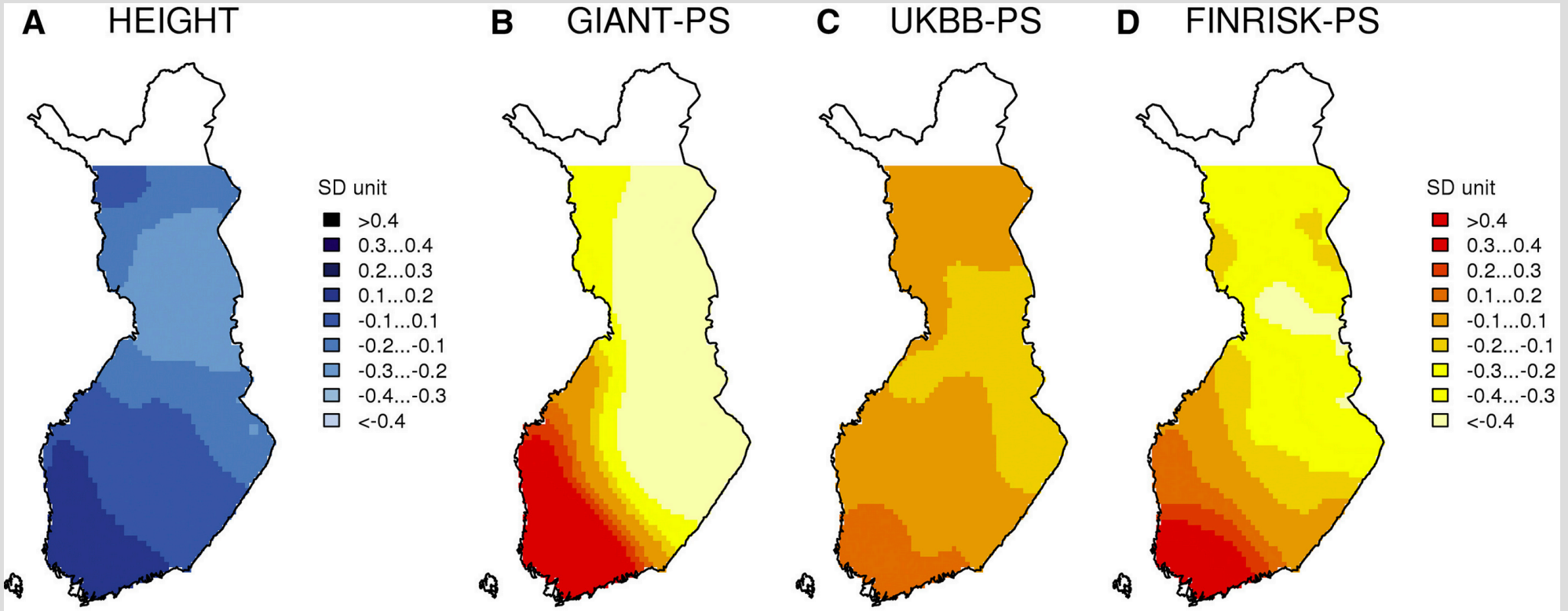
- If PGS is used to predict phenotypes of individuals who were included in the base GWAS, the prediction will be dramatically over optimistic
  - Make sure there is no overlap between GWAS and target sample
- Even if there is no overlap, relatedness and population structure can cause biases
- PGS based 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

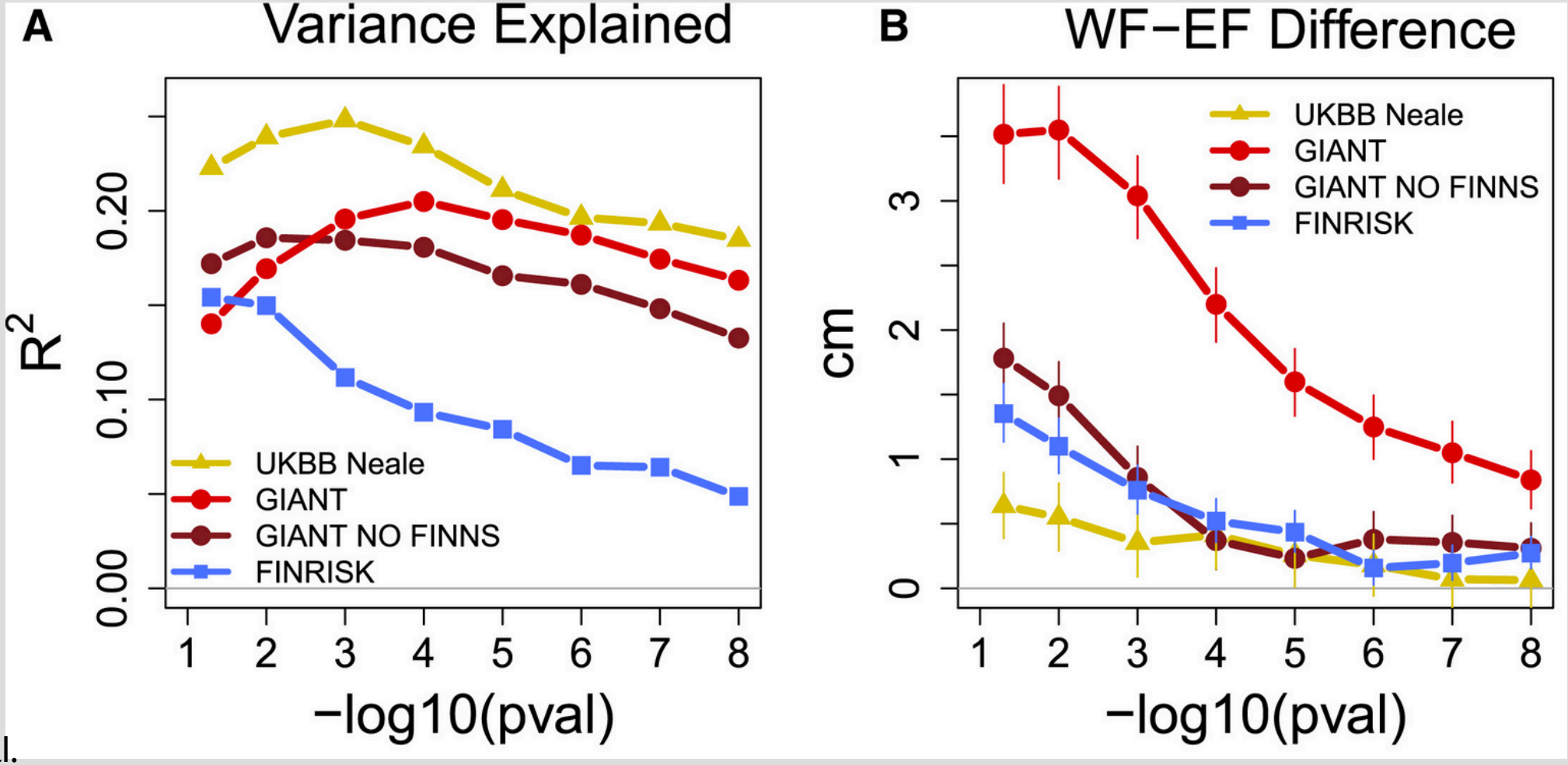
# COMPARING PREDICTIONS

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) |

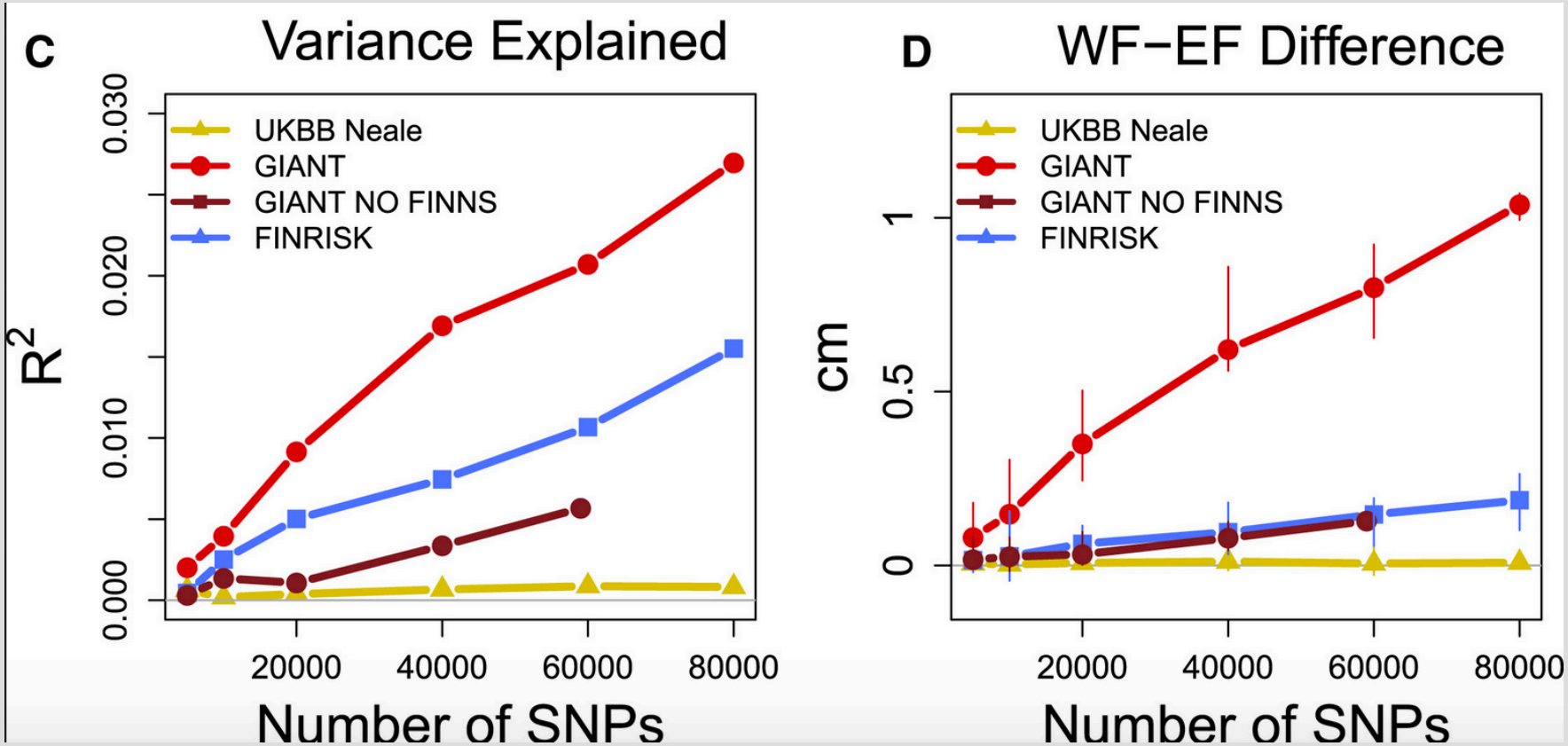
True East-West height difference is 1.6 cm, and these PGS should only predict < 1/3 of it.

# PERFORMANCE DEPENDS ON $P_{THR}$




Kerminen et al.
2019 AJHG

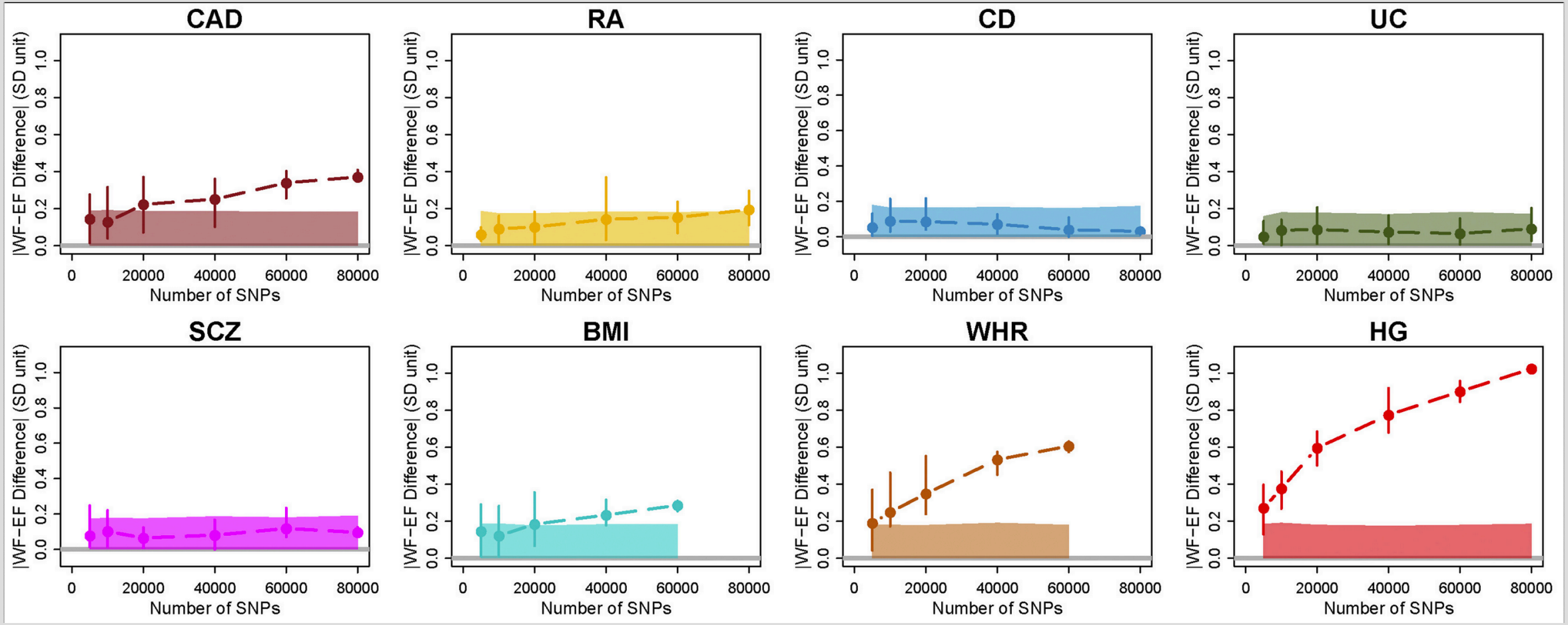
# "RANDOM" PGS




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.

Kerminen et al. 2019
AJHG

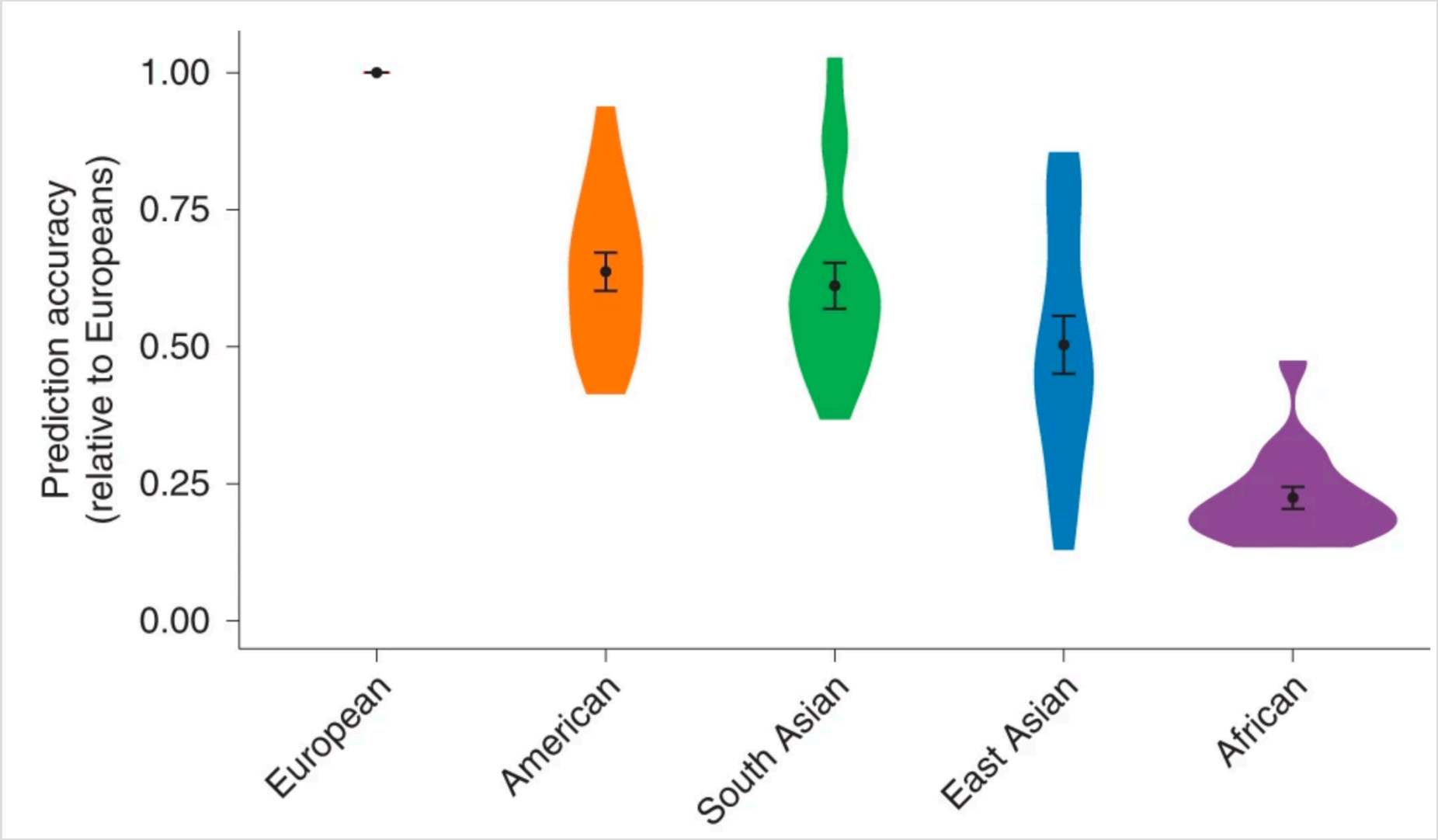
# RANDOM SCORES



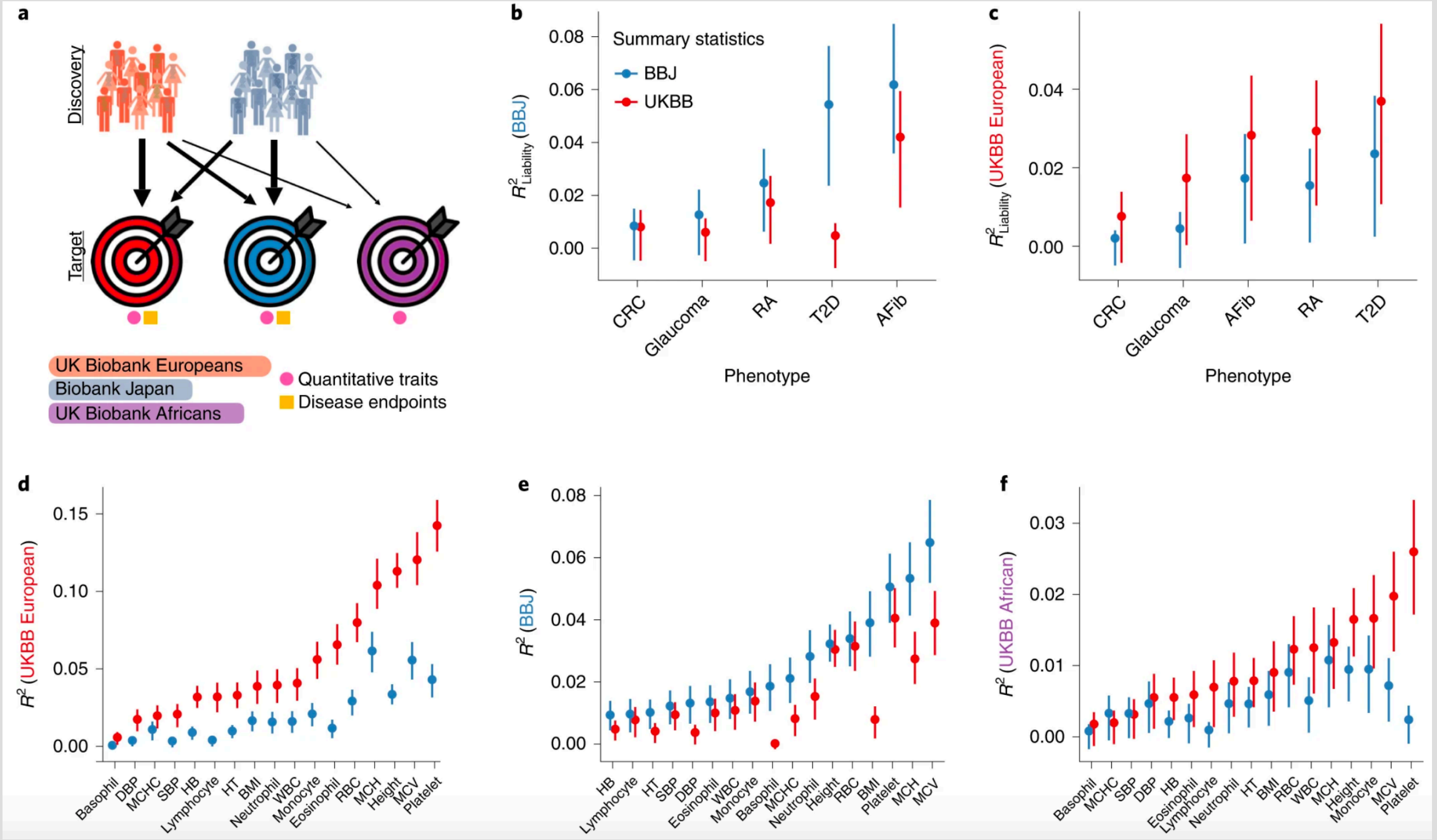

Kerminen et al. 2019 AJHG

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

# LACK OF TRANSFERRABILITY BTW POPULATIONS

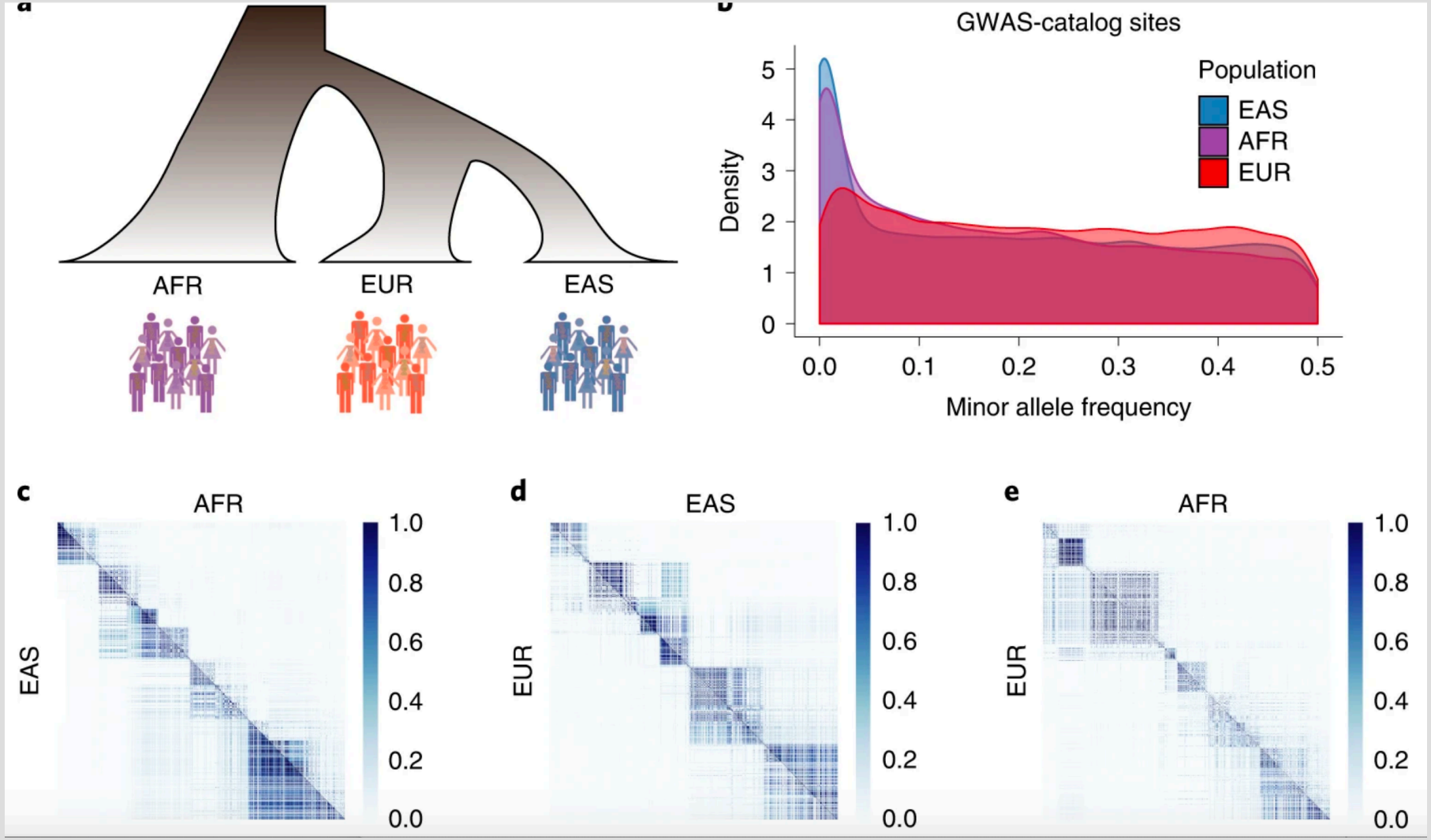



Martin et al. 2019 Nat Gen: **Clinical use of current polygenic risk scores may exacerbate health disparities**




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

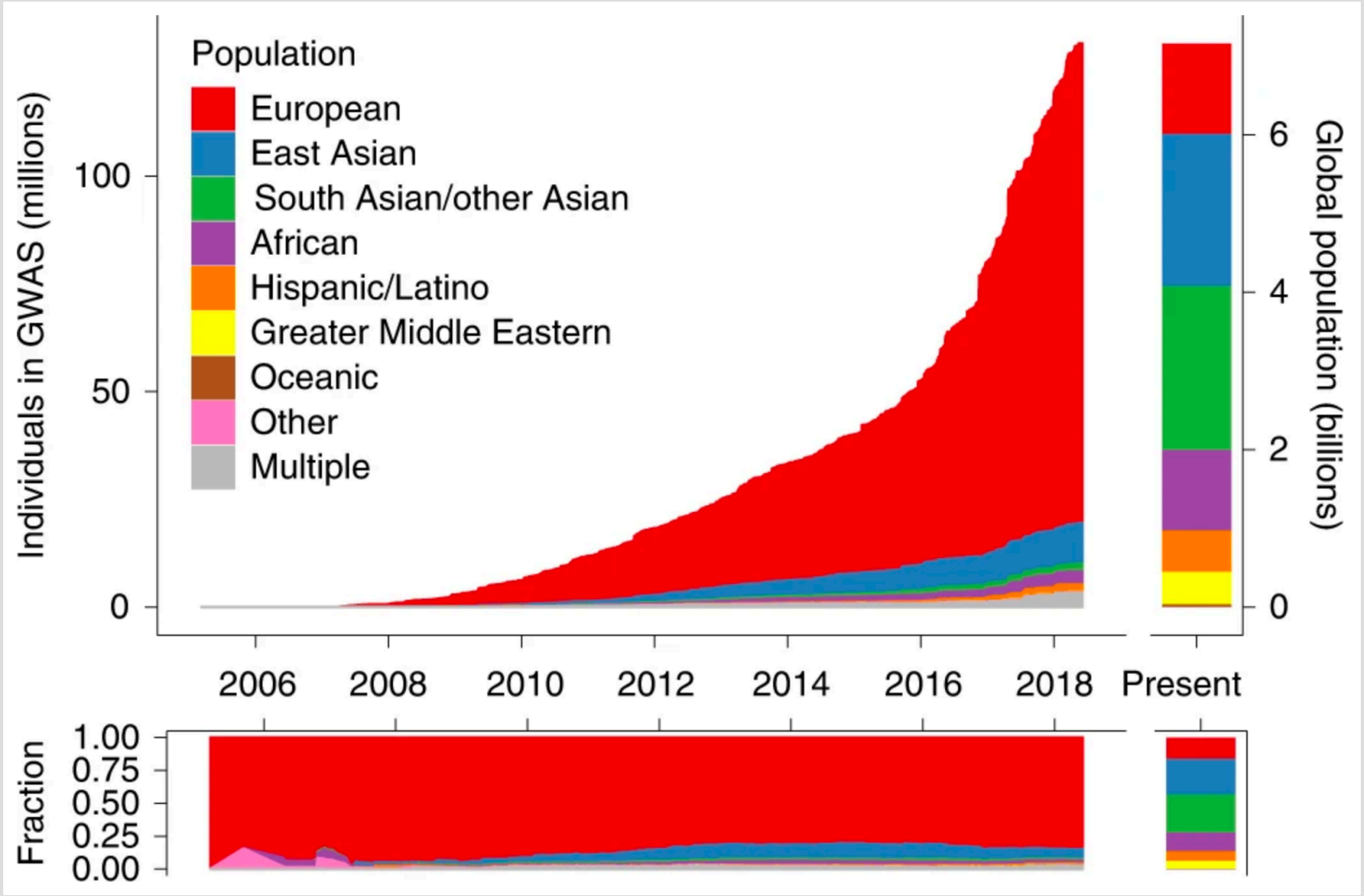
# SOME CAUSES FOR DISPARITIES




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

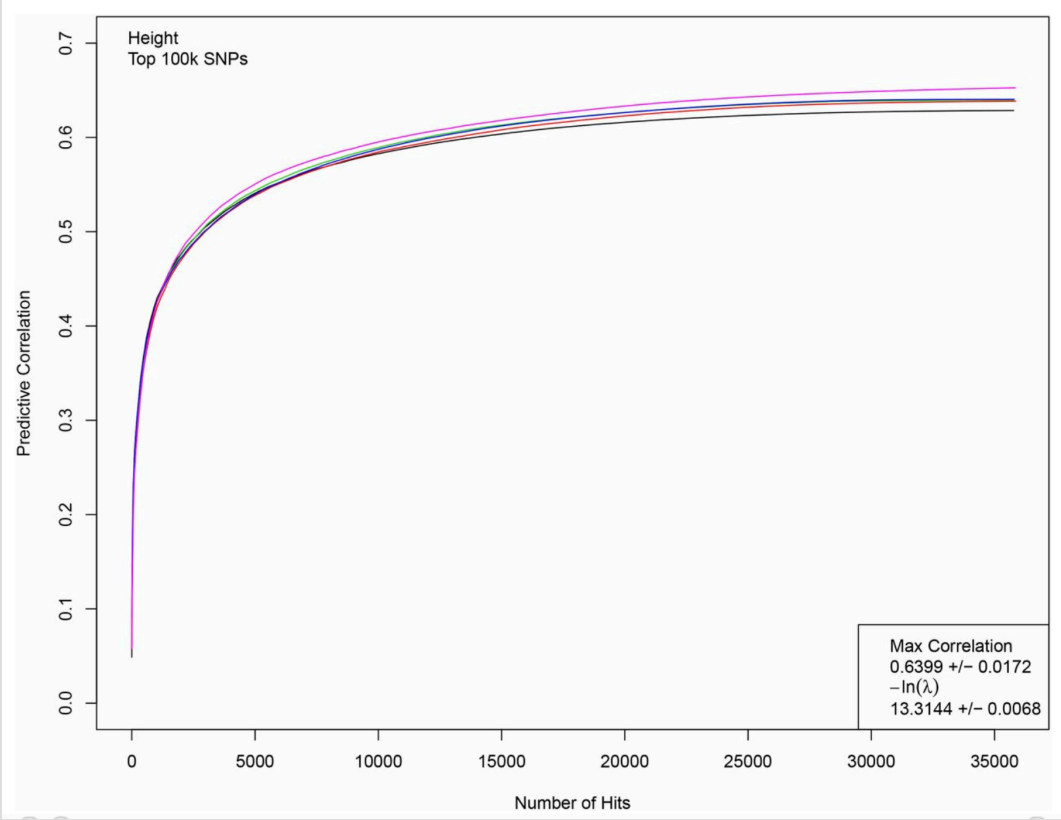
# Accurate Genomic Prediction of Human Height

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

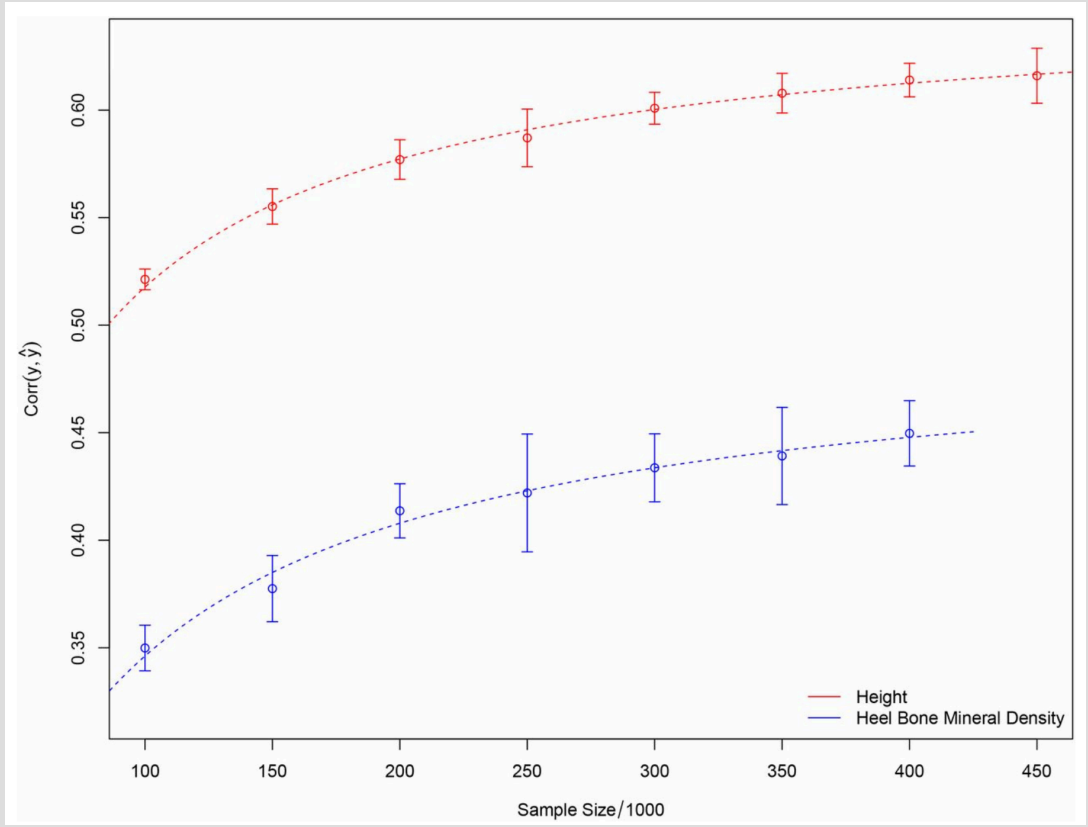
GENETICS *October 1, 2018 vol. 210 no. 2 477-497; https://doi.org/10.1534/genetics.118.301267*

- 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

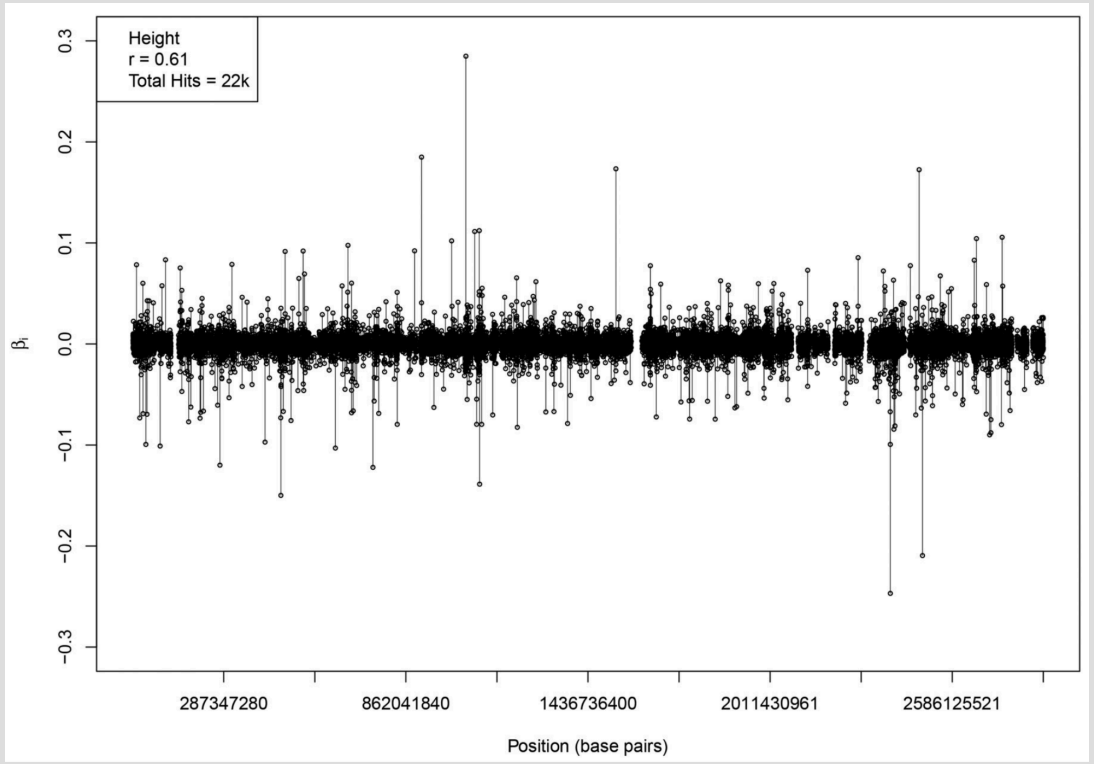


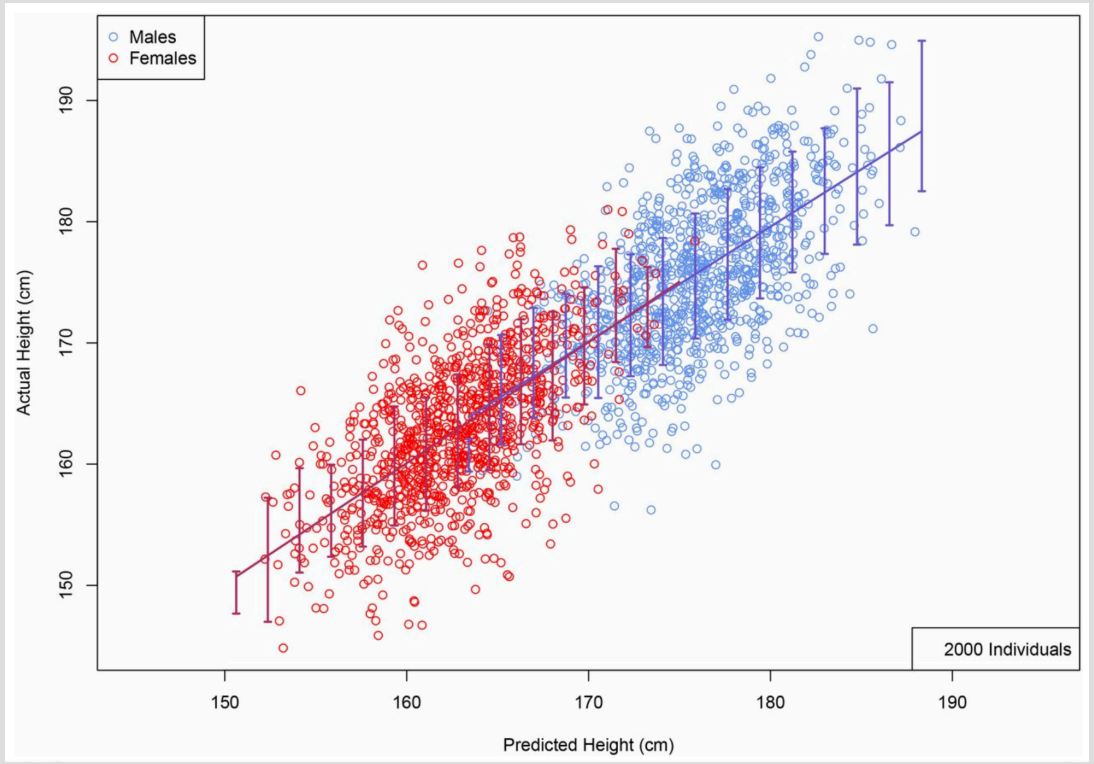

How many non-zero coefficients in the model?




How large GWAS was used?

# FINAL MODEL



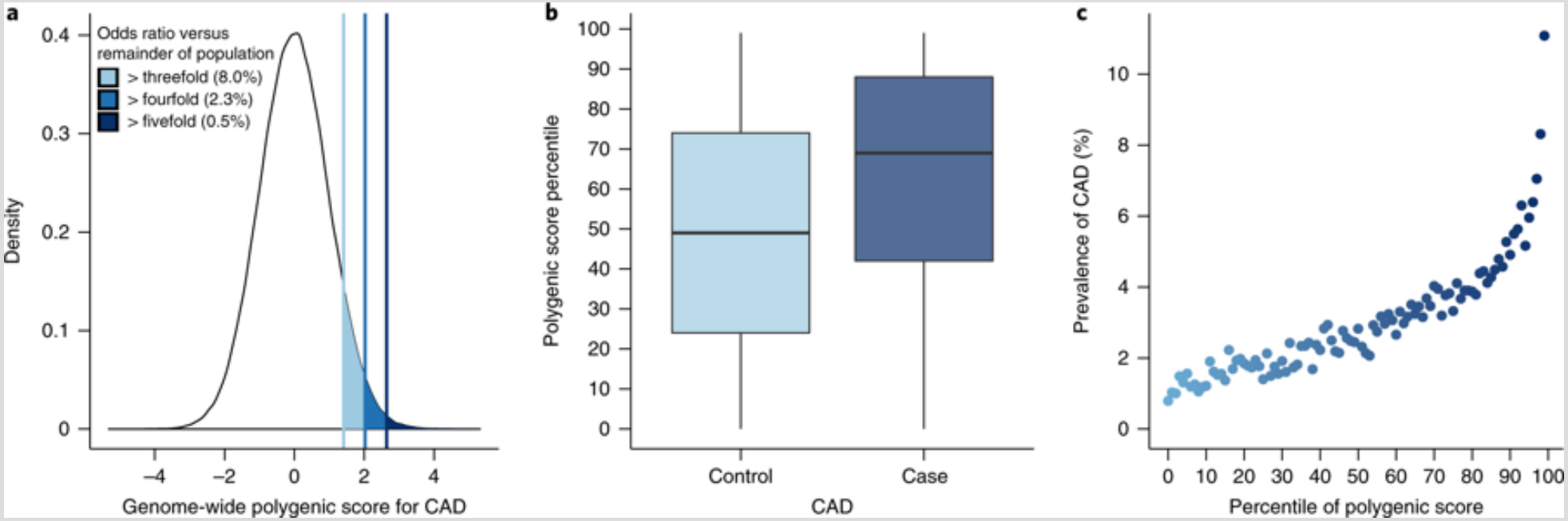




Uses 22,000 non-zero coefficients for SNPs across genome

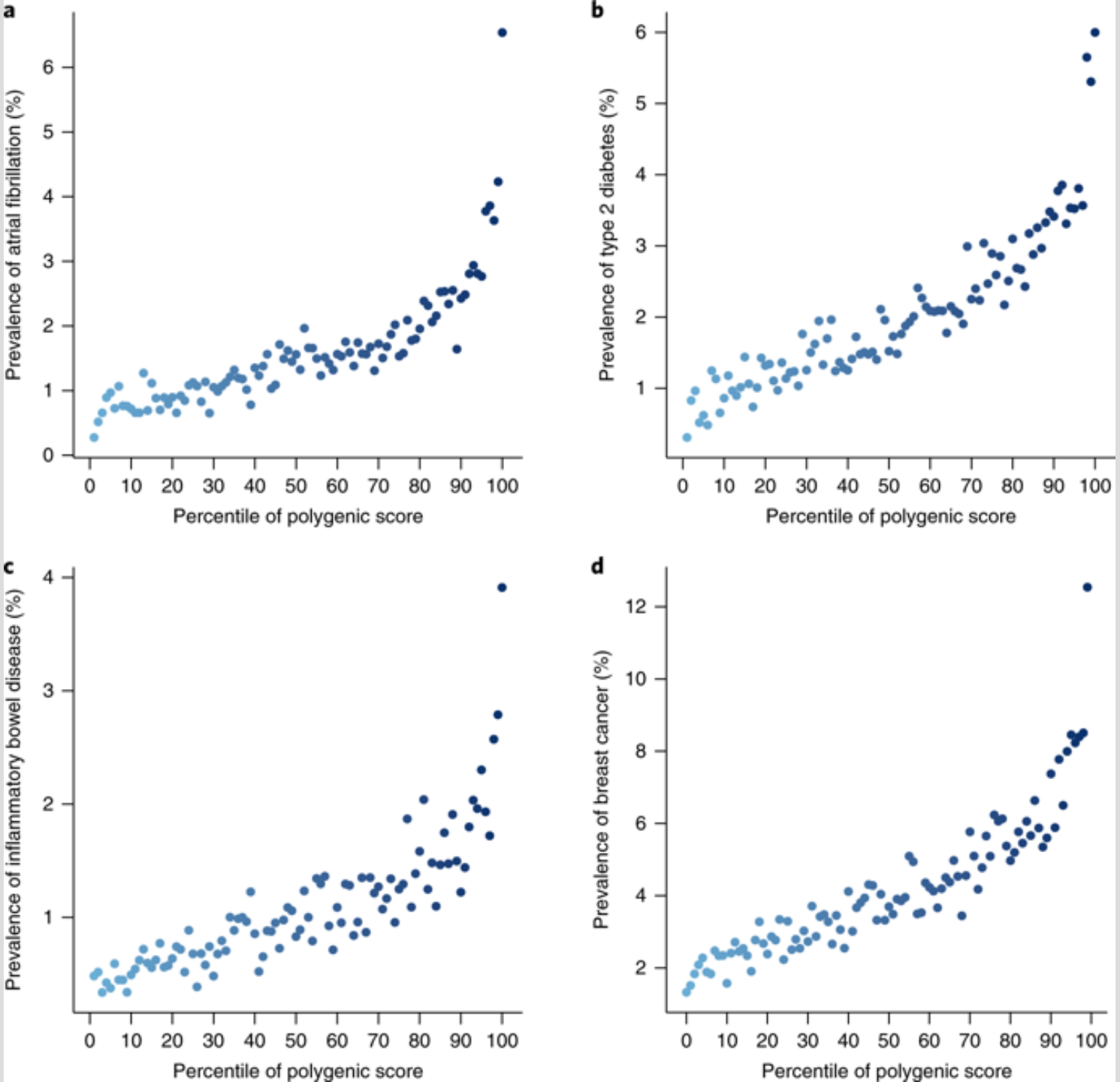
Achieves r = 0.58, i.e., $R^2$ = 0.34 in UKB test data set and r = 0.54, i.e., $R^2$ = 0.29 in ARIC data that are independent of UKB.

# 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 GPS. **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.

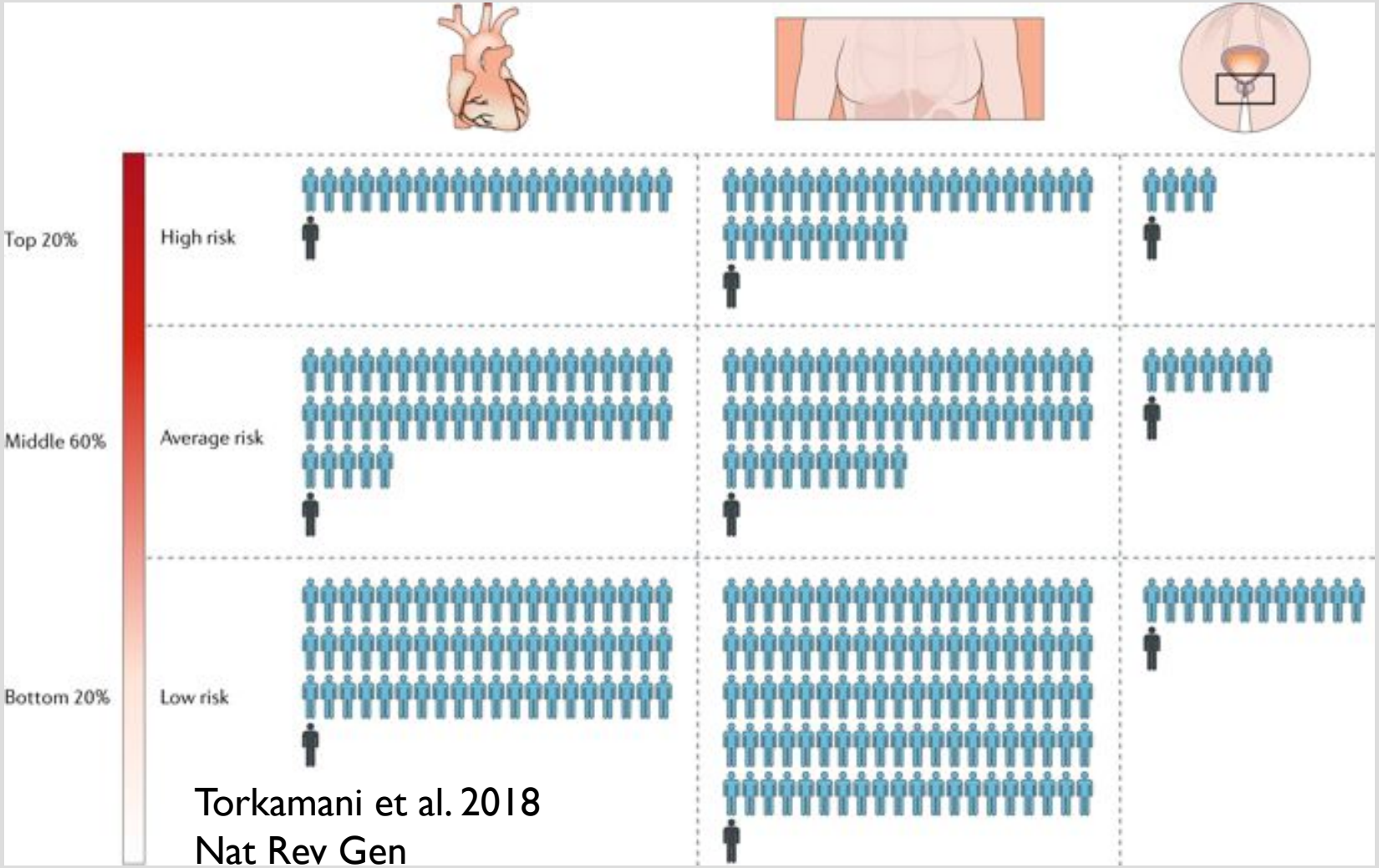
Khera et al. 2018 Nat Gen

# CLINICAL UTILITY

Khera et al. 2018
Nat Gen

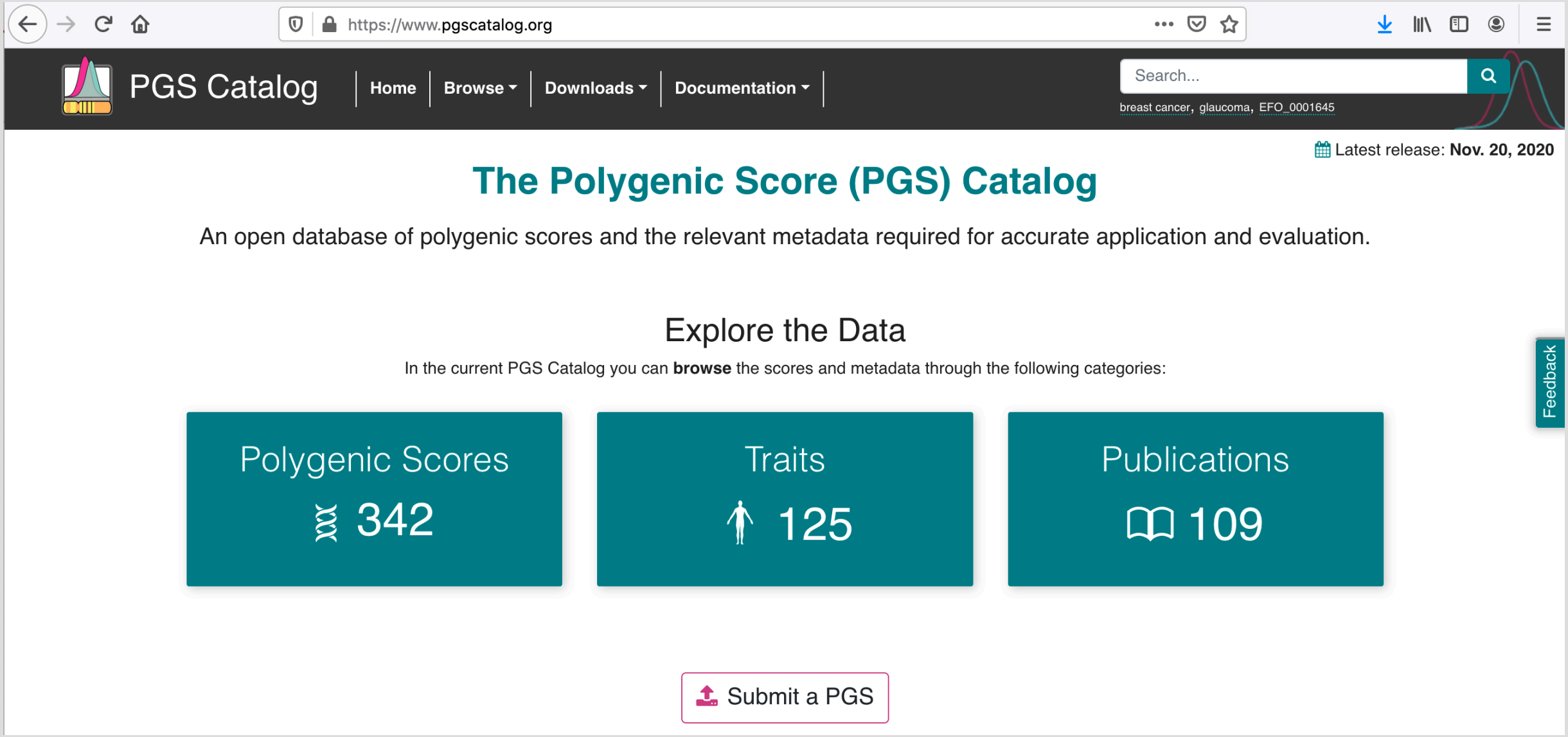
| High GPS definition | Reference group | Odds ratio | 95% CI | *P* value |
|---|---|---|---|---|
| **CAD** | | | | |
| Top 20% of distribution | Remaining 80% | 2.55 | 2.43–2.67 | $<1 \times 10^{-300}$ |
| Top 10% of distribution | Remaining 90% | 2.89 | 2.74–3.05 | $<1 \times 10^{-300}$ |
| Top 5% of distribution | Remaining 95% | 3.34 | 3.12–3.58 | $6.5 \times 10^{-264}$ |
| Top 1% of distribution | Remaining 99% | 4.83 | 4.25–5.46 | $1.0 \times 10^{-132}$ |
| Top 0.5% of distribution | Remaining 99.5% | 5.17 | 4.34–6.12 | $7.9 \times 10^{-78}$ |
| **Breast cancer** | | | | |
| Top 20% of distribution | Remaining 80% | 2.07 | 1.97–2.19 | $3.4 \times 10^{-159}$ |
| Top 10% of distribution | Remaining 90% | 2.32 | 2.18–2.48 | $2.3 \times 10^{-148}$ |
| Top 5% of distribution | Remaining 95% | 2.55 | 2.35–2.76 | $2.1 \times 10^{-112}$ |
| Top 1% of distribution | Remaining 99% | 3.36 | 2.88–3.91 | $1.3 \times 10^{-54}$ |
| Top 0.5% of distribution | Remaining 99.5% | 3.83 | 3.11–4.68 | $8.2 \times 10^{-38}$ |

# UTILITY OF PRS IN TREATMENT




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.


https://www.pgscatalog.org
PGS Catalog
Home
Browse
Downloads
Documentation
Search...
breast cancer, glaucoma, EFO_0001645
Latest release: Nov. 20, 2020
The Polygenic Score (PGS) Catalog
An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.
Explore the Data
In the current PGS Catalog you can browse the scores and metadata through the following categories:
Polygenic Scores
342
Traits
125
Publications
109
Submit a PGS
Feedback