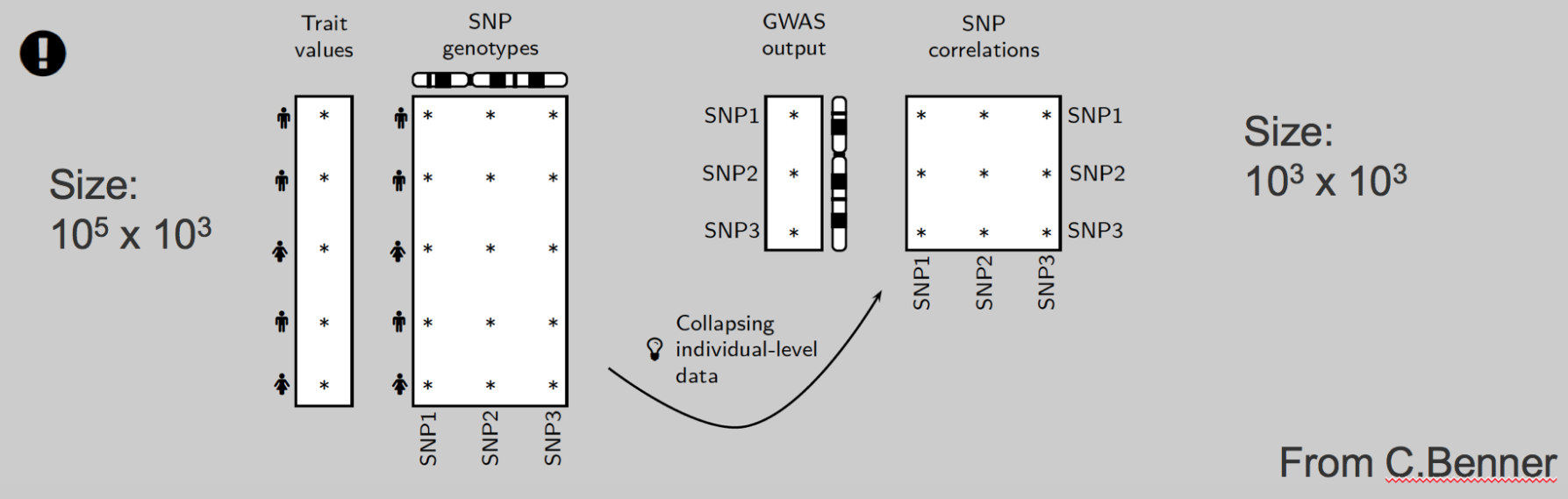
# GWAS 9

Matti Pirinen
University of Helsinki
Week 6

# SUMMARY STATISTICS

- Many downstream analyses, such as stepwise forward selection or fine-mapping ca be done using only summary statistics from GWAS rather than requiring access to original genotype-phenotype data
- Summary statistics include: $\hat{\beta}$, SE, *P*-value, MAF, LD, quality control values
- Benefits: saving storage and computation, avoiding privacy concerns of individual-level sensitive data



# 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 one 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

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

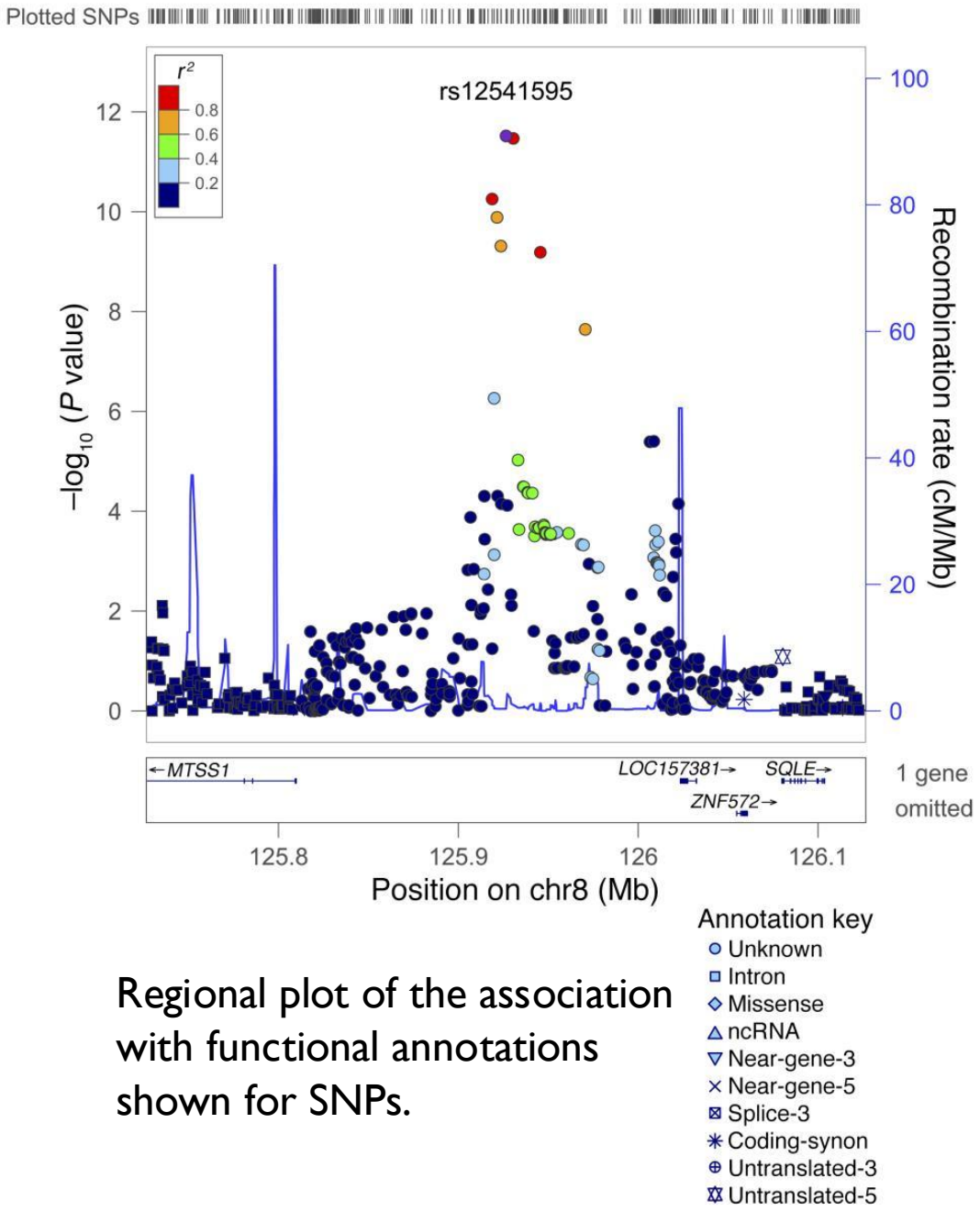
studies 1,…, K

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

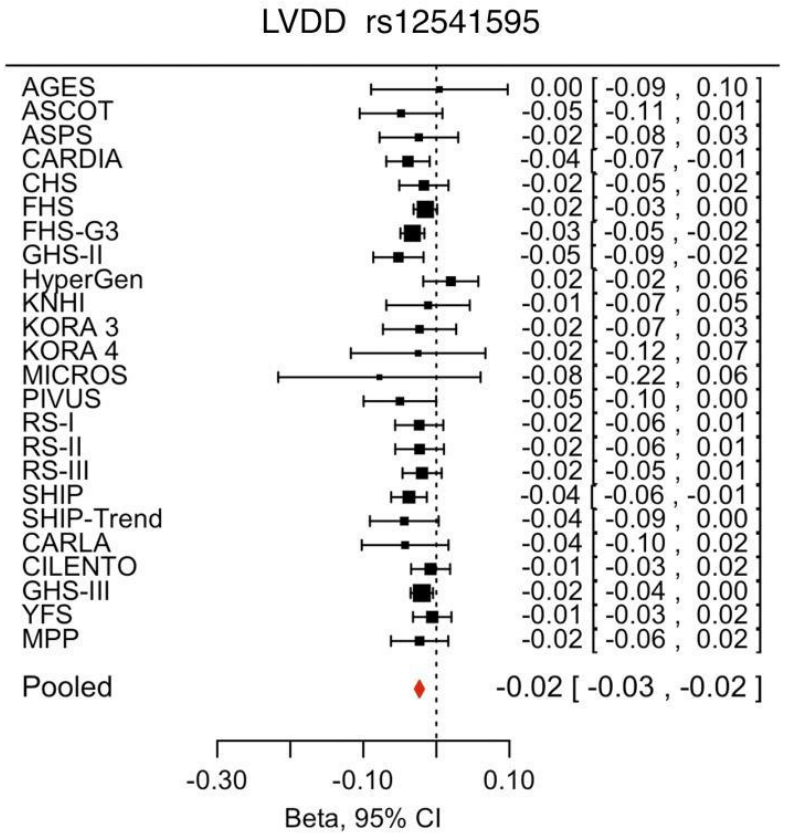
where the weight

$$w_{kl} = \frac{1}{\mathrm{SE}_{kl}^2}$$ 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, **not** on the odds-ratio scale




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

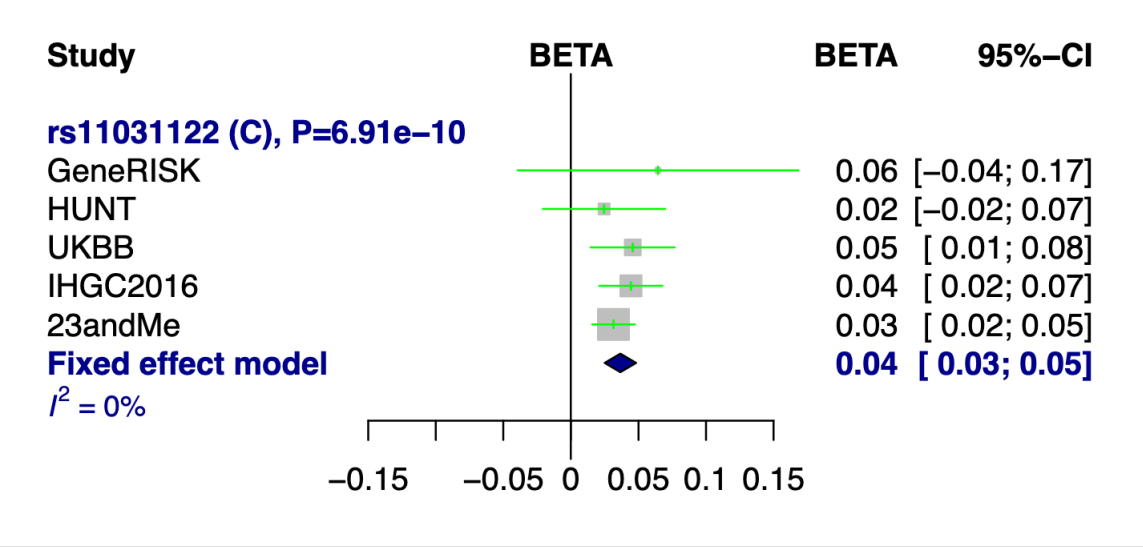



Forest plot for the meta-analysis of the association between rs12541595 and left ventricular diastolic internal dimension (LVDD)

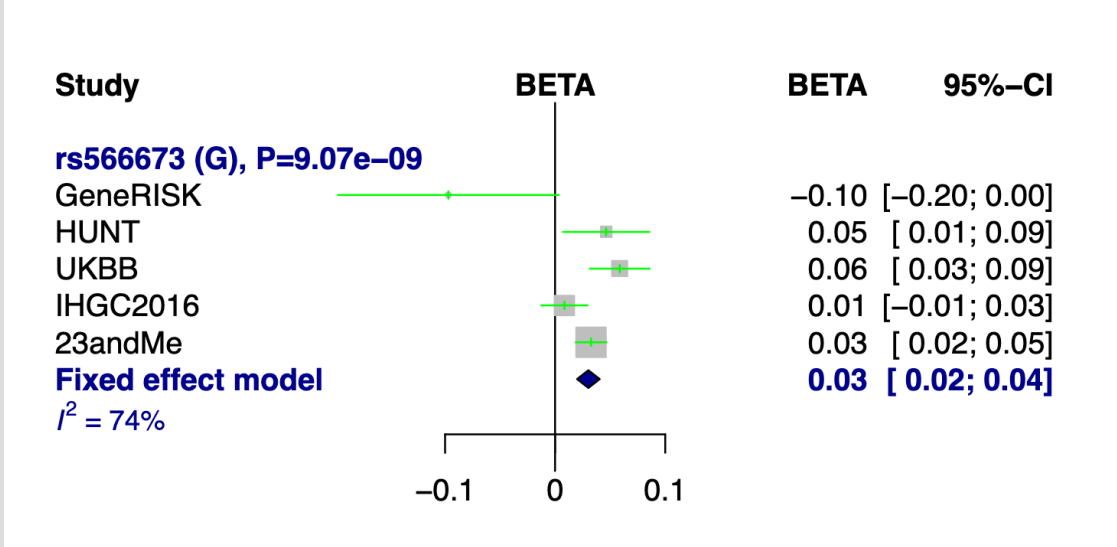
*J Clin Invest.* 2017;127(5):1798-1812

# FIXED EFFECT ASSUMPTION

- Inverse-variance weighted meta-analysis makes an assumption that every component study of the meta-analysis is estimating the same underlying effect size

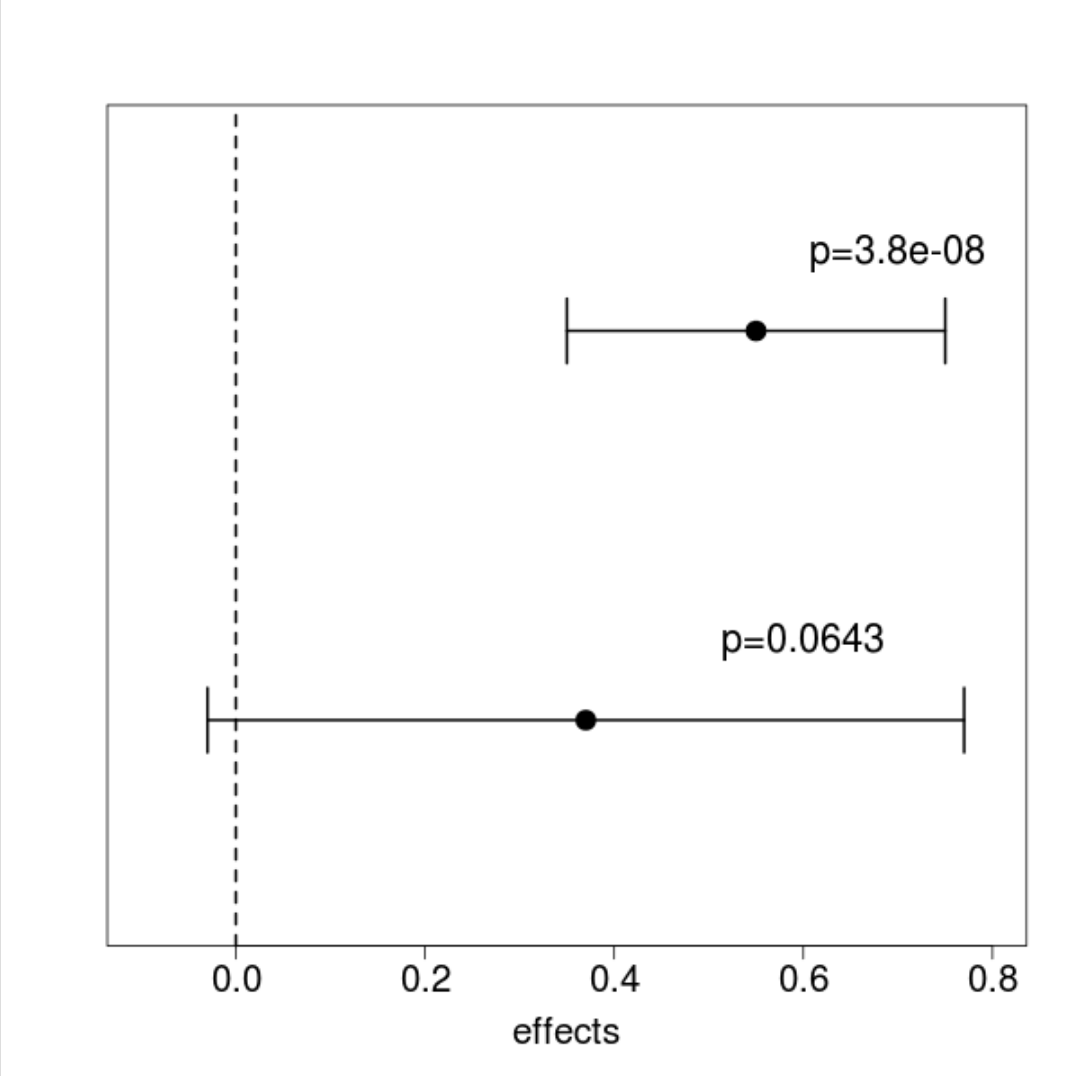



Fixed effect assumption seems valid.




Fixed effect assumption seems questionable.

# IS FIXED EFFECT ASSUMPTION REASONABLE?



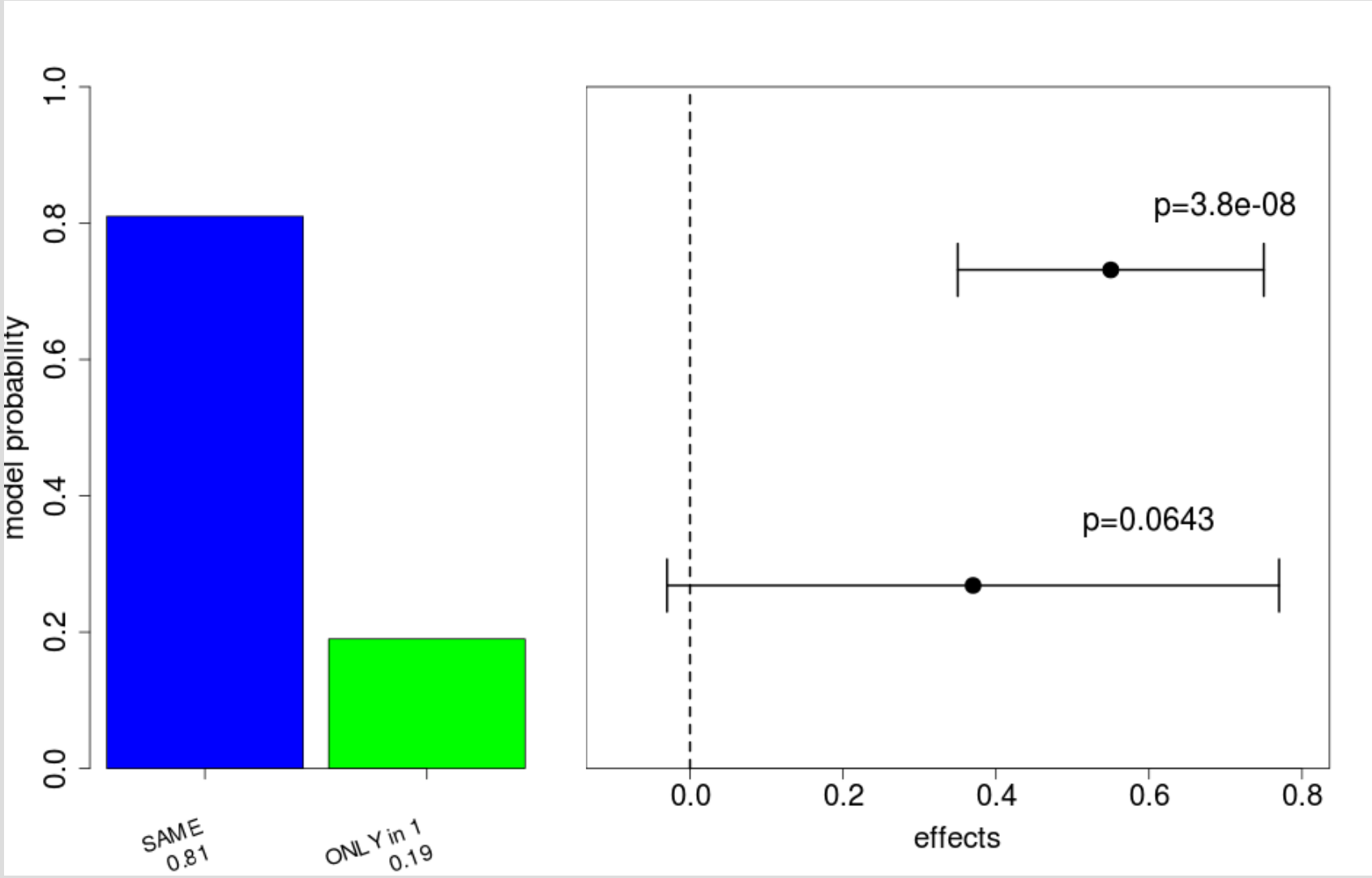

Suppose we have two estimates.

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?

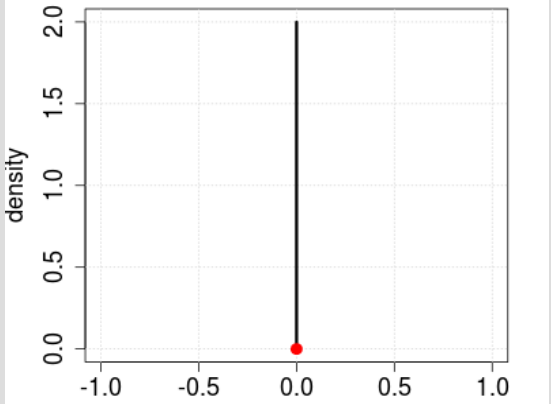



We write each of the possible explanations of the data in terms of a statistical model and compare how well each of the models describes the data by using a Bayesian model comparison framework.

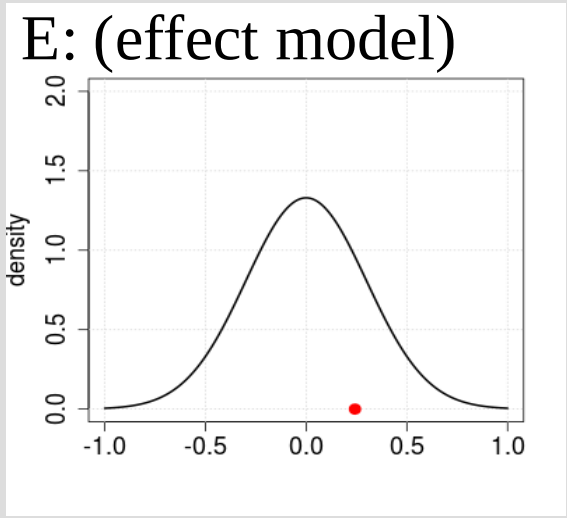
# BUILDING MODELS FOR 2 EFFECTS

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

N: (null model)




E: (effect model)



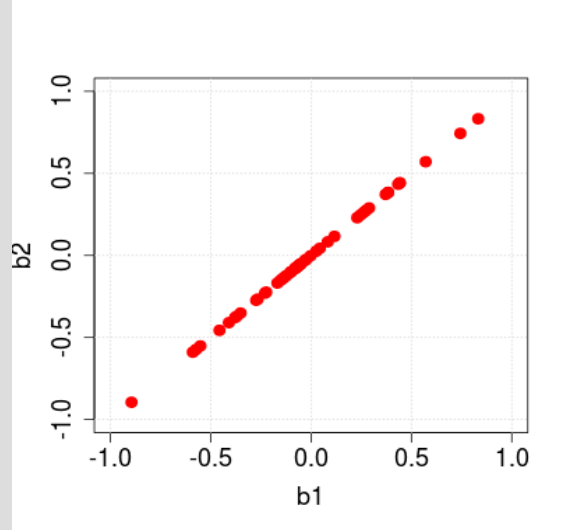

Here we have two SNPs and build joint models for them.

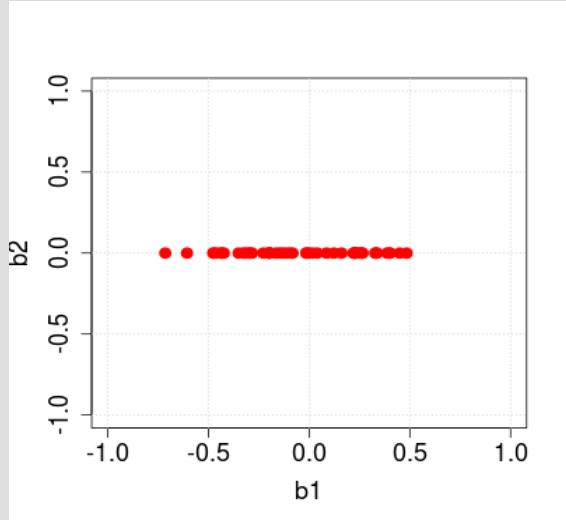
SAME EFFECT:
- Pick value $x \sim E$
- Set $\beta_1 = \beta_2 = x$

EFFECT IN ONLY 1:
- Pick $\beta_1 \sim E$
- Pick $\beta_2 \sim N$

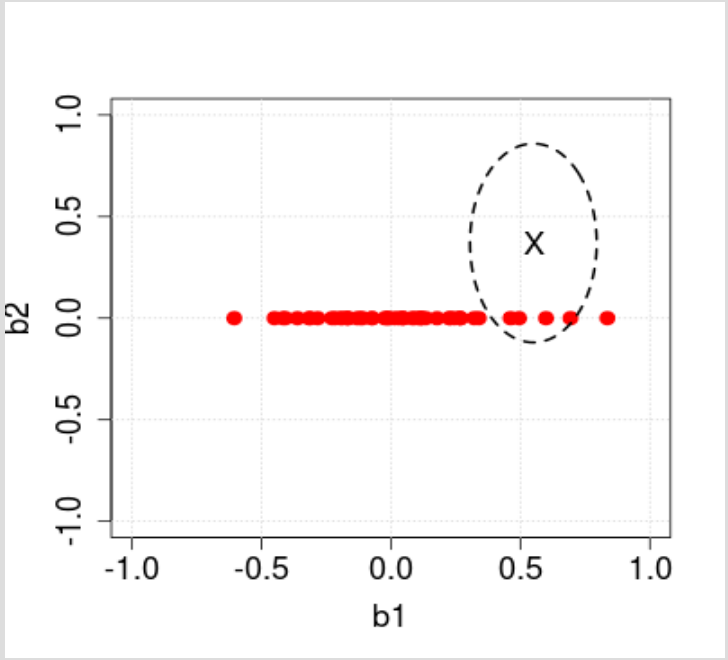
i.e. $\beta_2 = 0$

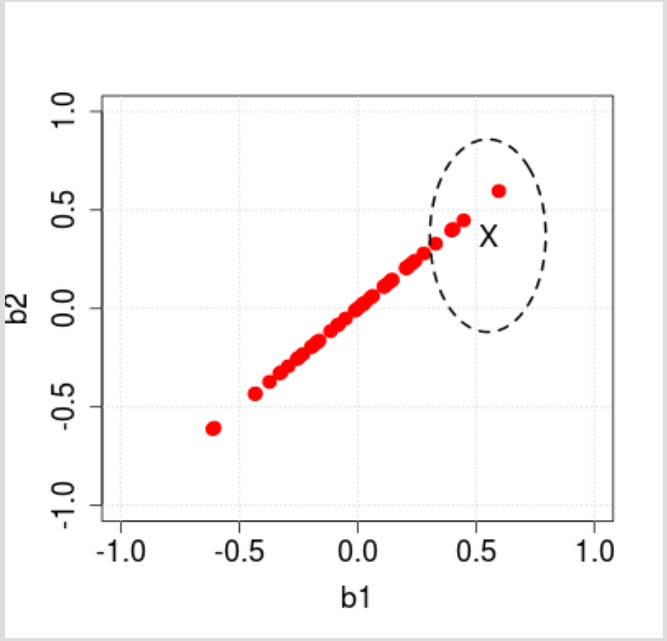
Example data from each model



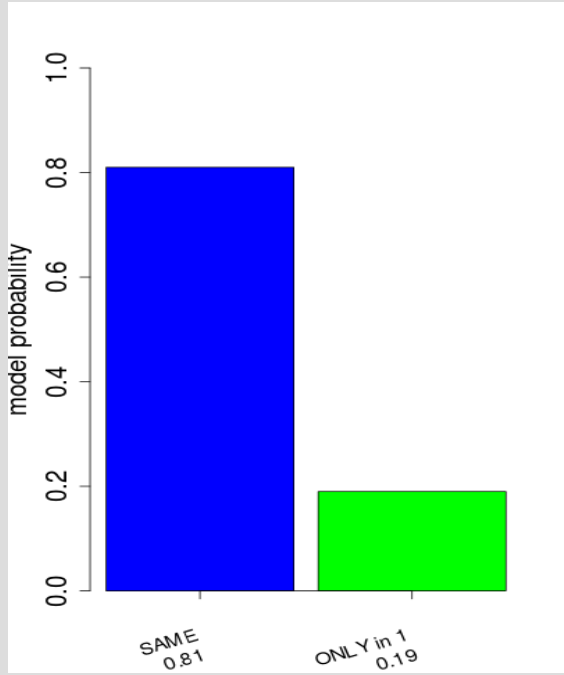
# HOW WELL THE MODELS EXPLAIN THE DATA?




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

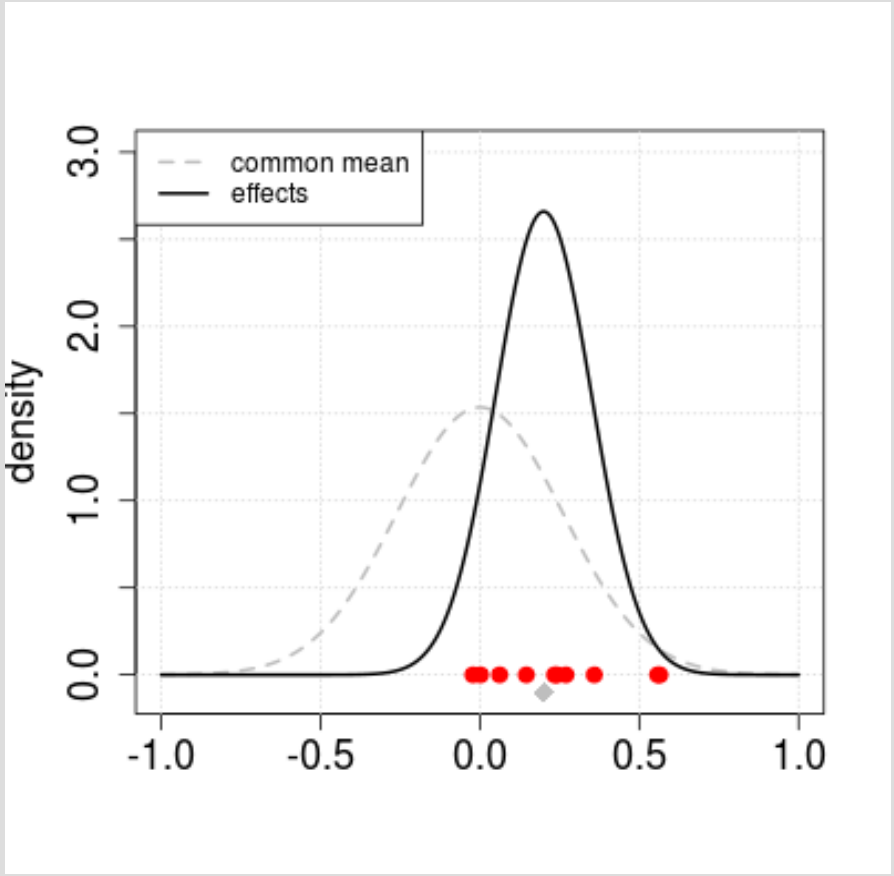
Same effect model is a better explanation here.




Bayes factor compares two models

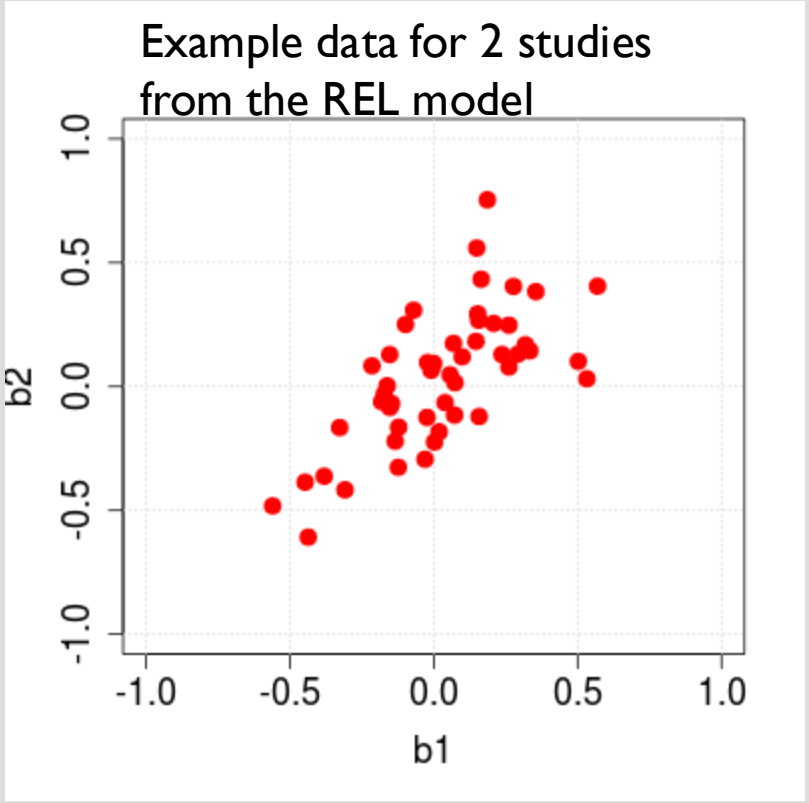
$$\mathrm{BF}_{S:1} = \frac{\Pr(\text{Data} \mid \text{Model "Same"})}{\Pr(\text{Data} \mid \text{Model "Only in 1"})}$$

# RELATED EFFECTS MODEL

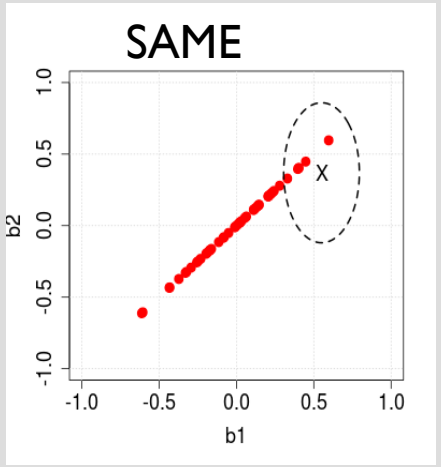


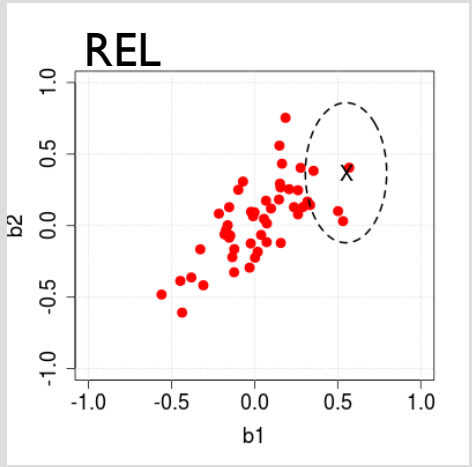

Model "REL"
- Pick $\mu \sim N(0, s^2)$
- Given $\mu$, pick each $\beta_i \sim N(\mu, t^2)$

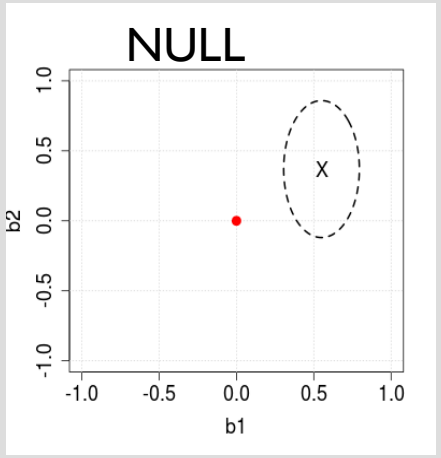
Example data for 2 studies from the REL model

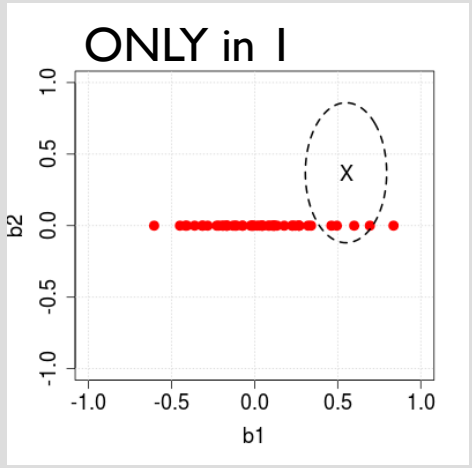


# MODEL COMPARISON

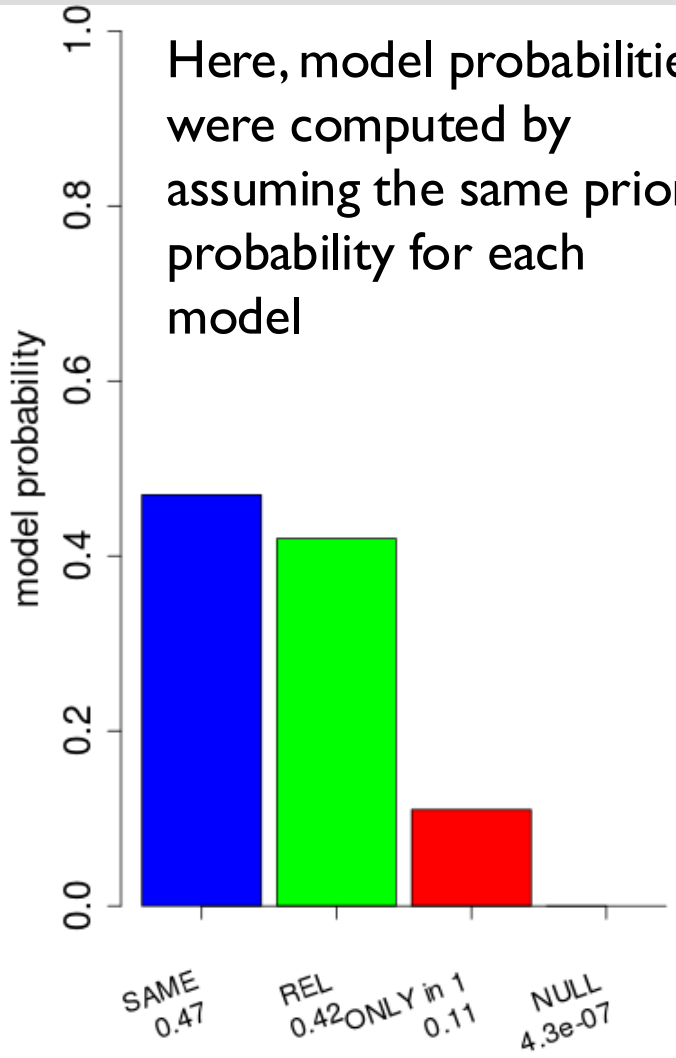


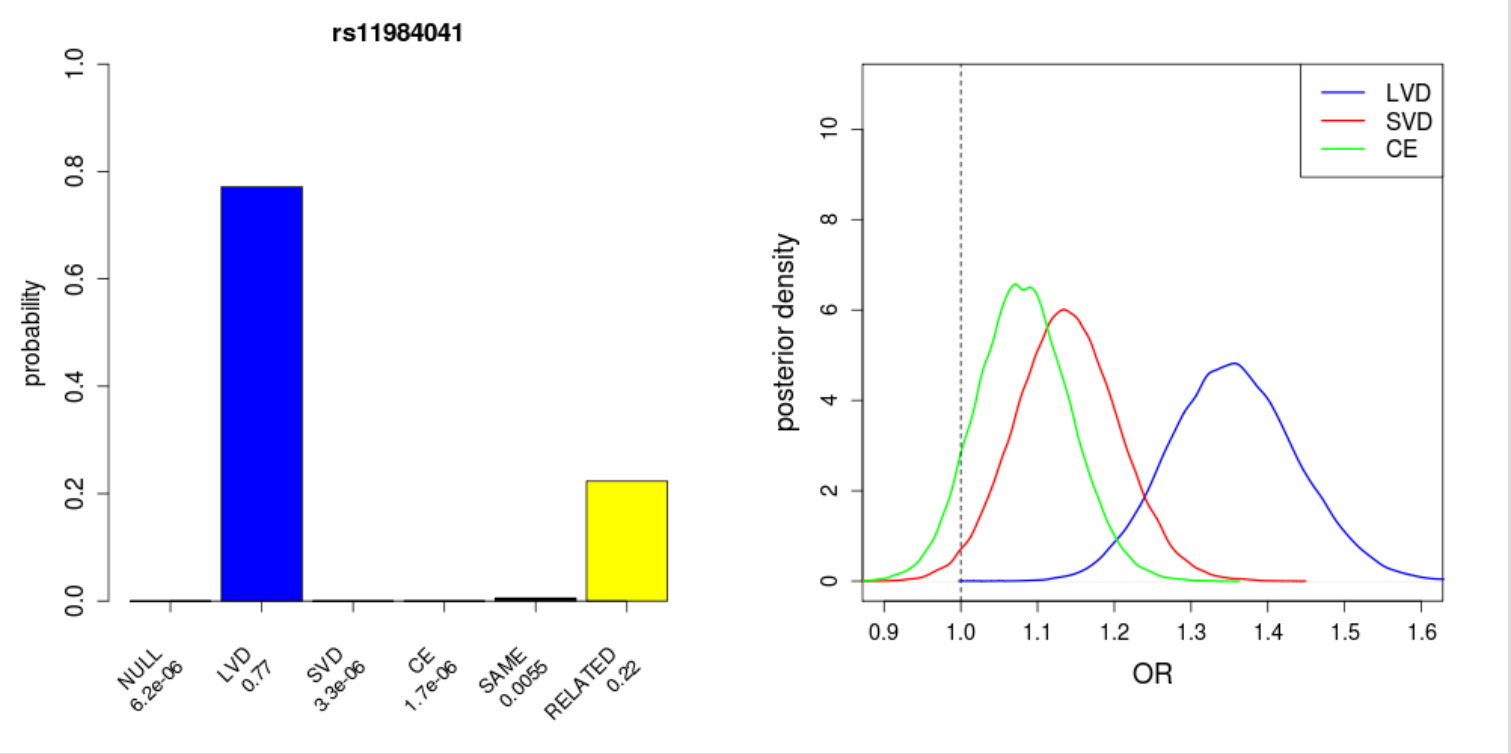

Here, model probabilities were computed by assuming the same prior probability for each model



# 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

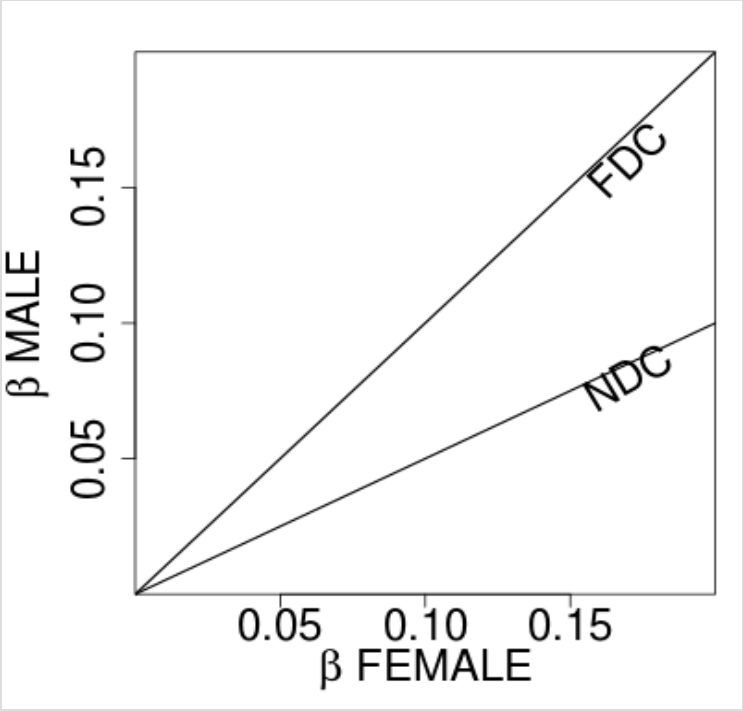



*HDAC9* SNP has a strong effect in LVD
and not the same effect in the 2 other subtypes.

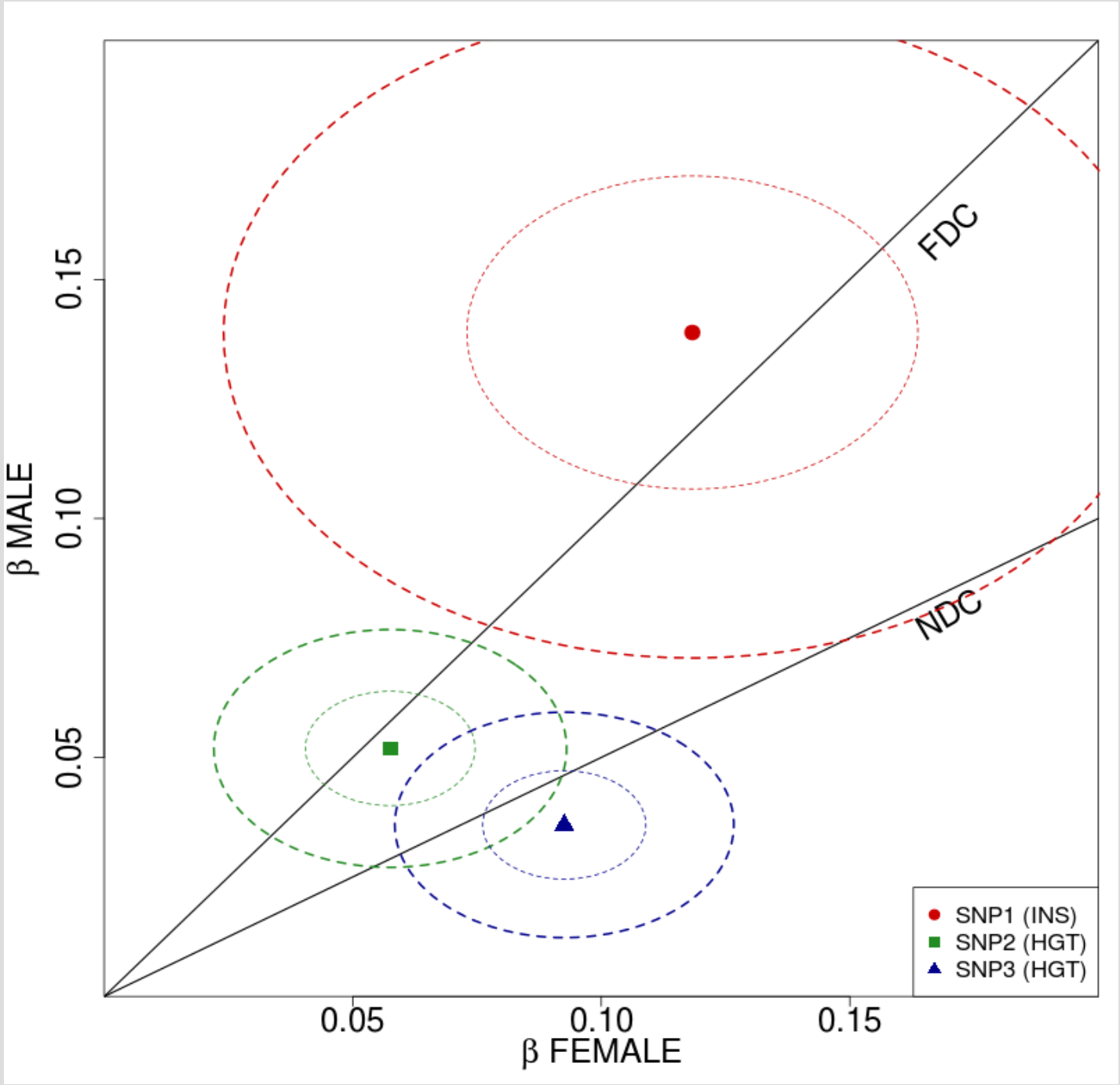
Bellenguez et al. 2012 Nat Gen

# CHR X DOSAGE COMPENSATION

- One of the female's X chrs in each cell is inactivated
  - Balances out the difference in chr X count 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

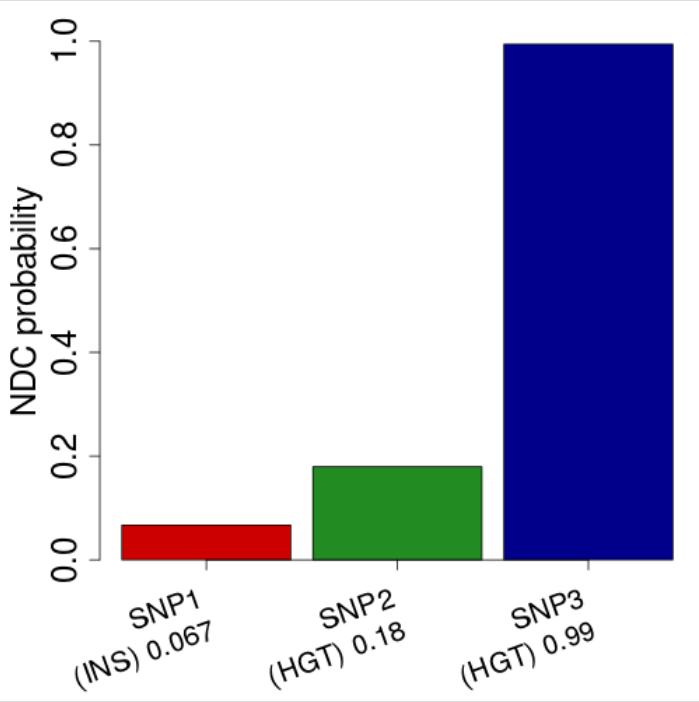


Tukiainen et al. 2014 PLoS Gen



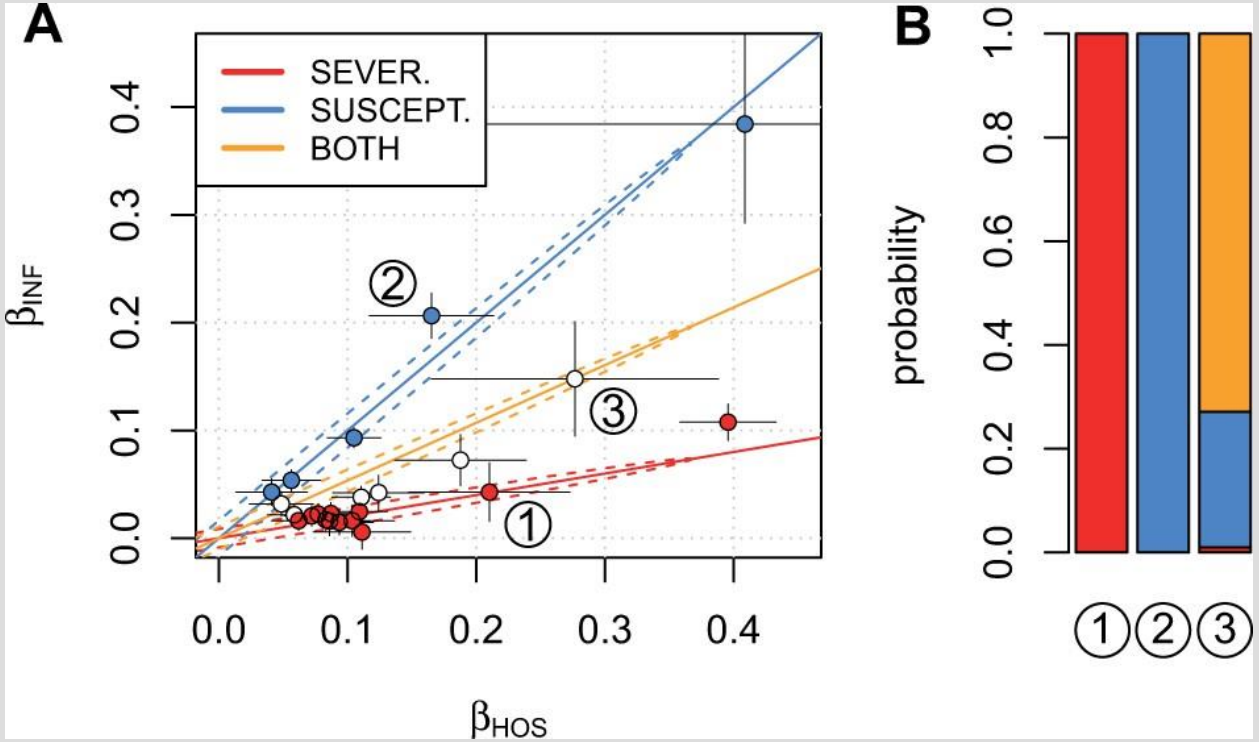

We have 3 chr X associations, 1 with insulin levels and 2 with height.

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

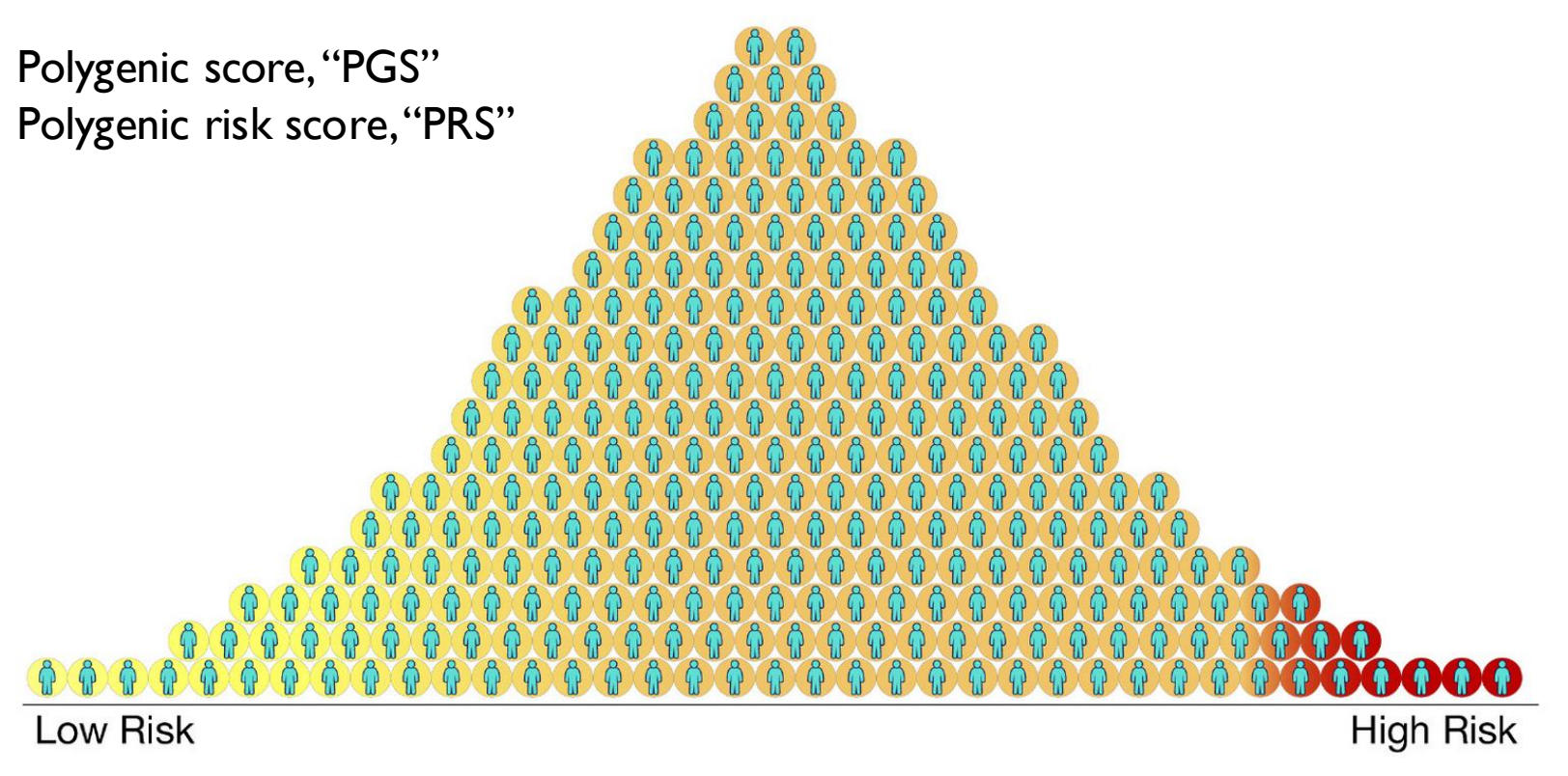



Question:
Which variants affect susceptibility to infection and which severity of the disease?
Basis for inference are differences between effect sizes from infection GWAS and hospitalization GWAS.
Figure defines line models for susceptibility and severity variants, and variants that affect both.

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




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

Figure: NIH

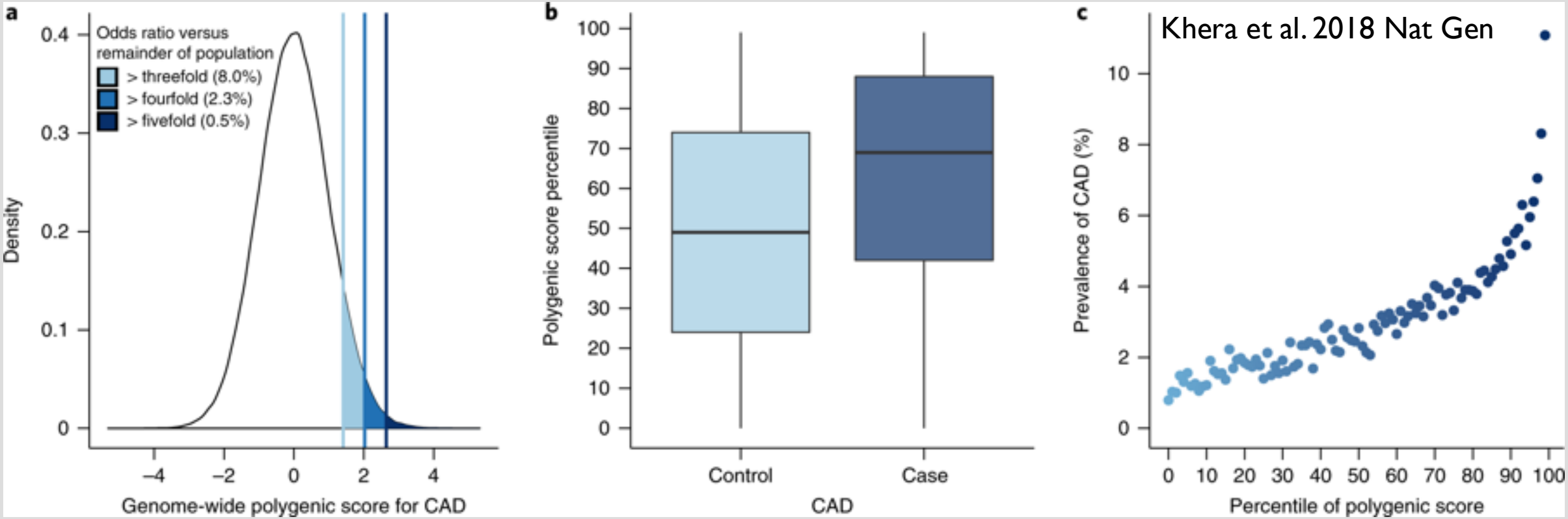
# (FUTURE) USES OF GENETIC SCORES


Help in prevention

- lifestyle change
- screening programmes

How best to treat this person?

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



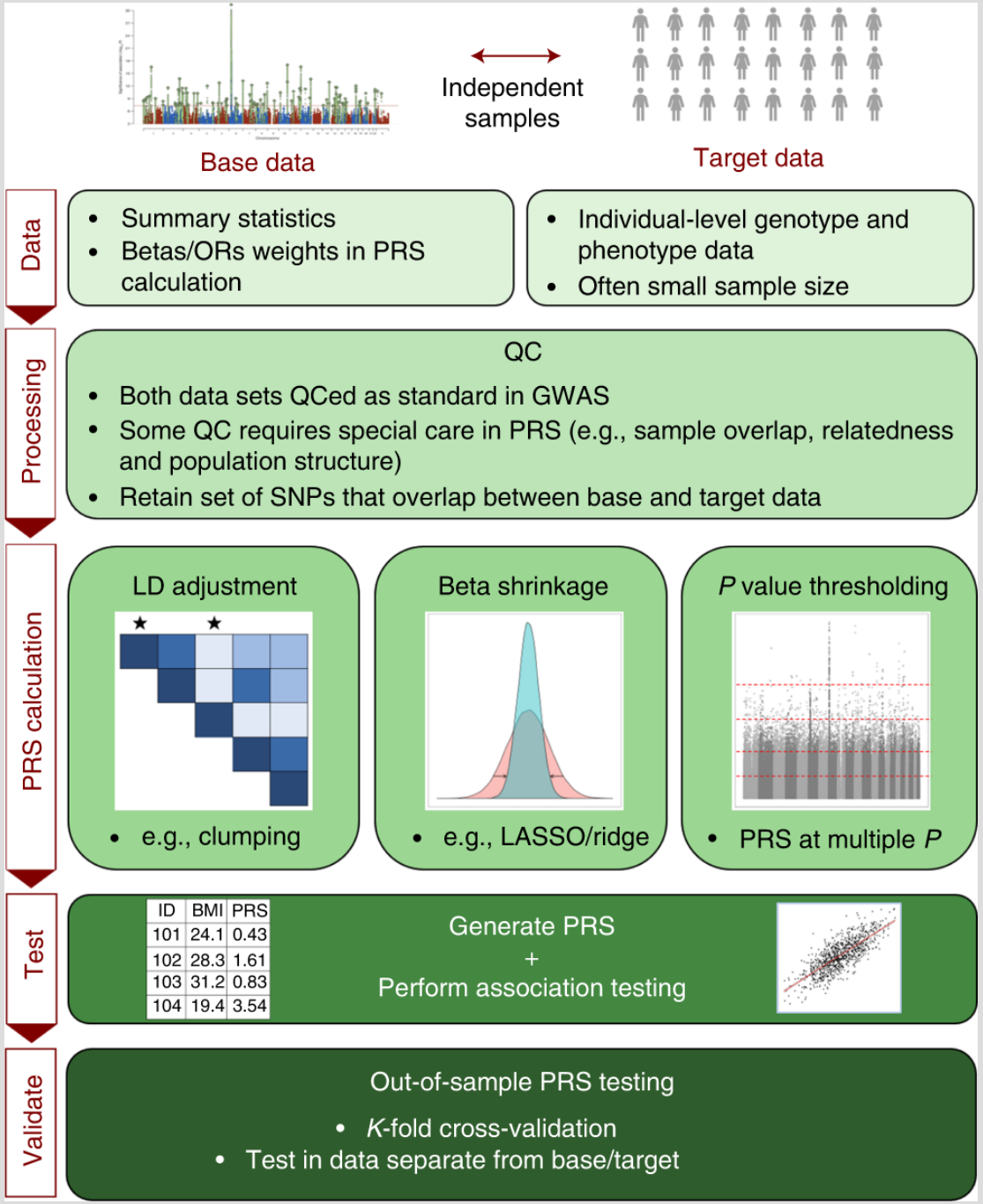

**a.** Distribution of $PGS_{CAD}$ in the UK Biobank testing dataset ($n$ = 288,978). The $x$ axis represents $PGS_{CAD}$, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation.
Shading reflects the proportion of the population with three-, four-, and fivefold increased 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 ($g_{ik}$) at variants $k = 1, \ldots K$
- Compute PRS for individual *i* as sum

$$PRS_i = \sum_{k=1}^{K} g_{ik}\,\hat{\beta}_k$$

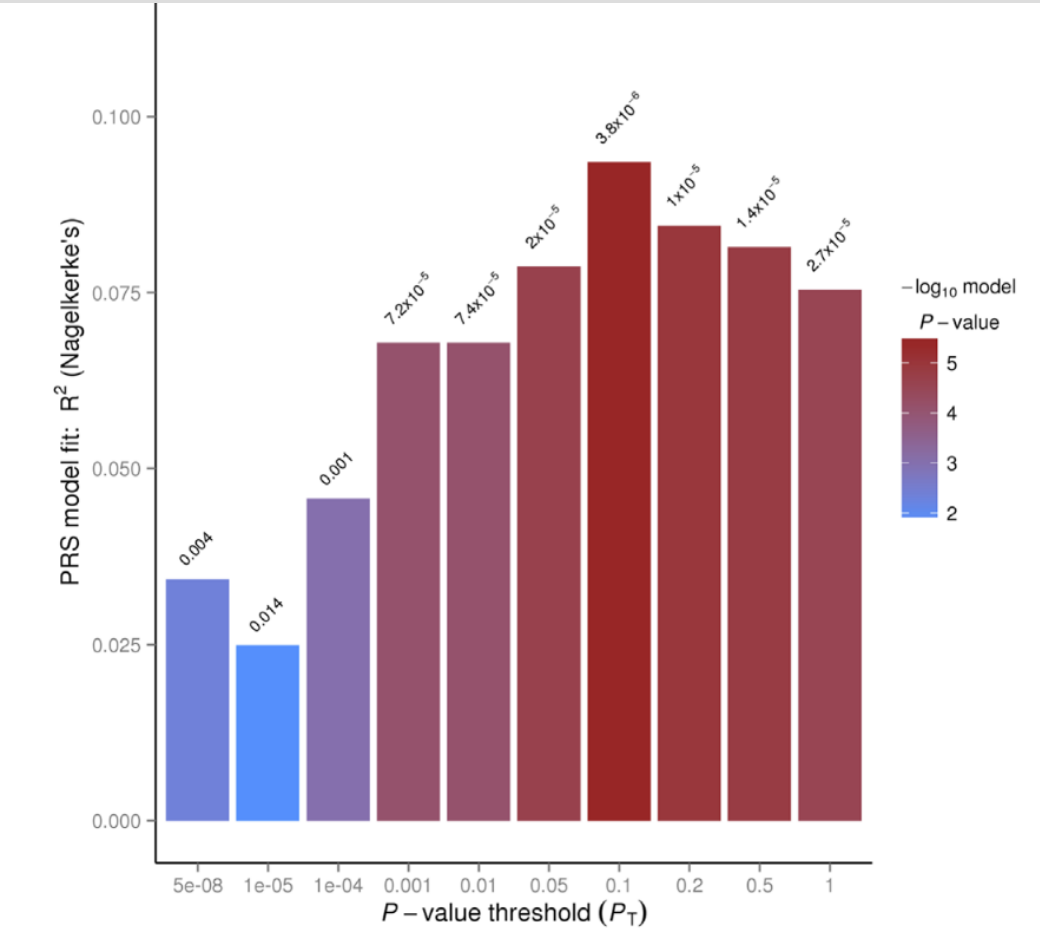
Choi et al. 2020 Nat Protocols



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

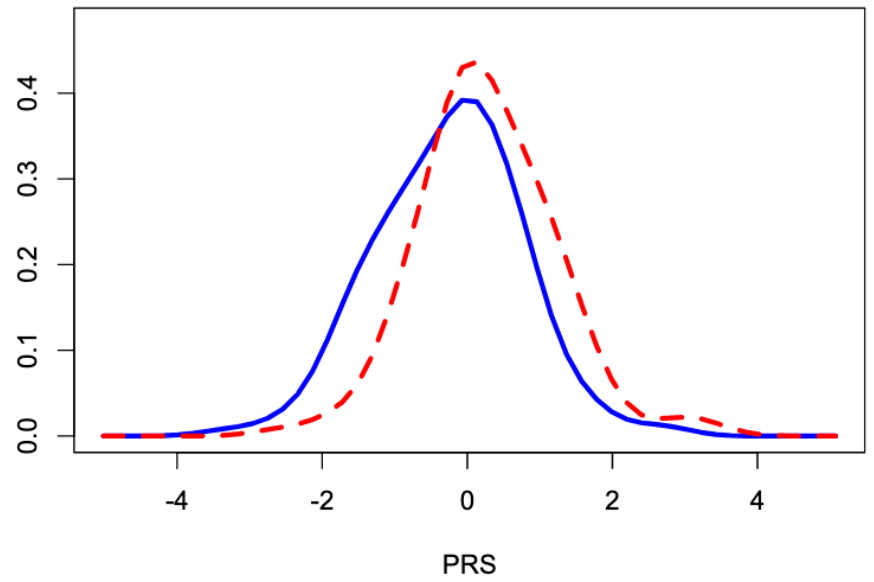



**Figure 2.** Density distribution of polygenic risk score (PRS) in European first-episode psychosis case and control subjects. PRS represents the standardized residuals of PRS after adjustment for the 10 principal components. Blue line indicates control subjects; red line, case subjects.

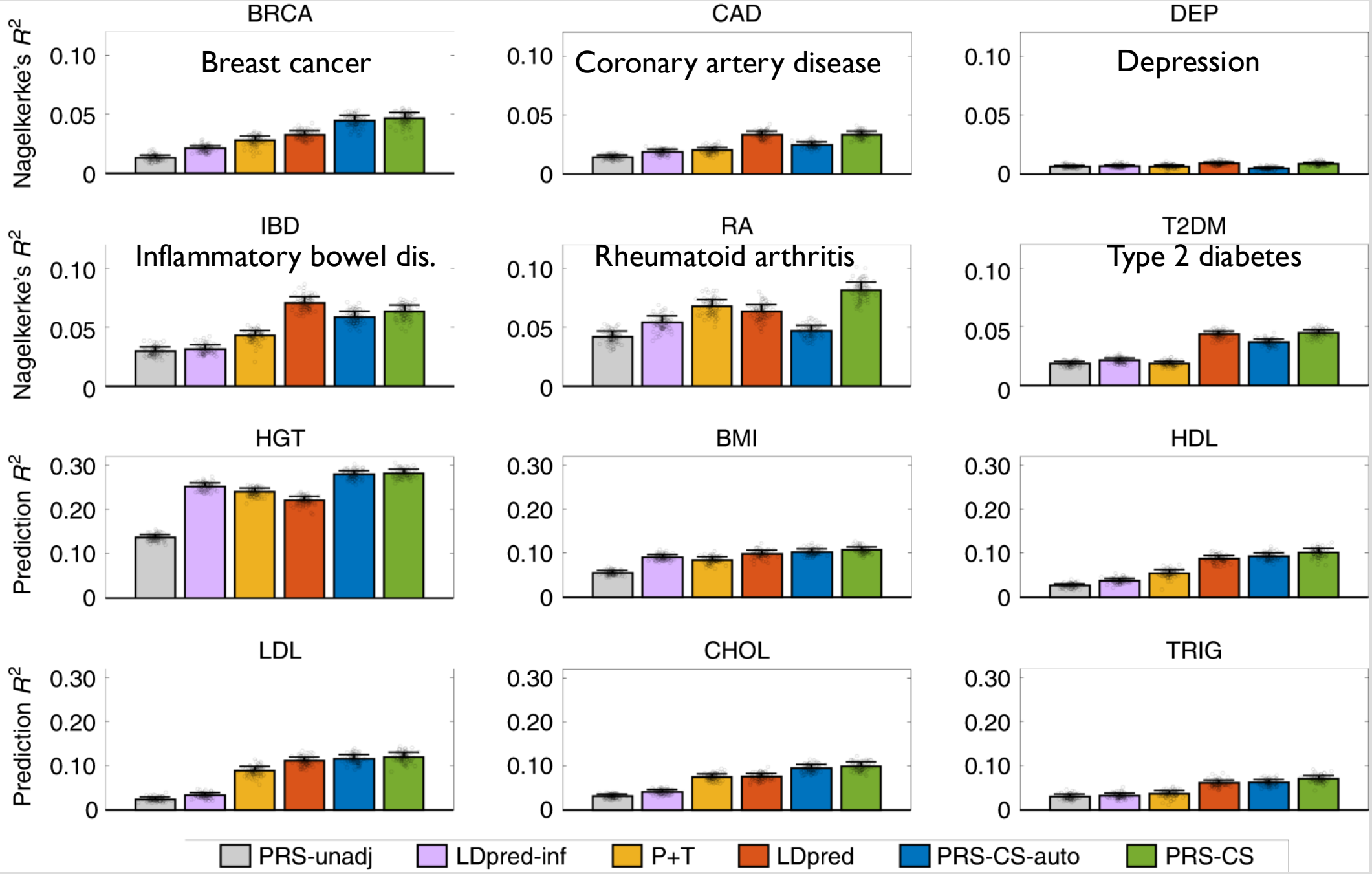
Goal: Predicting psychosis cases by schizophrenia PRS.
Barplot on left: Optimal PRS uses $P_{thr}$ = 0.1.
$r^2$ threshold was fixed to 0.1 (not tuned).
Computed using PRSice software.

# 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 $E(\boldsymbol{\lambda} | \widehat{\boldsymbol{\beta}}, \boldsymbol{R}, h^2, \theta)$, 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 the best performing model is chosen
- These estimates of causal effects are used as weights in PRS computation



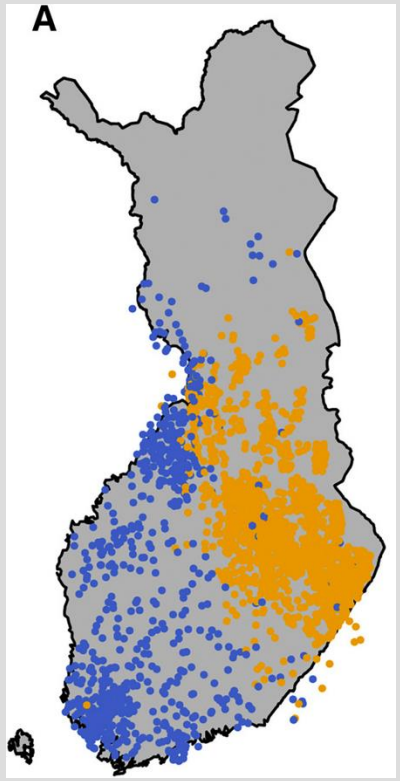

Ge et al. Nat Comm 2019

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

# BIASES

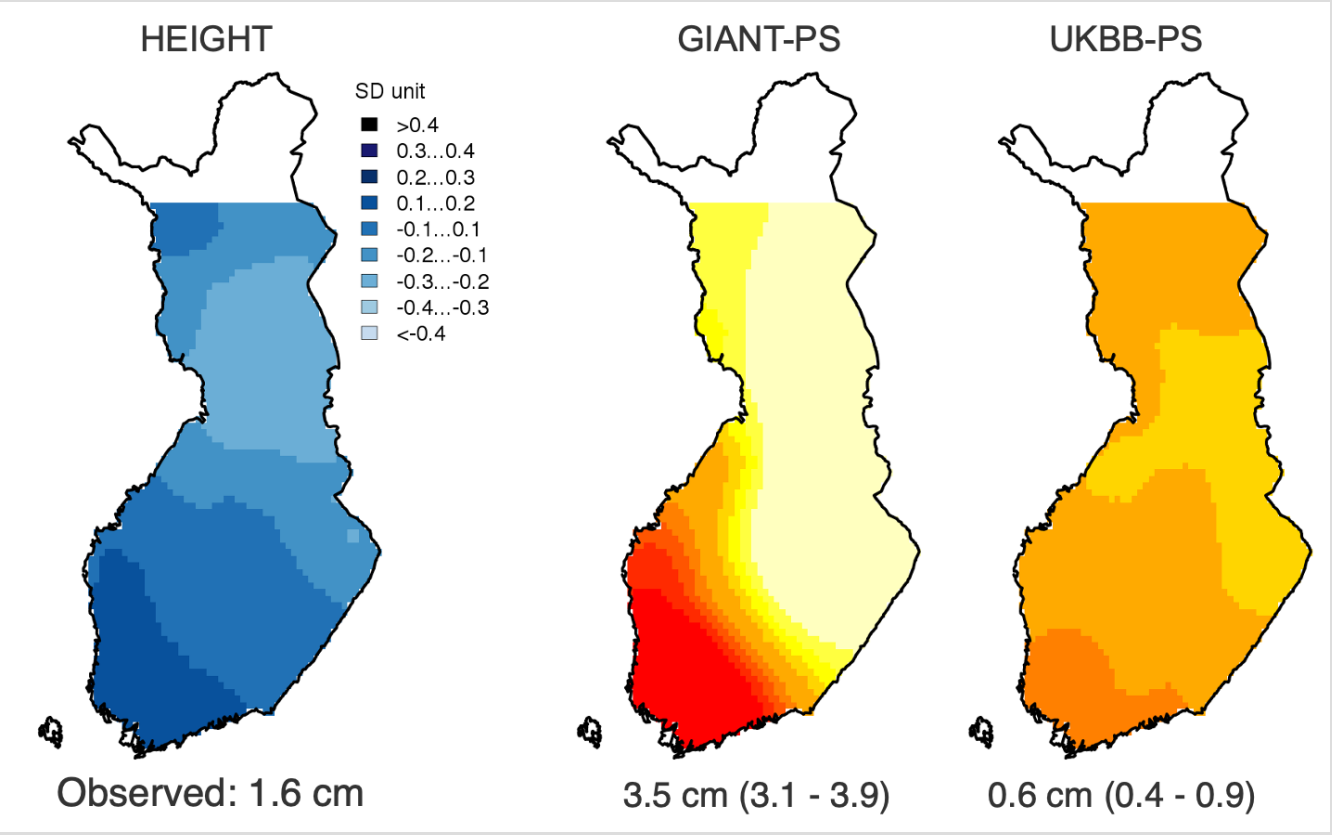
- If PGS is used to predict phenotypes of individuals who were included in the base GWAS, the prediction will be dramatically over optimistic
  - Make sure there is no overlap between GWAS and target sample
- Even if there is no overlap, relatedness and population structure can cause biases
- PGS based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

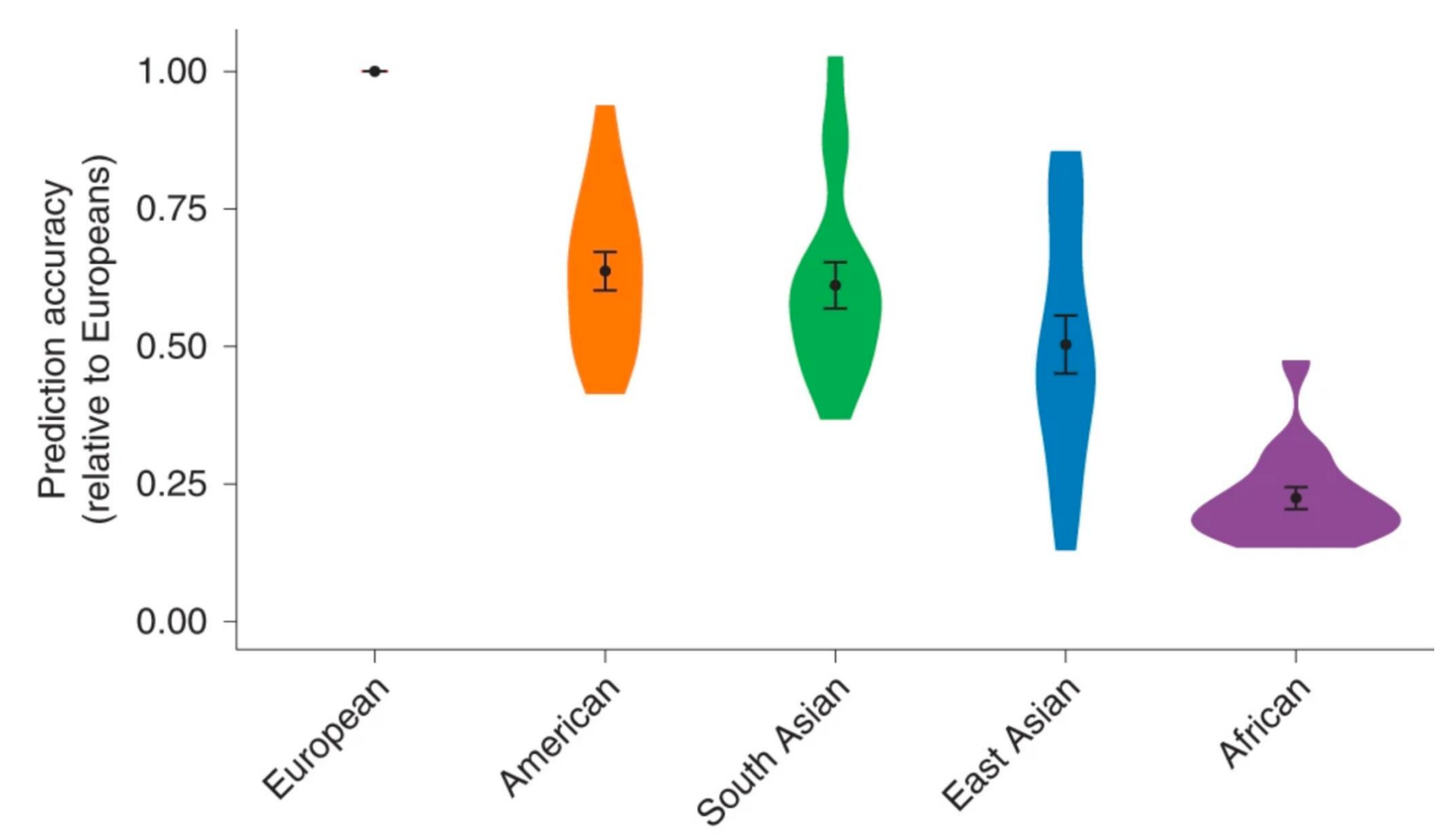
UKB GWAS
N = 337,000
No Finns

There is a 1.6cm average difference in height between SW and NE Finland.

PGS for height based on the GIANT GWAS dramatically overestimaes the SW and NE difference in height, likely because population structure and overlap of samples.

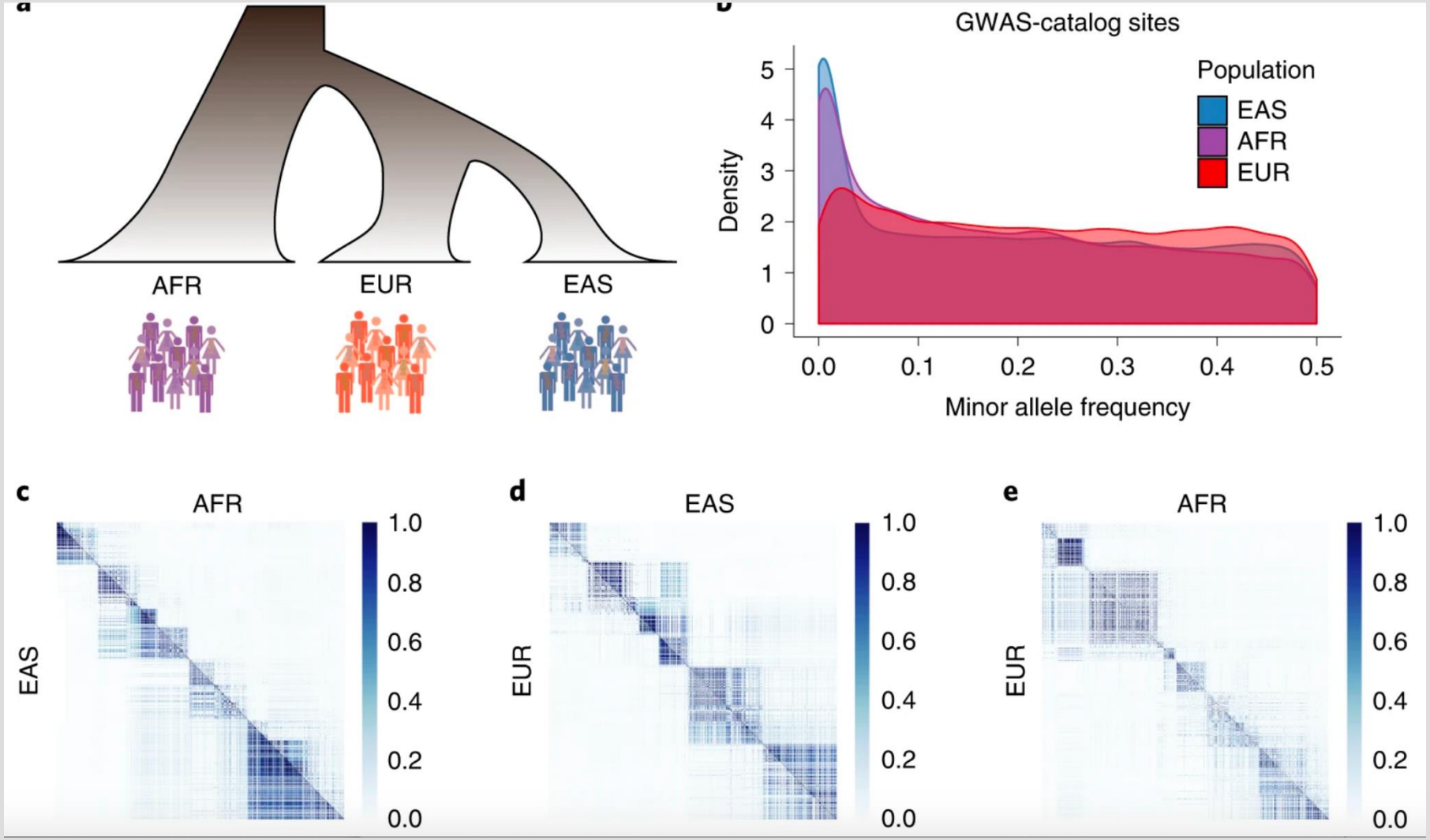
UK biobank based PGS gives a reasonable prediction.

# LACK OF TRANSFERRABILITY BTW POPULATIONS




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

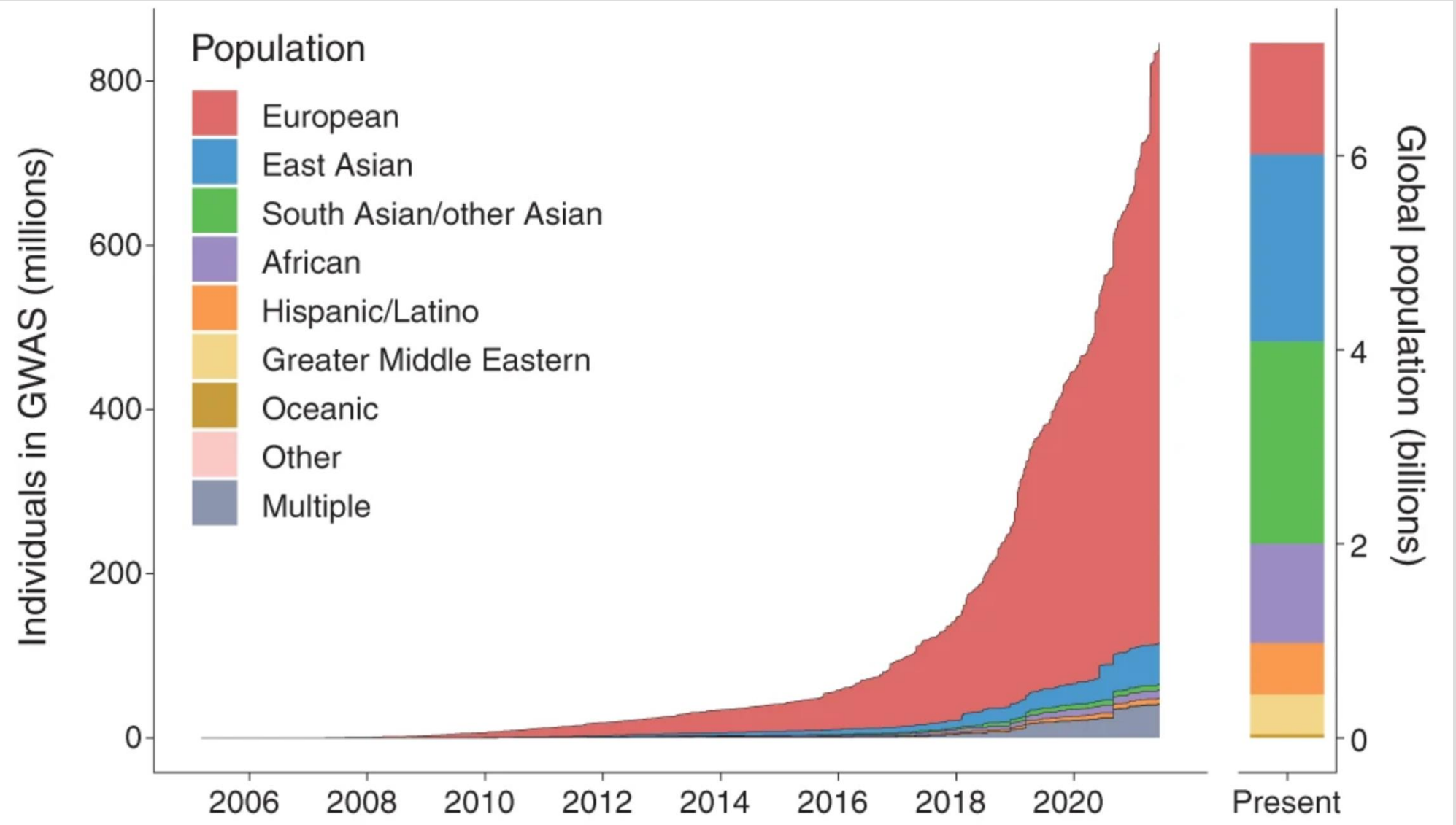
# SOME CAUSES FOR DISPARITIES




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. 2019 Nat Gen

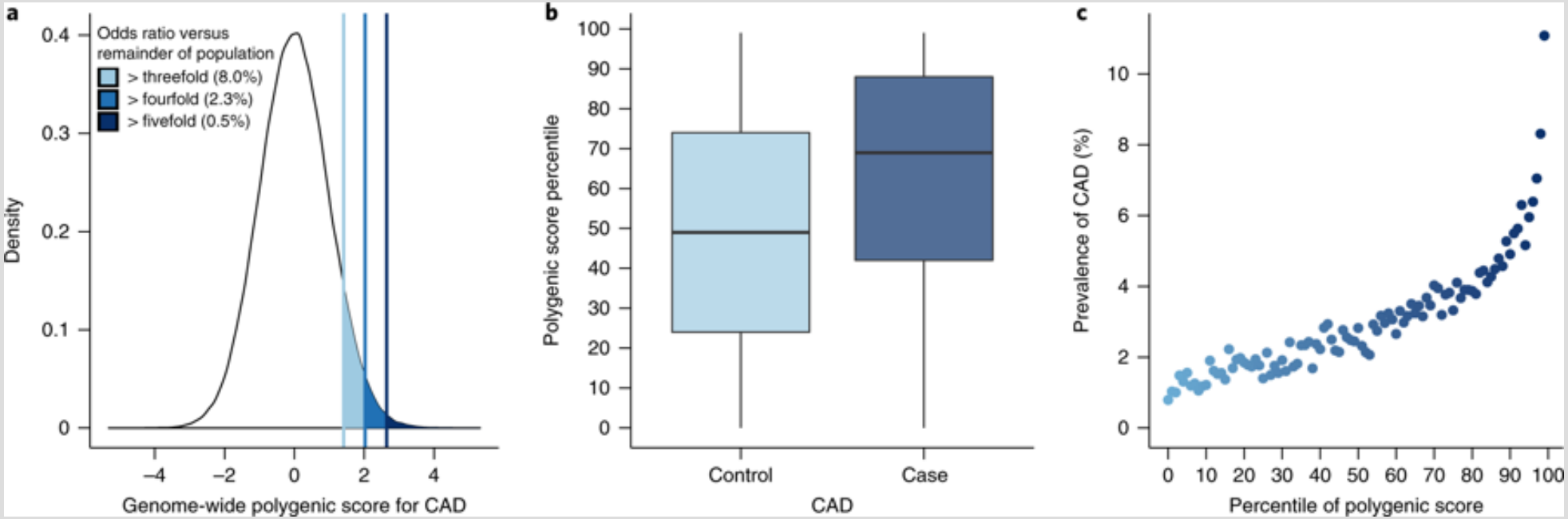
# DIVERSITY CURRENTLY LACKING IN GWAS DATA



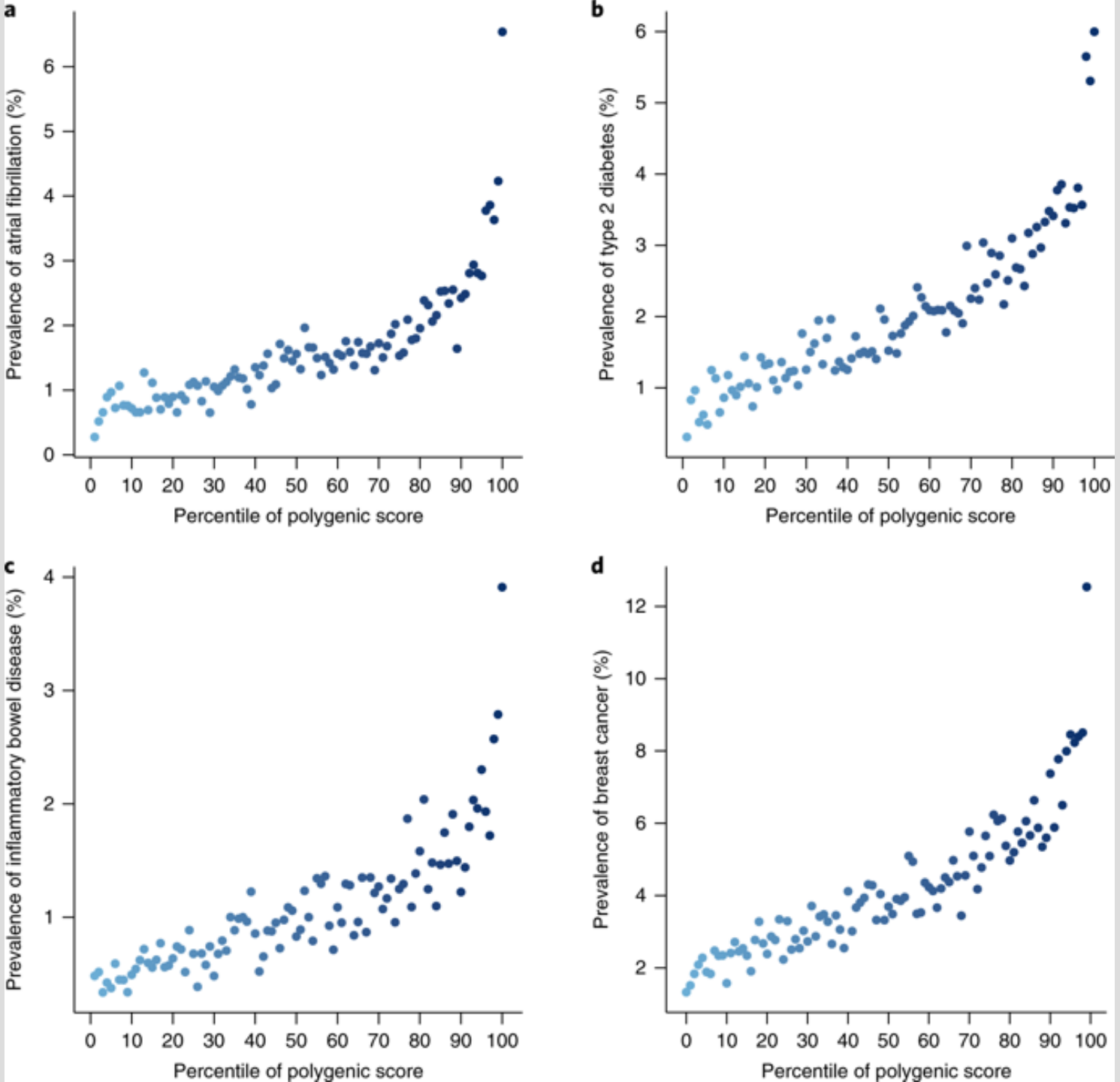

Fatumo et al.
Nat Med 2022

# 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

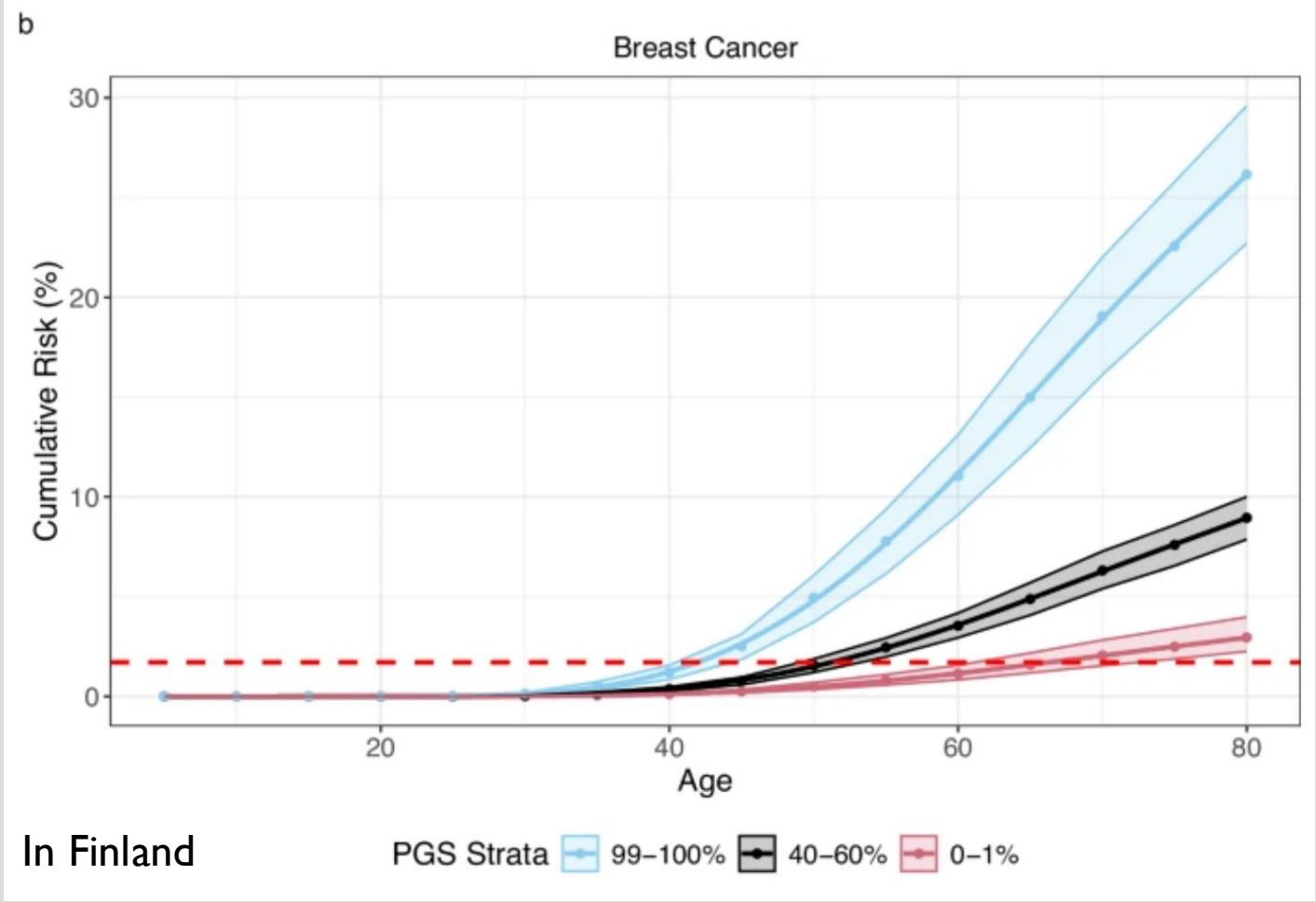



100 groups of the testing dataset were derived according to the percentile of the disease-specific PGS. **a**–**d**, Prevalence of disease displayed for the risk of atrial fibrillation (**a**), type 2 diabetes (**b**), inflammatory bowel disease (**c**), and breast cancer (**d**) according to the PGS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

Khera et al. 2018 Nat Gen

# PGS AND AGE

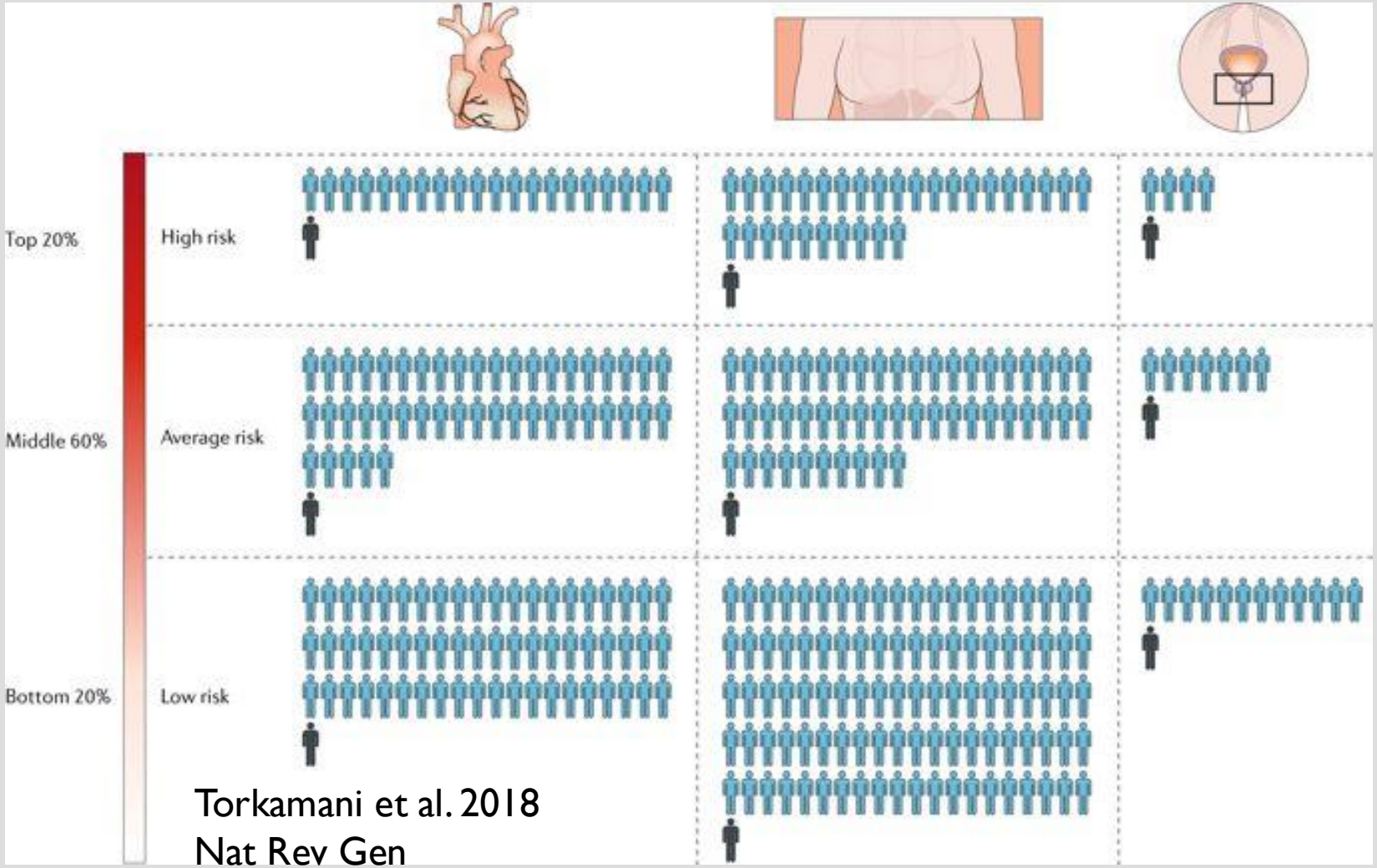
- PRS could inform screening practices for cancers and other diseases where prevention is possible




In Finland

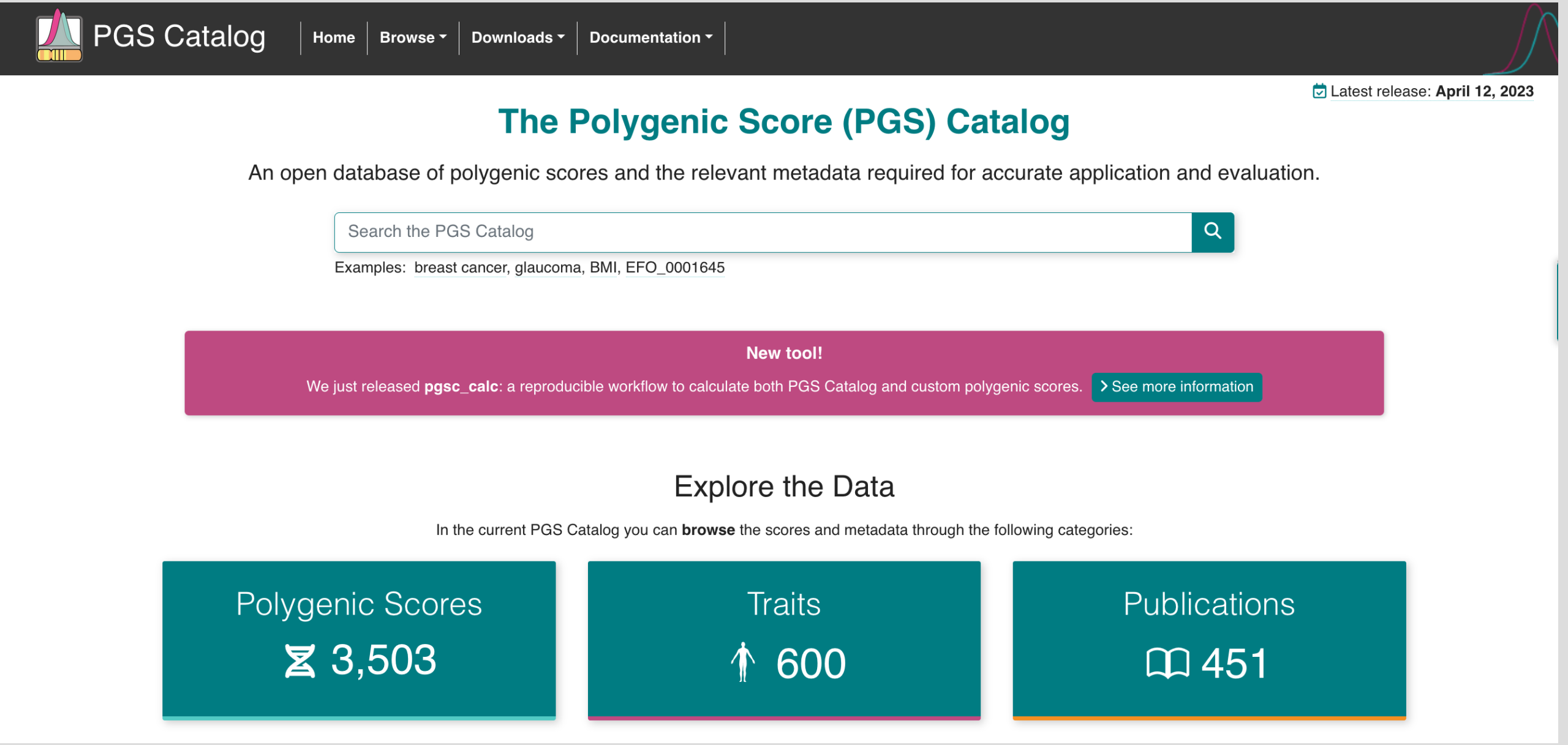
Jermy et al. 2024 Nat Commun

# 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.




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