Polygenic prediction and sparse regression analysis

Yongjin Park, UBC Path + Stat, BC Cancer

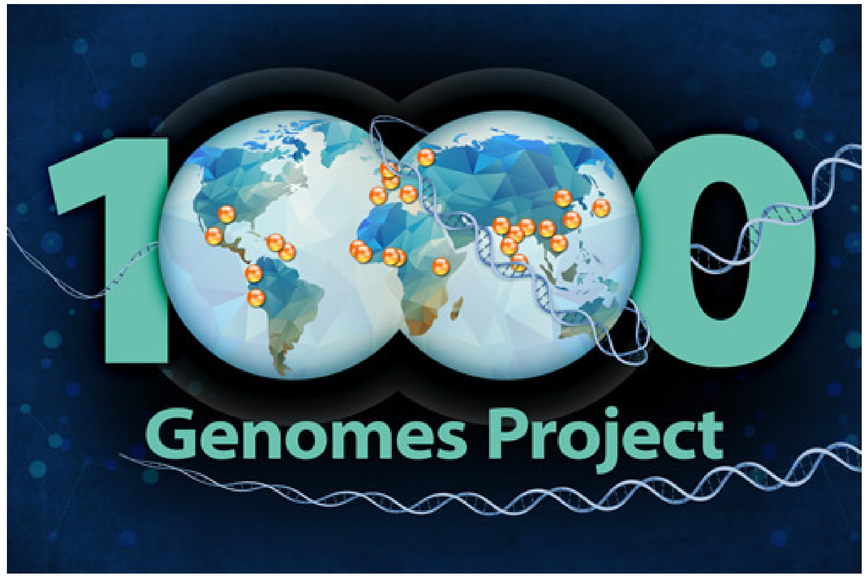
04 March 2024

Learning objective

- Gain insights into human genetics problems.
- Understand statistical challenges in a **high-dimensional** prediction problem.
- Survey of **statistical approaches** to coping with $n \ll p$ settings.

n: sample size; p: dimension (predictors)

Human genetics data – The 1000 Genomes Project



- International consortium
- Goal: find common genetic variants with frequencies of at least 1%
- The project planned to sequence each sample to 4x genomic coverage.

The 1KG data using `bigsnpr` library.

```
library(bigsnpr)
download_1000G("../data/genotype/")
```

https://www.internationalgenome.org/

How these genotypes instantiated in 1KG data

By calling `snp_readBed(.bed.file)`, we can convert the “BED”-formatted data to a “RDS” file for faster access. Later, we need to “attach” that RDS.

```
data <- snp_attach(.bk.file)
str(data, max.level = 1, strict.width = "cut")
```

```
## List of 3
##  $ genotypes:Reference class 'FBM.code256' [package "bigstatsr"] with 16 fields
##   ..and 26 methods, of which 12 are  possibly relevant
##  $ fam      :'data.frame':   2490 obs. of  6 variables:
##  $ map      :'data.frame':   1664852 obs. of  6 variables:
##  - attr(*, "class")= chr "bigSNP"
```

Definitions – a quick review of the previous lecture

Allele

- A different form of a gene
- A Greek word "allos," $\alpha\lambda\lambda\eta\lambda o$, meaning "other"

Variant and locus

- A specific region of the genome differs across two or more genomes
- A result of mutation
- A locus: a location where many variants lie (plural: loci).

Ploidy

- The number of copies of chromosomes within a cell/organism
- Haploid: one copy
- Diploid: two copies

More definitions

Biallelic variant

- bi + allelic
- Two forms for a variant
- Reference (more frequently observed) vs. alternative allele

Polymorphism

- Poly + morph
- Occurrence of different forms

SNP

- Single Nucleotide Polymorphism
- A place in the genome where people differ by a single base pair

We pronounce SNP “snip” in North America.

Genetic variants cheat sheet

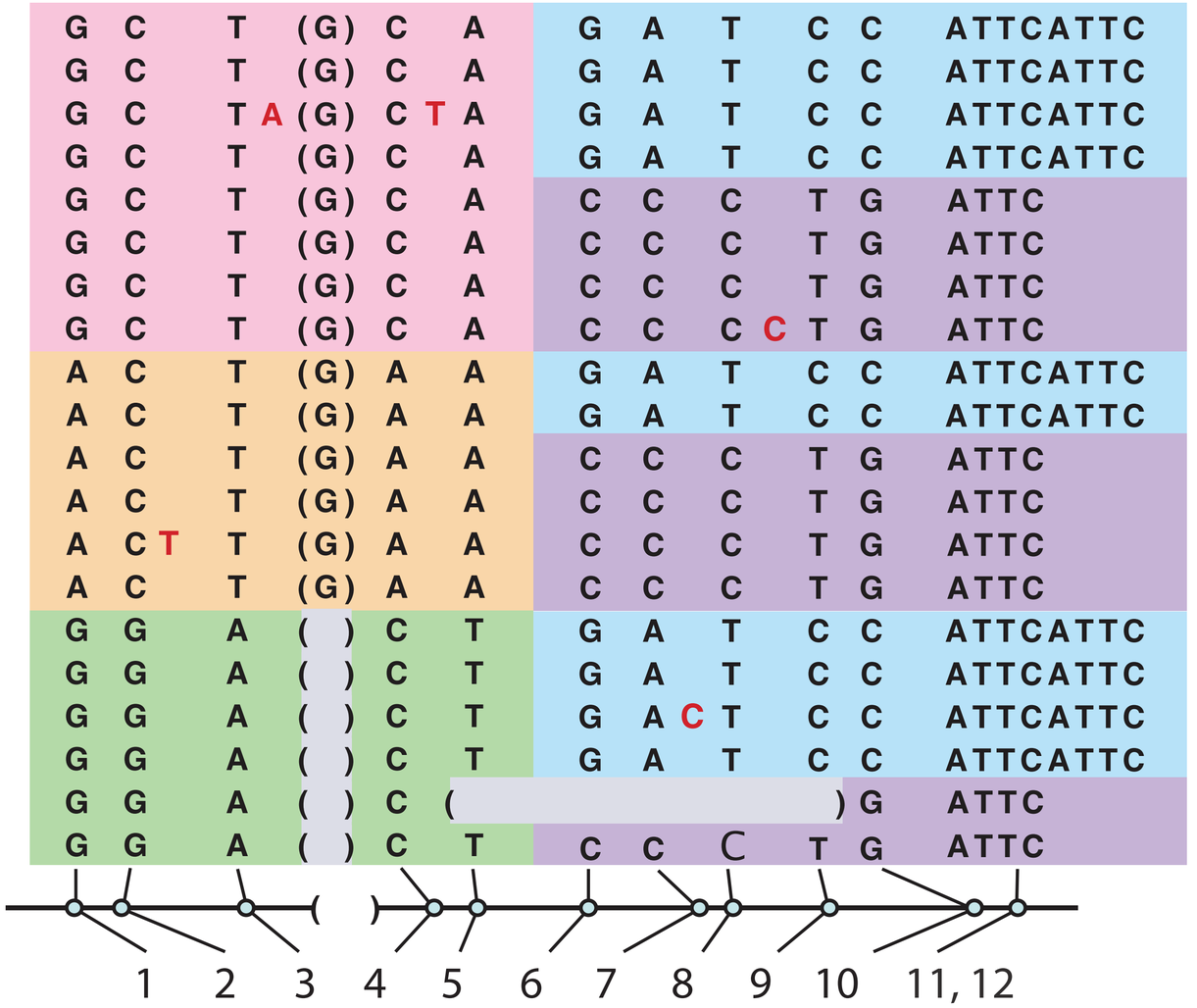


Image: Altshuler, Daly, Lander (Science)

Genetic variants cheat sheet

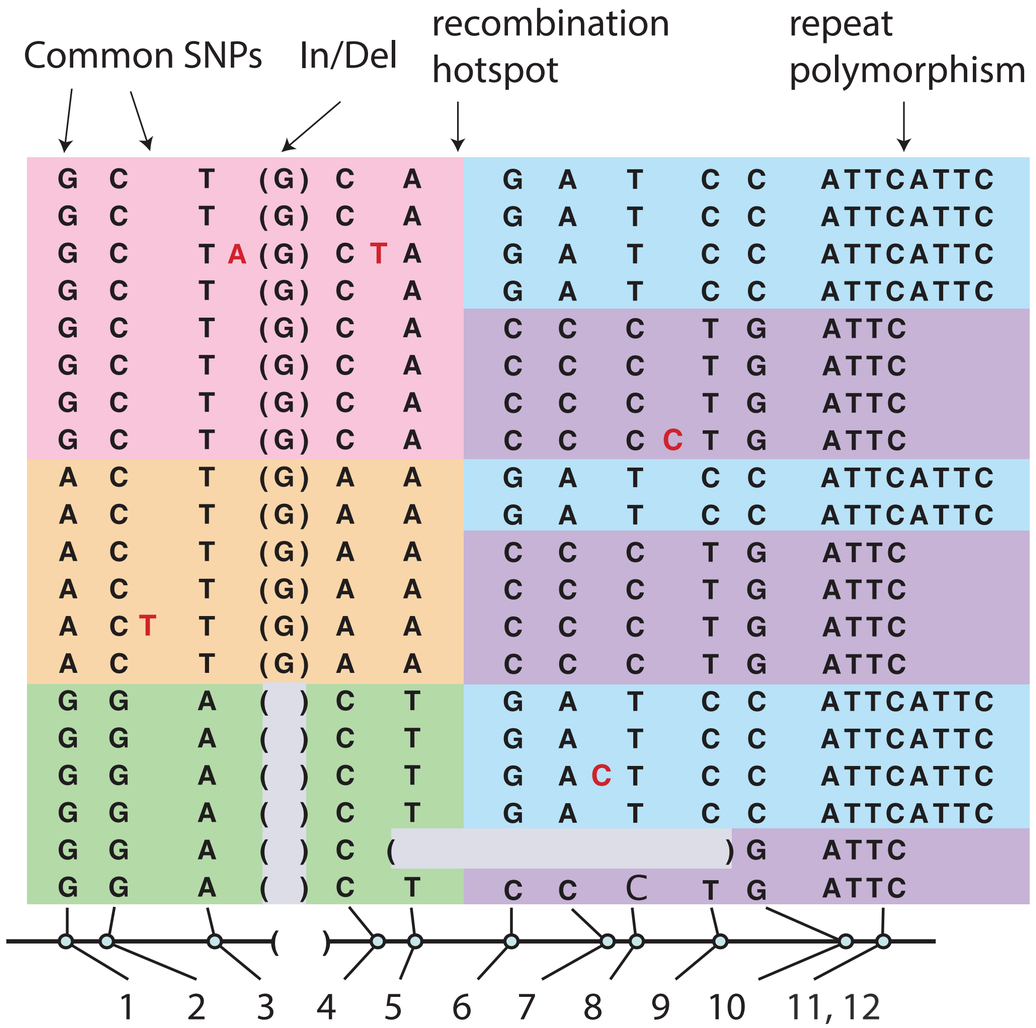
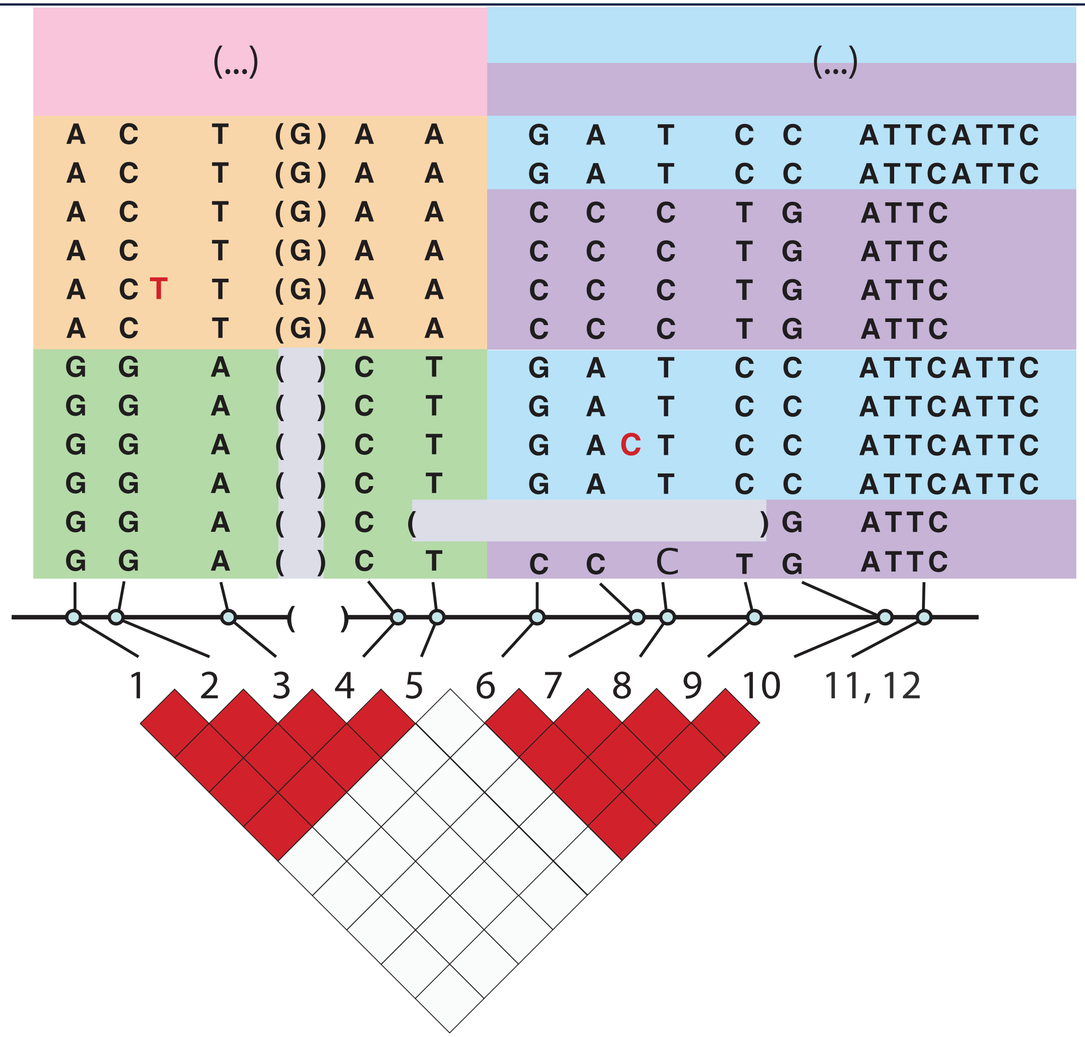


Image: Altshuler, Daly, Lander (Science)

Genetic variants cheat sheet

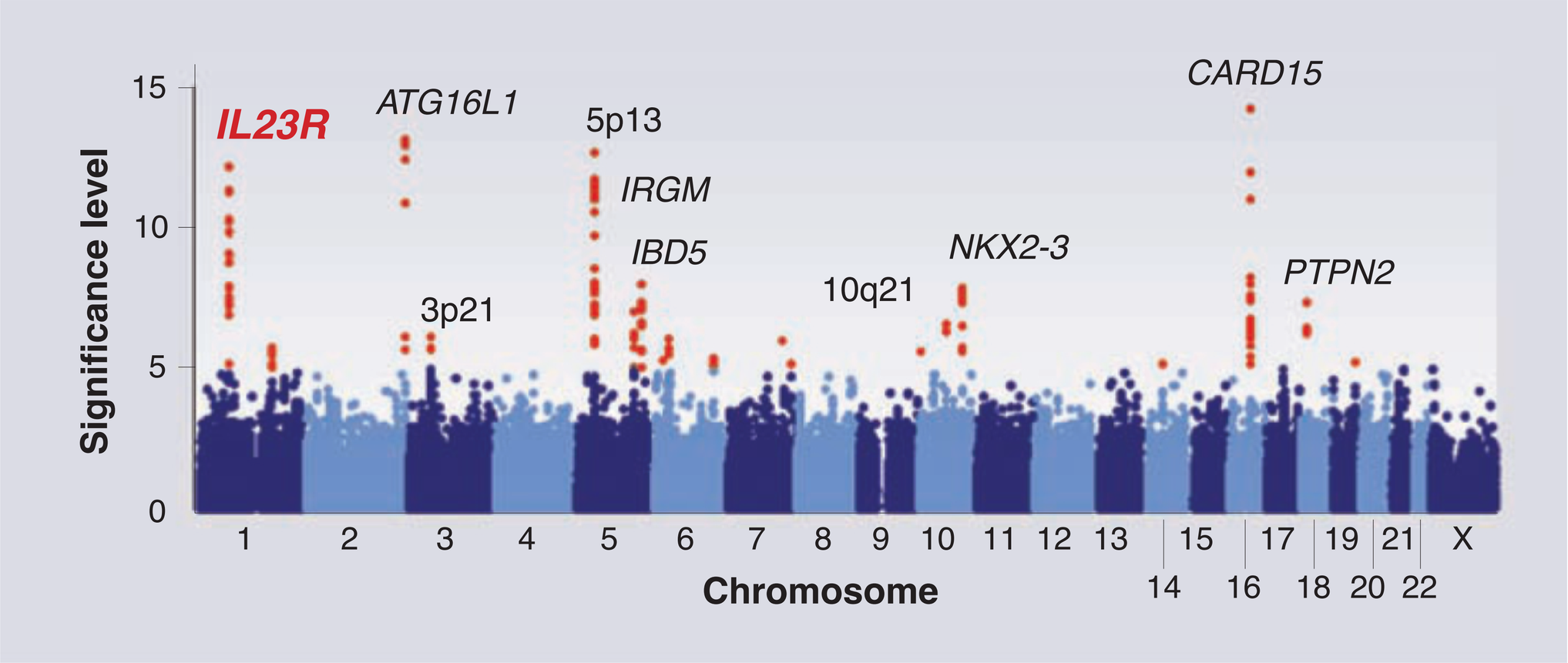


Linkage disequilibrium ≈ correlation block

Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. What are the statistical challenges in PGS estimation?
4. Statistical fine-mapping to handle LD structures
5. Other topics

GWAS (previous lecture), *then*, what's next?



GWAS (previous lecture), *then*, what's next?

- GWAS until 2010s: heavy focuses on **mapping**
  - GWAS map: genetic variants $\longrightarrow$ a phenotype
  - Stringent genome-wide p-value cutoff
  - Study design, meta analysis
- NHGRI-EBI GWAS Catalog:
  `https://www.ebi.ac.uk/gwas/`

Let’s take a look at the GWAS Catalog:

`https://www.ebi.ac.uk/gwas/`

GWAS (previous lecture), then, *what's next?*

- GWAS since 2010s: more emphases on **prediction**
  - Can we turn GWAS results to a prediction model?
  - Can we understand the mechanisms?
  - Machine learning, data integration, causal inference

Polygenic score models predict disease prevalence based on genome

- Poly + genic = many genes

A (linear) polygenic score (PGS)

$$Y_i = \sum_j X_{ij}\beta_j$$

for an individual i.

Polygenic score models predict disease prevalence based on genome

- Poly + genic = many genes

A (linear) polygenic score (PGS)

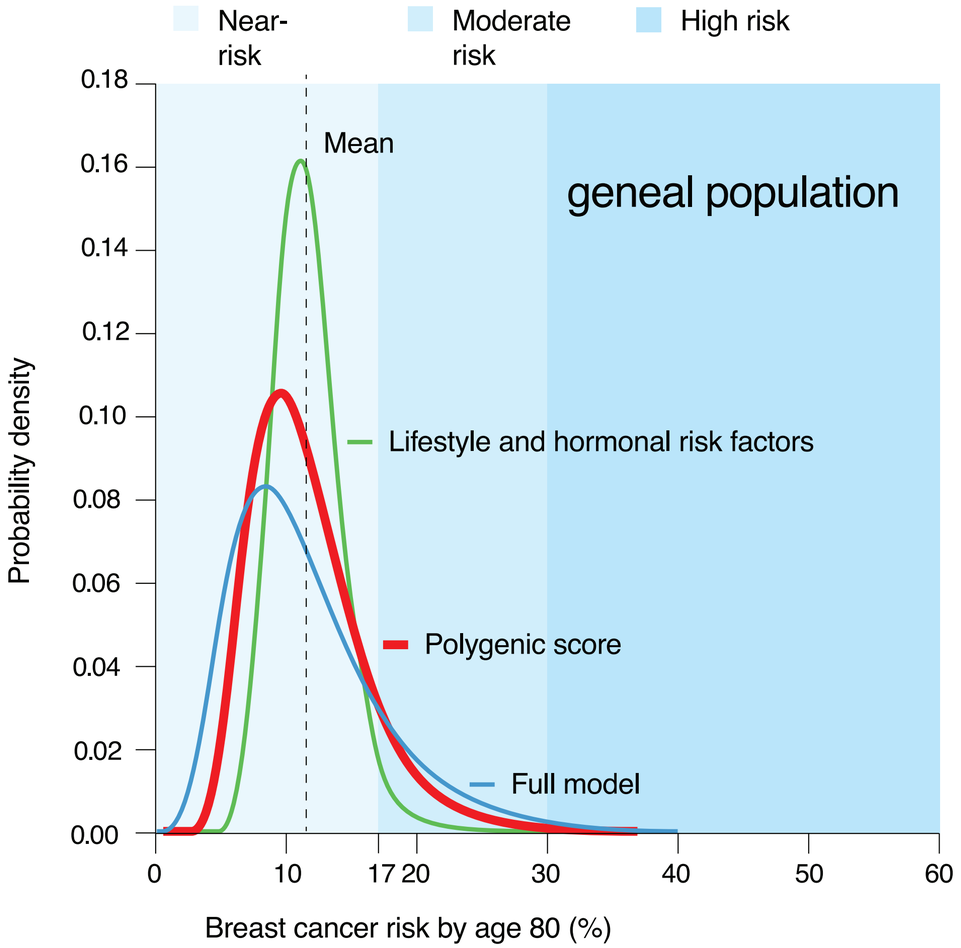
$$Y_i = \sum_j X_{ij}\beta_j$$

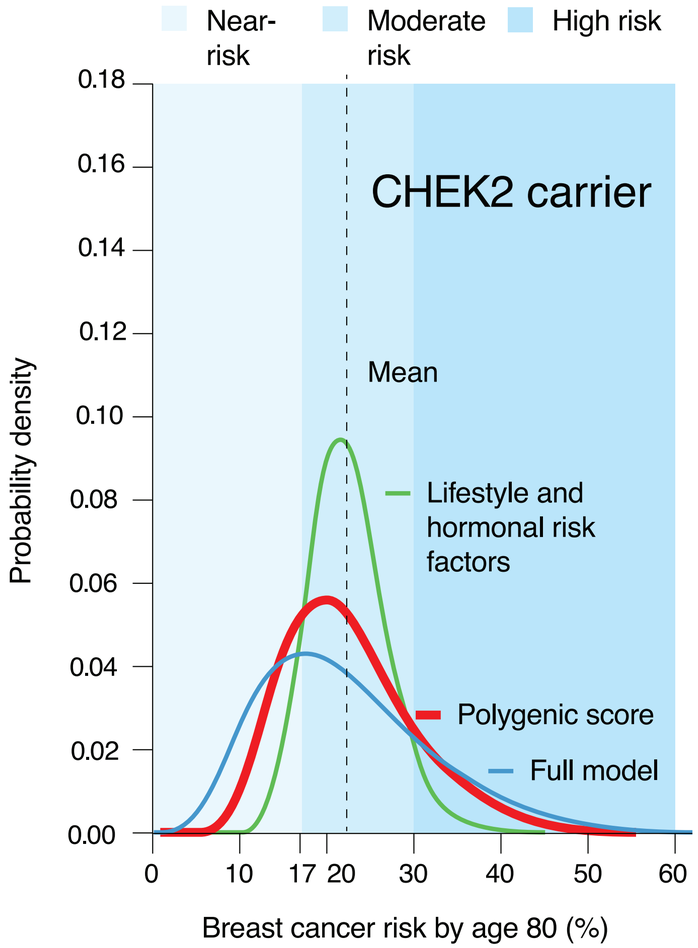
for an individual i.

Knowing β's, we can predict:

$$Y^\star{}_i \leftarrow \sum_j X^\star_{ij}\hat{\beta}_j$$

Example: PGS for breast cancer occurrence

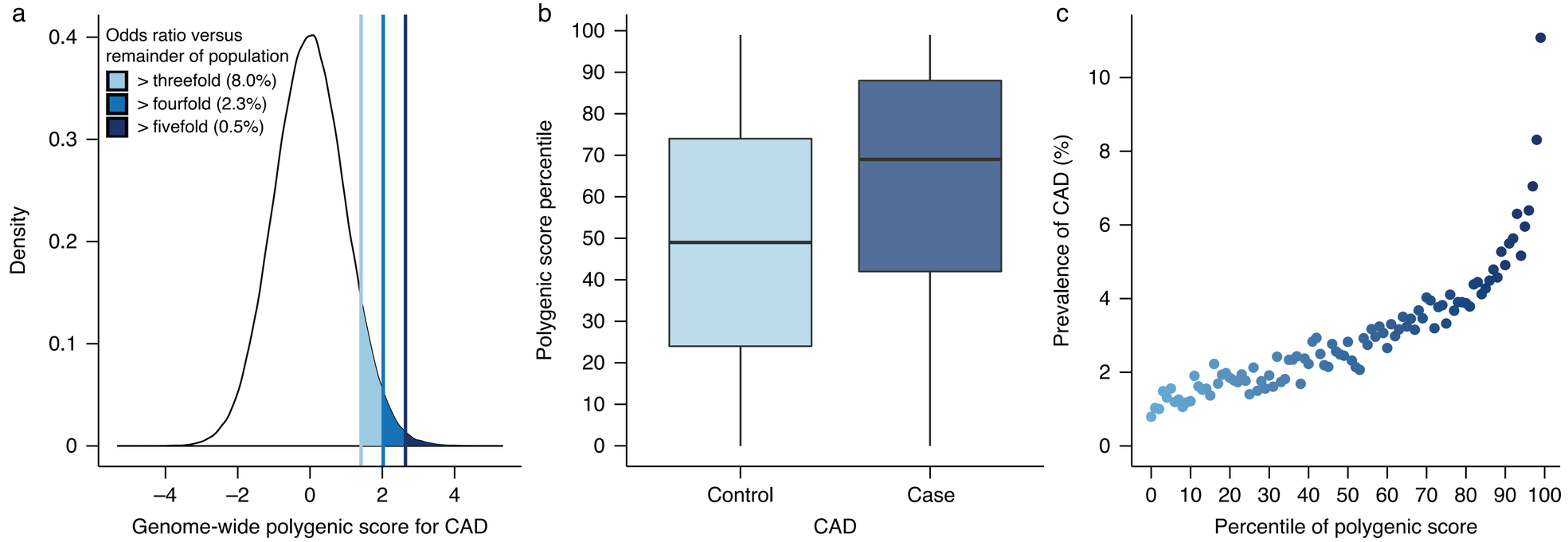




- Strong heritability (proportion of disease risk variation explained by genetics): 35% - 80%
- Prediction more accurate than life style and other environmental factors
- More powerful if combined with rare risk factors

Yang .. Pharoah, Nature Review Genetics (2023)

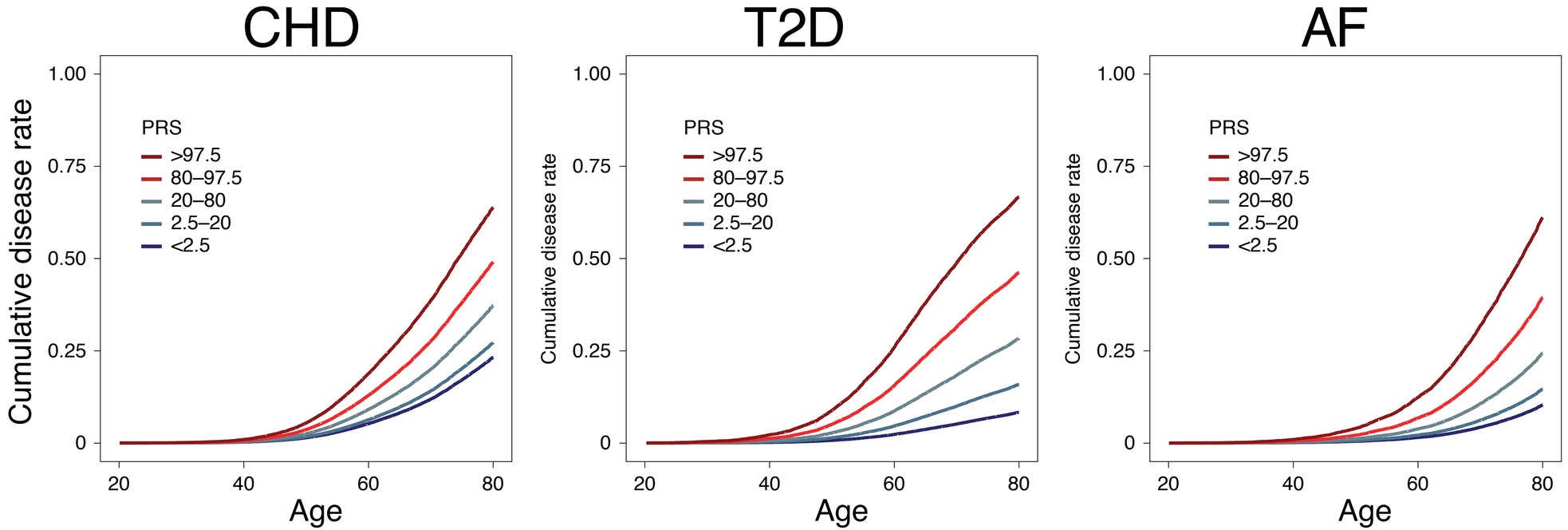
Example: PGS for coronary artery disorder $\rightarrow$ prevention



- Top .5% of PGS values $\rightarrow$ five fold increase of CAD

Khera .. Kathiresan, Nature Genetics (2018)

Example: PGS stratifies individuals' disease onset and risk



- PRS $\approx$ PGS.
- We can partition cohorts based on PGS profiles.

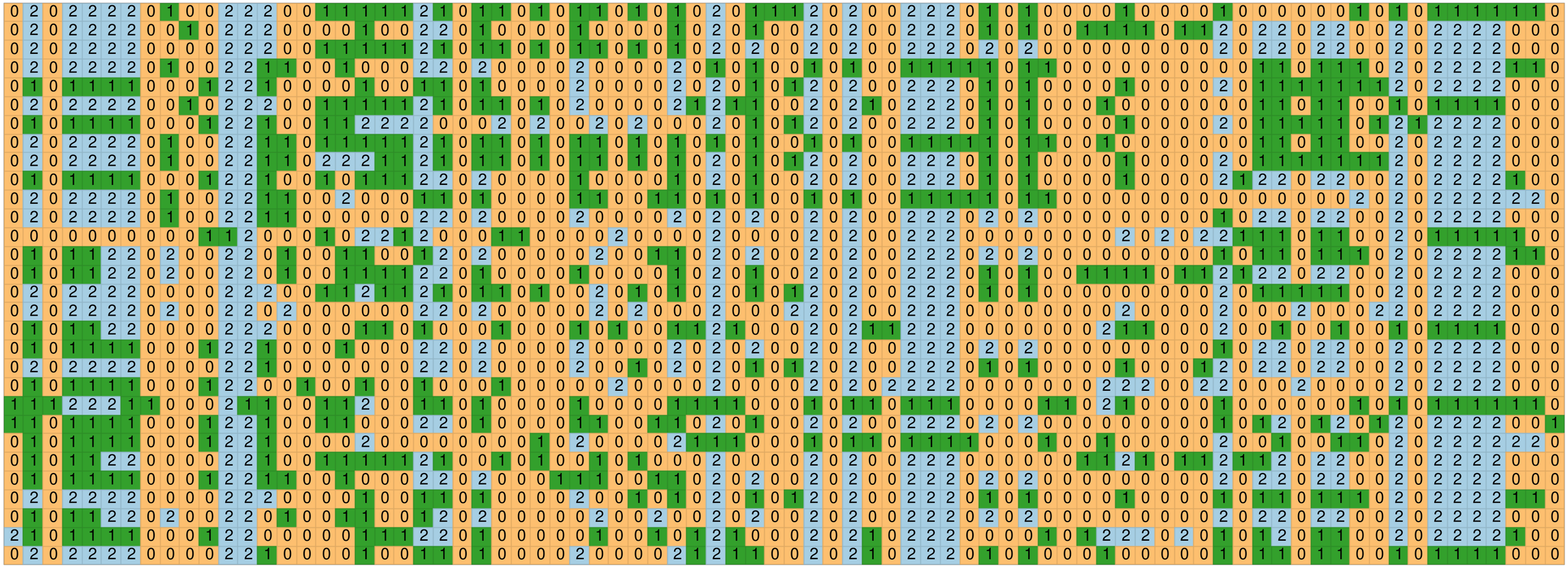
Mars .. Ripatti, Nature Medicine (2020)

Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. What are the statistical challenges in PGS estimation?
4. Statistical fine-mapping to handle LD structures
5. Other topics

A toy example: GWAS for “Obsessive `ggplot` Disorder”

A genotype matrix X ($X_{ij} \in \{0, 1, 2\}$):



[1] just an illustration purpose

In GWAS, we can have case vs. control phenotype

$Y_i = 1$ if case vs. $Y_i = 0$ if control

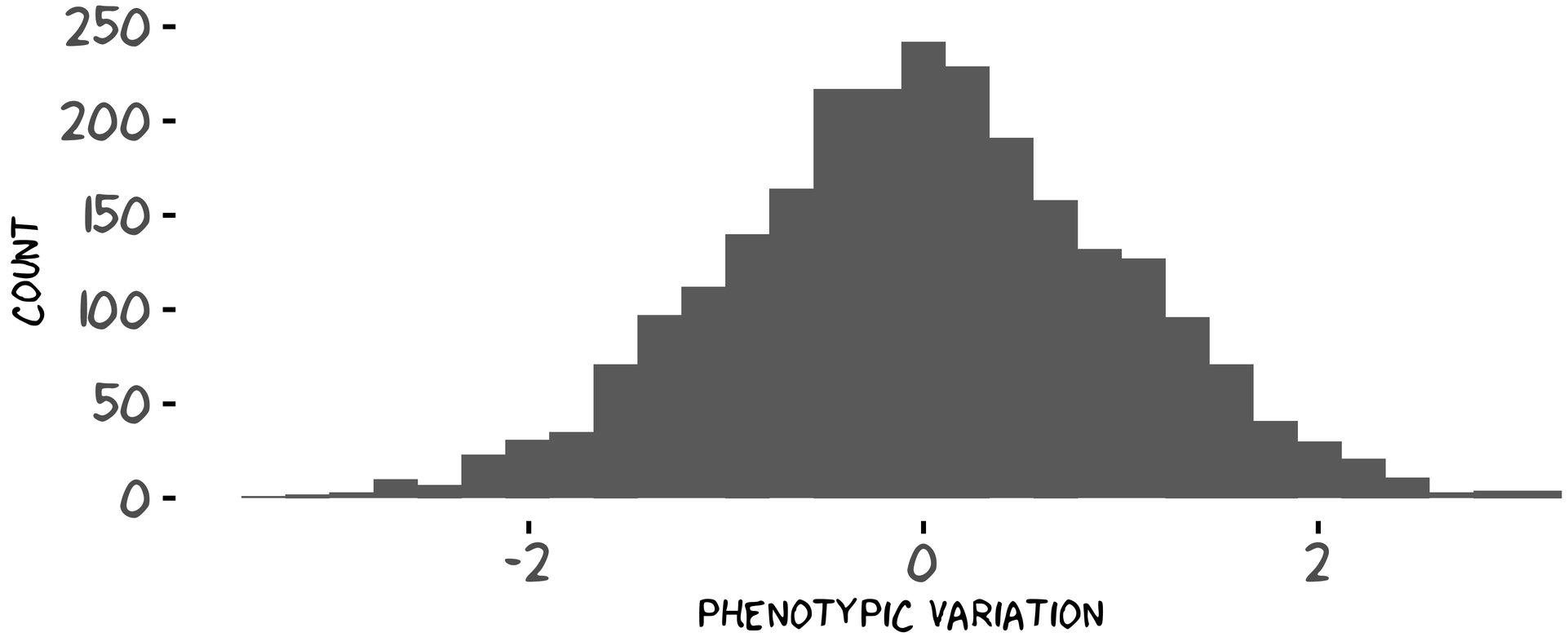
For our *OGD* GWAS:

```
table(y)
```

```
## y
##   0   1
## 731 769
```

- E.g., Cancer vs. non-cancer, schizophrenia vs. no mental disorder
- The fundamental question is, “Which individuals/subjects can be truly labelled control/wild type?”

The under the hood, there are quantitative “risk” scores



- E.g., height, BMI, parents’ age of death, how man pack of cigarettes, etc.

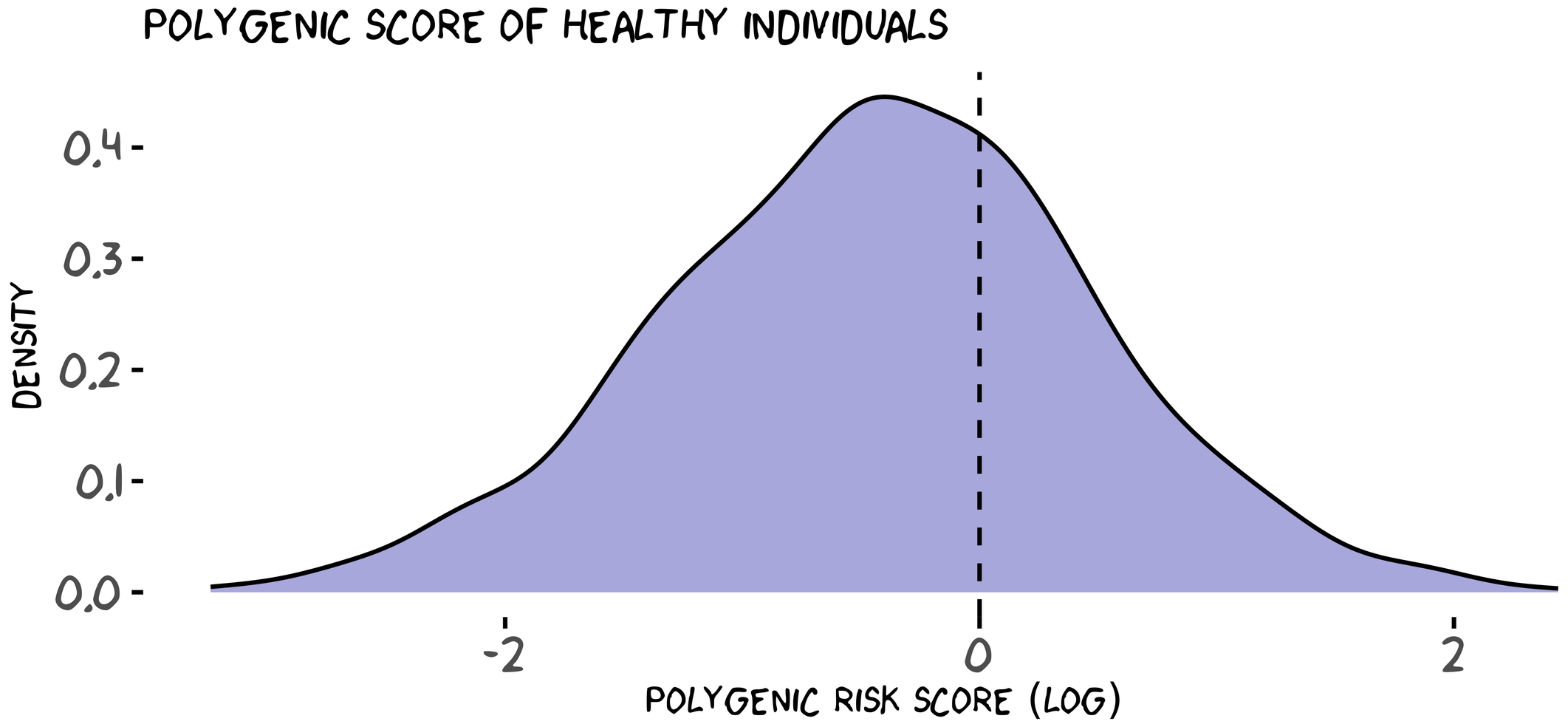
Side note: Almost all complex traits are polygenic

Polygenic effects of common SNPs $\longrightarrow$ polymorphism in phenotypes

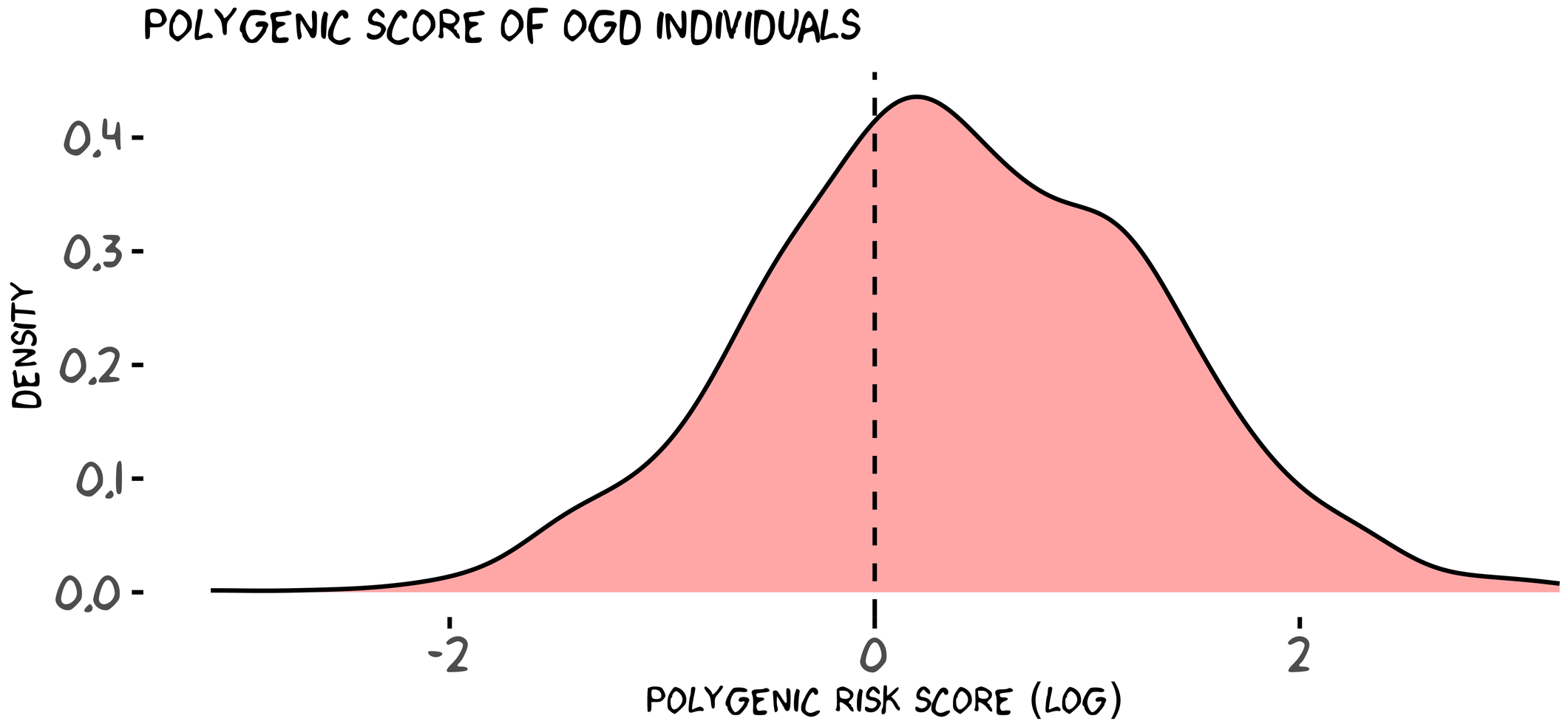
> *A polygenic trait is a characteristic, such as height or skin color, that is influenced by two or more genes.* ***Because multiple genes are involved, polygenic traits do not follow the patterns of Mendelian inheritance.*** *Many polygenic traits are also influenced by the environment and are called multifactorial.*

https://www.genome.gov/genetics-glossary/Polygenic-Trait

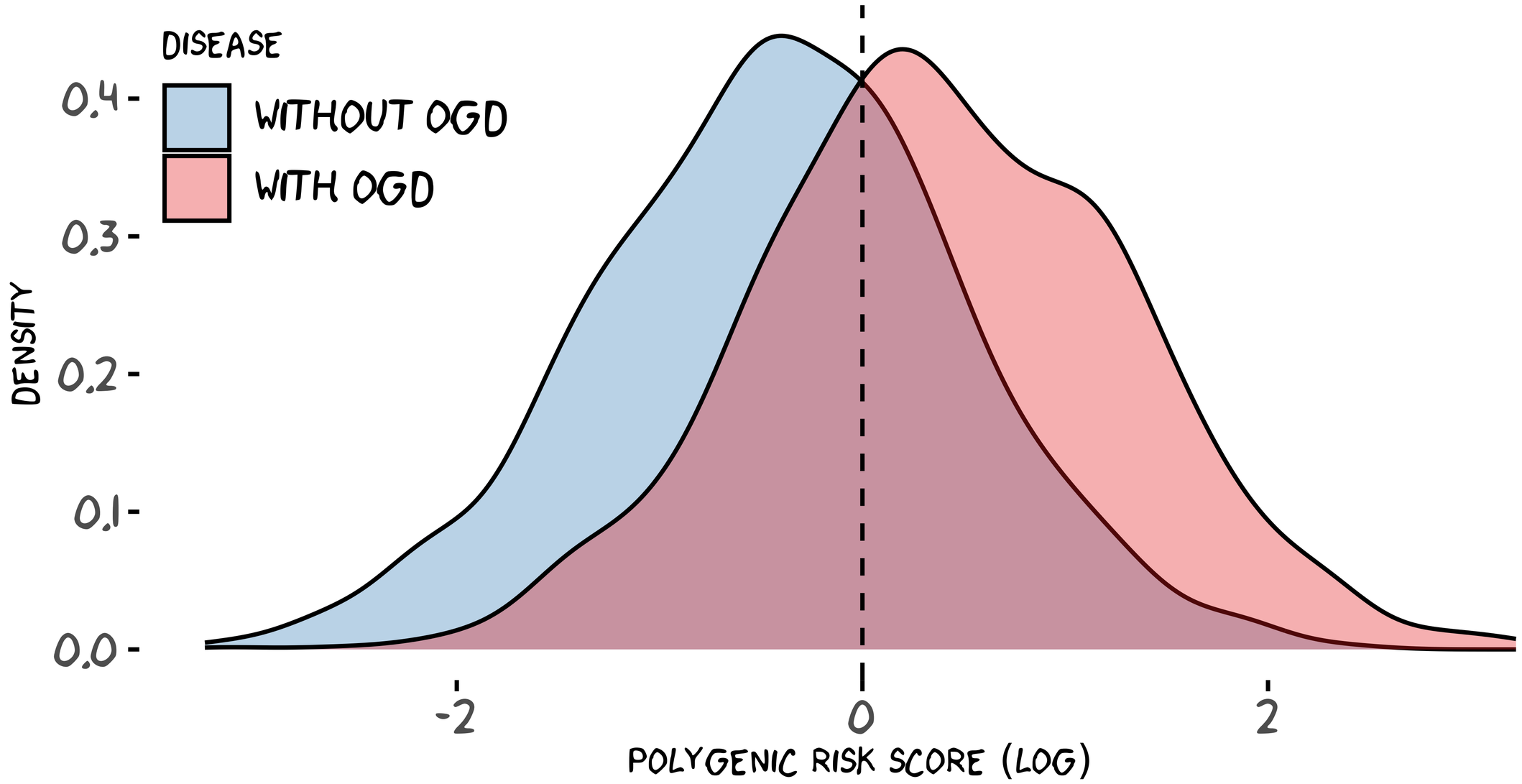
A polygenic score to predict disease prevalence



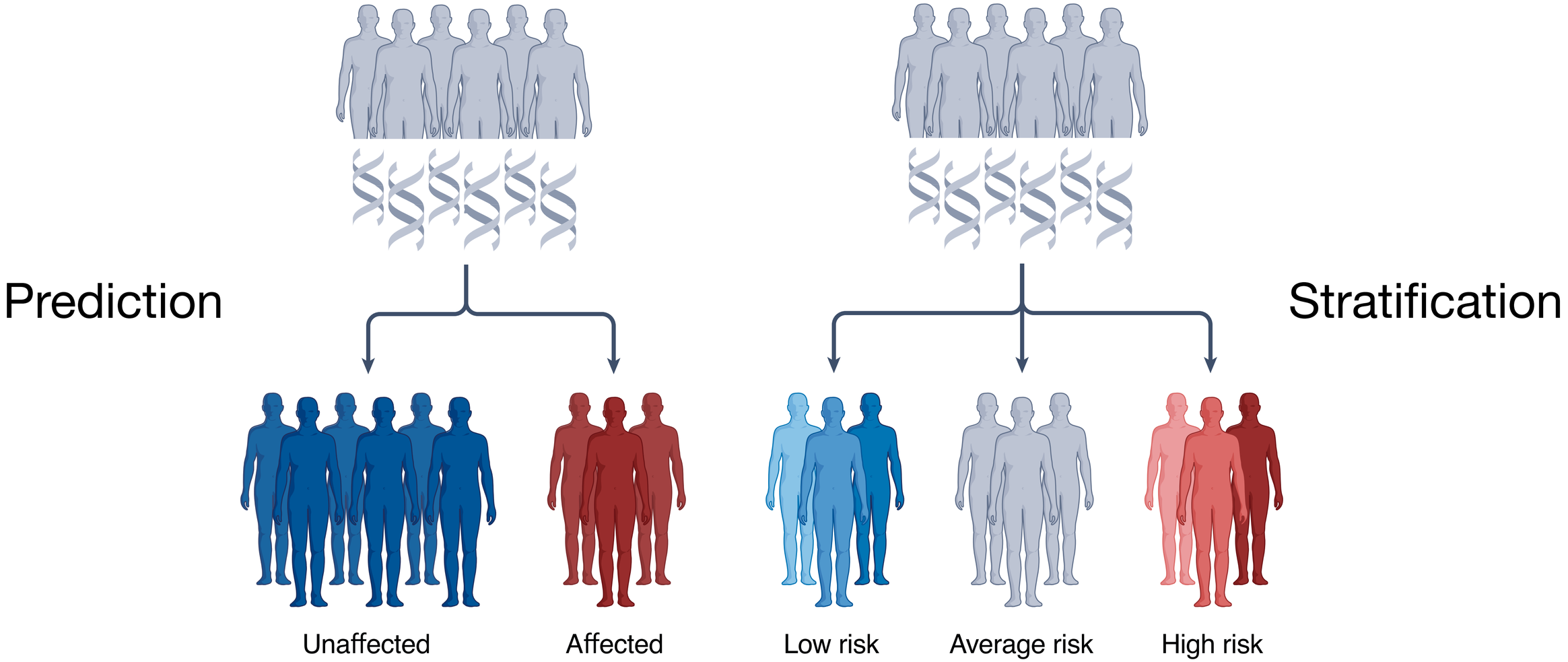
A polygenic score to predict disease prevalence



A polygenic score to predict disease prevalence



Risk case-control vs. stratification



Kachuri et al. Nature Review Genetics (2023)

Polygenic score $\propto$ the odds of the case vs. control in GWAS

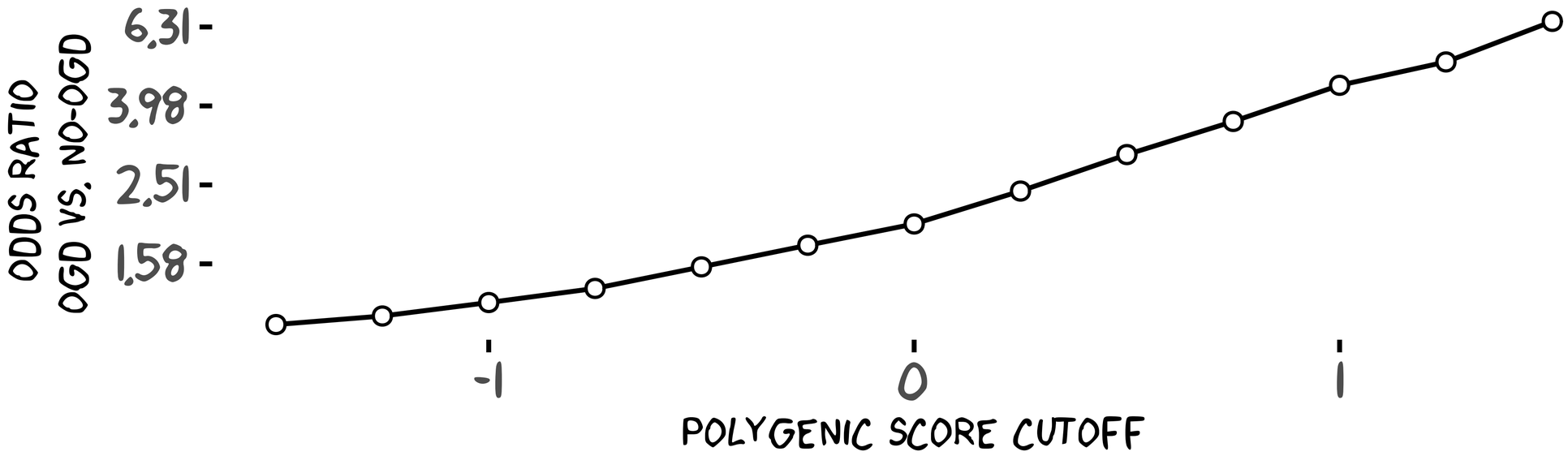
Log-odds ratio:

$$g(y) = \log \frac{p(\text{disease} | Y > y)}{p(\text{no disease} | Y > y)}$$

Polygenic score $\propto$ the odds of the case vs. control in GWAS

Log-odds ratio:

$$g(y) = \log \frac{p(\text{disease} | Y > y)}{p(\text{no disease} | Y > y)}$$



A PGS model to explain disease prevalence with genetics

Log-odds ratio:

$$g(y) = \log \frac{p(\text{disease} | Y > y)}{p(\text{no disease} | Y > y)}$$

Goal: Estimate a function to predict this log-odds ratio.

$$g(y_i) \sim \beta_1 X_{i1} + \beta_2 X_{i2} \cdots + \epsilon$$

A PGS model to explain disease prevalence with genetics

Log-odds ratio:

$$g(y) = \log \frac{p(\text{disease} | Y > y)}{p(\text{no disease} | Y > y)}$$

Goal: Estimate a function to predict this log-odds ratio.

$$g(y_i) \sim \underbrace{\beta_1 X_{i1} + \beta_2 X_{i2} \cdots}_{\textbf{genetic effects}} + \epsilon$$

PGS estimation is a linear modelling with $p \gg n$

$$Y_i \sim X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots$$

PGS estimation is a linear modelling with $p \gg n$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \sim \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} \beta_1 + \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{n2} \end{pmatrix} \beta_2 + \cdots$$

PGS estimation is a linear modelling with $p \gg n$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \sim \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} \beta_1 + \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{n2} \end{pmatrix} \beta_2 + \cdots + \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix} \beta_p$$

PGS estimation is a linear modelling with $p \gg n$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \sim \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} \beta_1 + \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{n2} \end{pmatrix} \beta_2 + \cdots + \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix} \beta_p$$

$$\mathbf{y} \sim \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_p\beta_p$$

$$p \gg n$$

PGS estimation is a linear modelling with $p \gg n$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \sim \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} \beta_1 + \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{n2} \end{pmatrix} \beta_2 + \cdots + \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix} \beta_p$$

$$\mathbf{y} \sim \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_p\beta_p$$

$$p \gg n$$

$n \approx 10^4$ but $p \approx 10^6$ for many large-scale GWAS

$p \gg n$: Why can't we just fit a model?

```
## lm.out <- lm(Y ~ X)
```

- $p \gg n$: need to estimate $\beta_1, \ldots, \beta_p$

$p \gg n$: Why can't we just fit a model?

```
## lm.out <- lm(Y ~ X)
```

- $p \gg n$: need to estimate $\beta_1, \ldots, \beta_p$
- How many samples? How many unknowns?

$p \gg n$: Why can't we just fit a model?

```
## lm.out <- lm(Y ~ X)
```

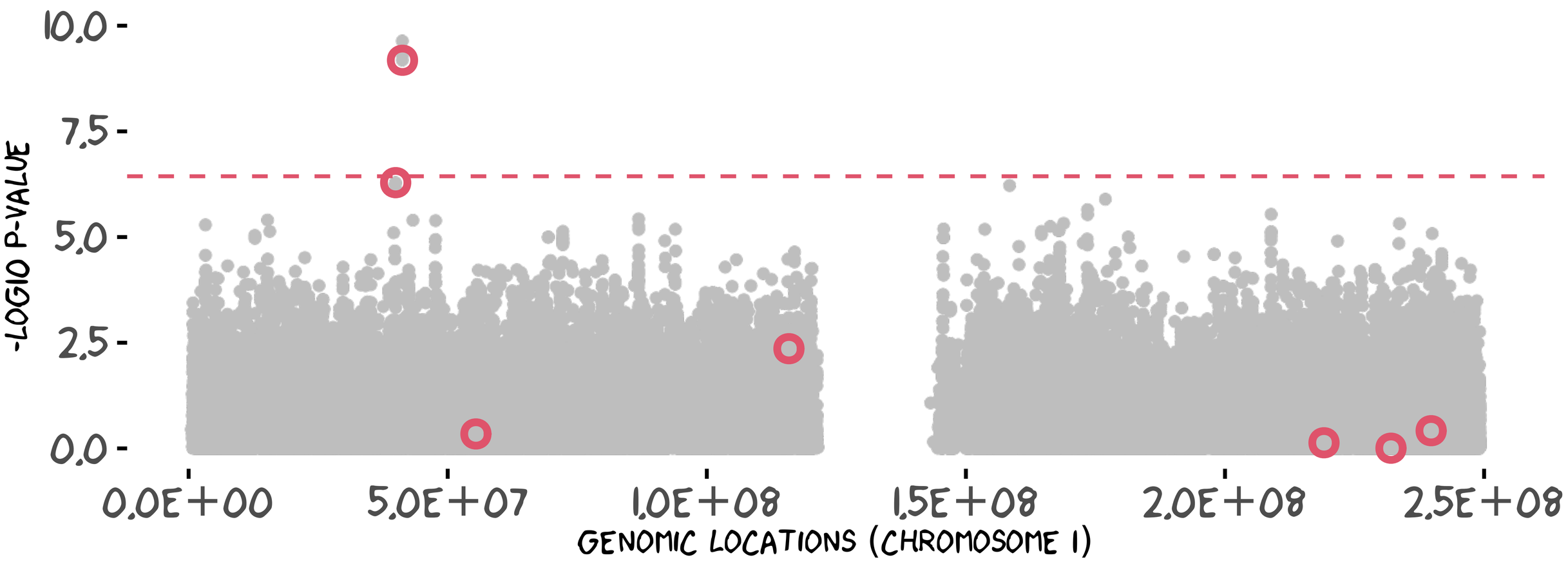
- $p \gg n$: need to estimate $\beta_1, \ldots, \beta_p$
- How many samples? How many unknowns?
- Is it computationally feasible?

$p \gg n$: Variant-by-variant GWAS is not a bad idea

```
.gwas <- col_t_welch(X[Y == 0, ], X[Y == 1, ])  # univar.
```

$p \gg n$: Variant-by-variant GWAS is not a bad idea

```
.gwas <- col_t_welch(X[Y == 0, ], X[Y == 1, ])  # univar.
```



A linear model including only these “GWAS” variants

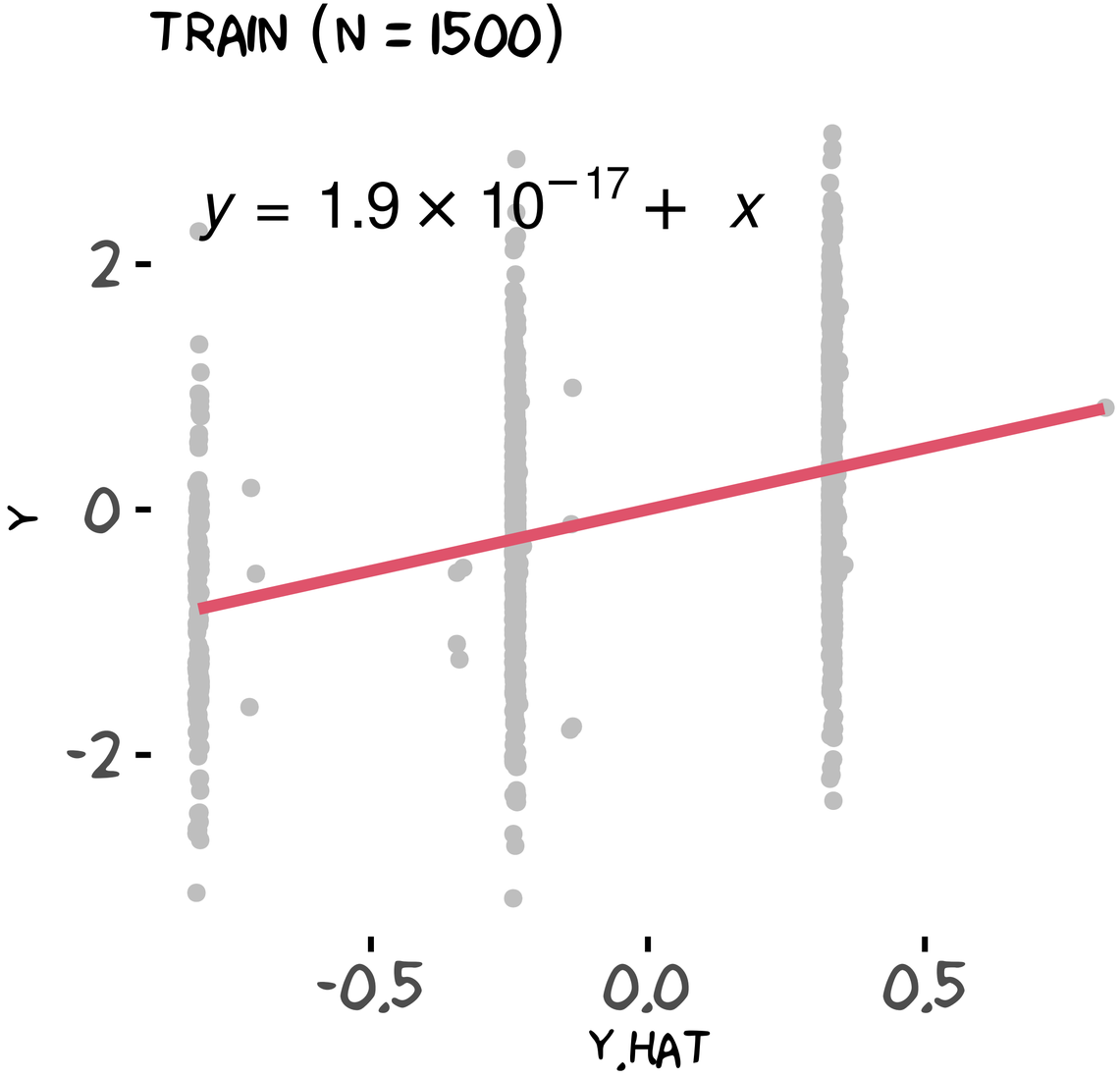
```
X.gwas <- X[, p.adjust(.gwas$pvalue) < 0.05]
head(X.gwas, 4)
```

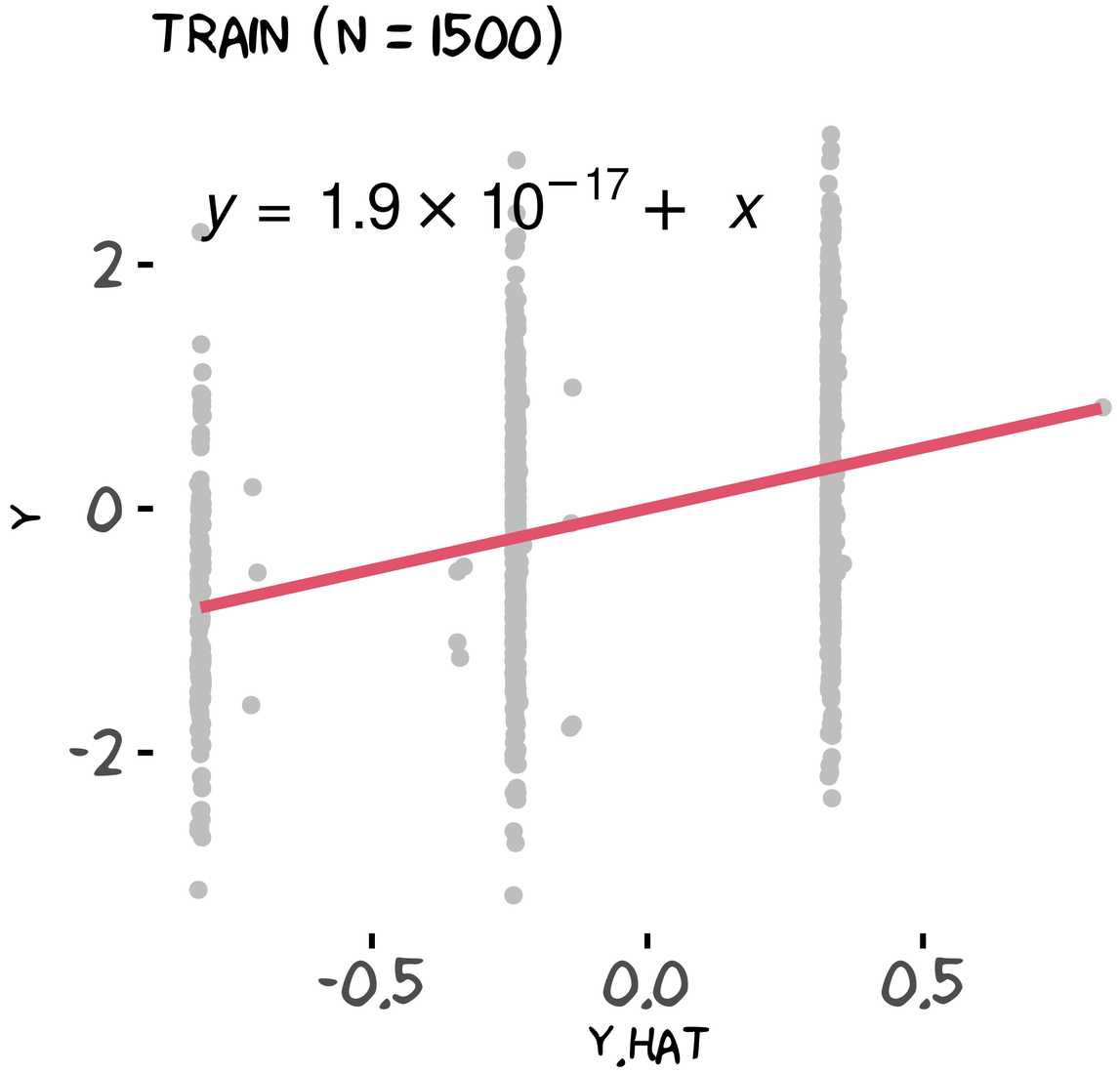
```
##         41270163 41270937 41272533 41274422
## HG00159        0        0        0        0
## HG01495        1        1        1        1
## HG01519        2        2        2        2
## HG03643        1        1        1        1
```

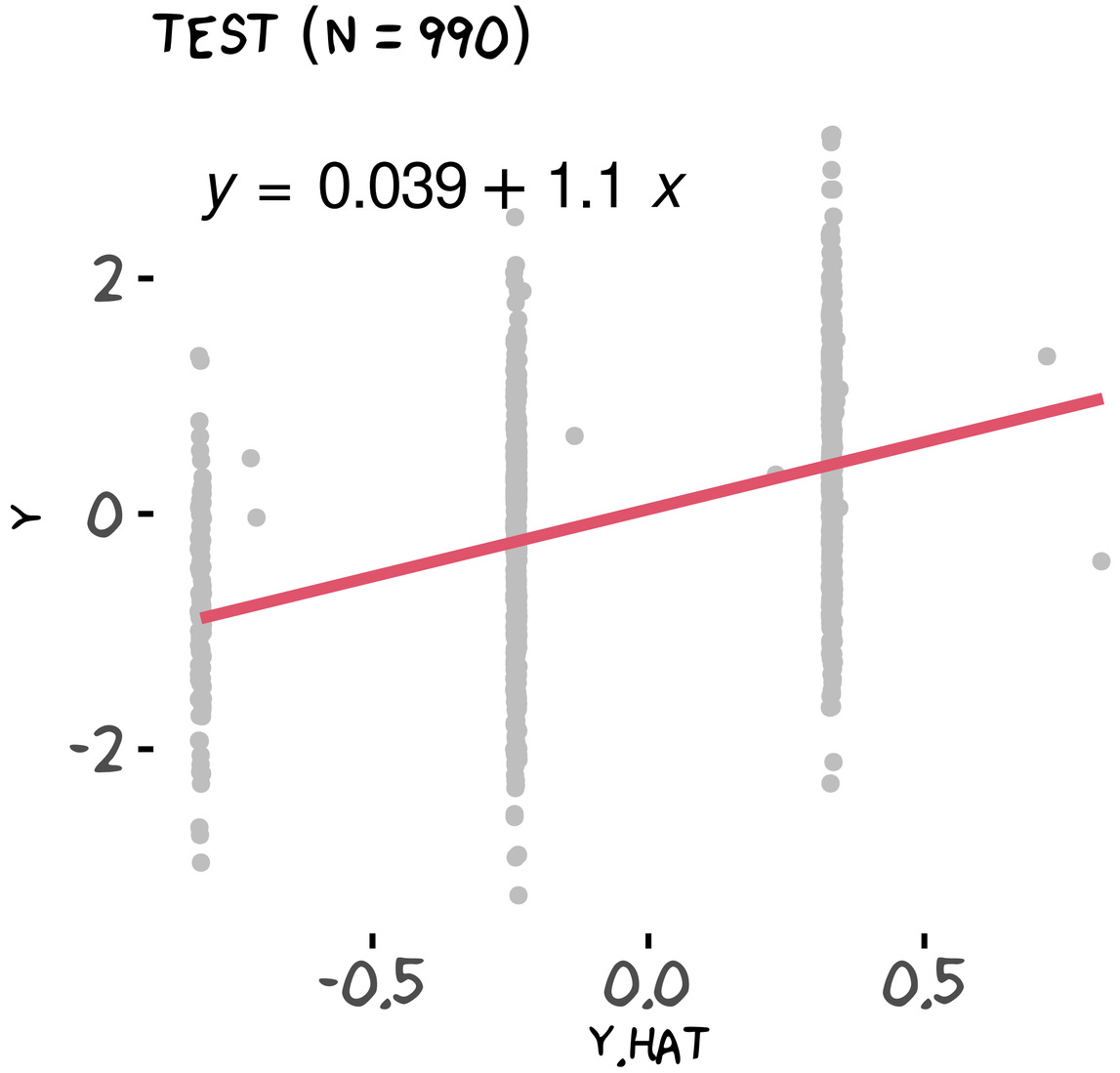
```
lm.out <- lm(Y ~ X.gwas)
```

```
anova(lm.out)
```

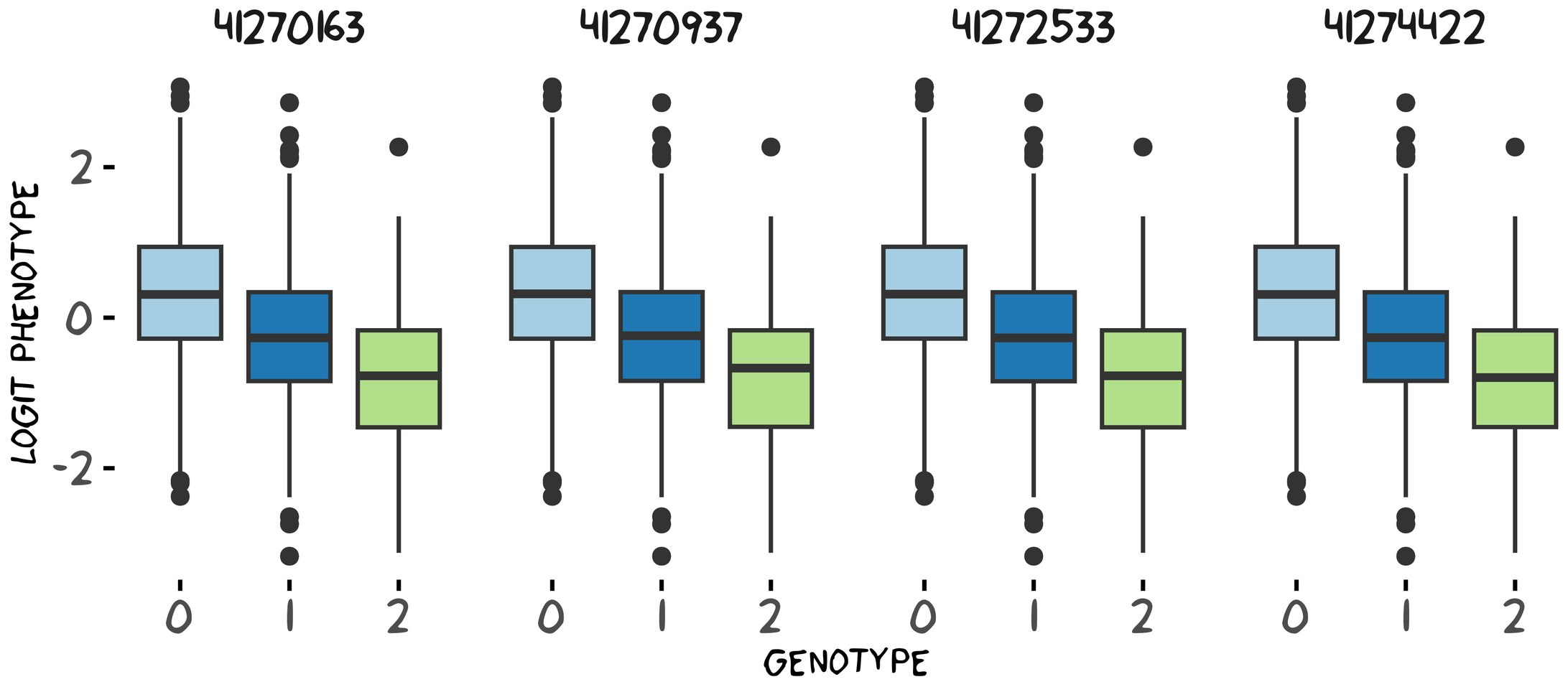
```
## Analysis of Variance Table
##
## Response: Y
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## X.gwas        4  212.77  53.194  63.965 < 2.2e-16 ***
## Residuals  1495 1243.25   0.832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


TRAIN (N = 1500)
$y = 1.9 \times 10^{-17} + x$
2
0
-2
Y
-0.5
0.0
0.5
Y.HAT

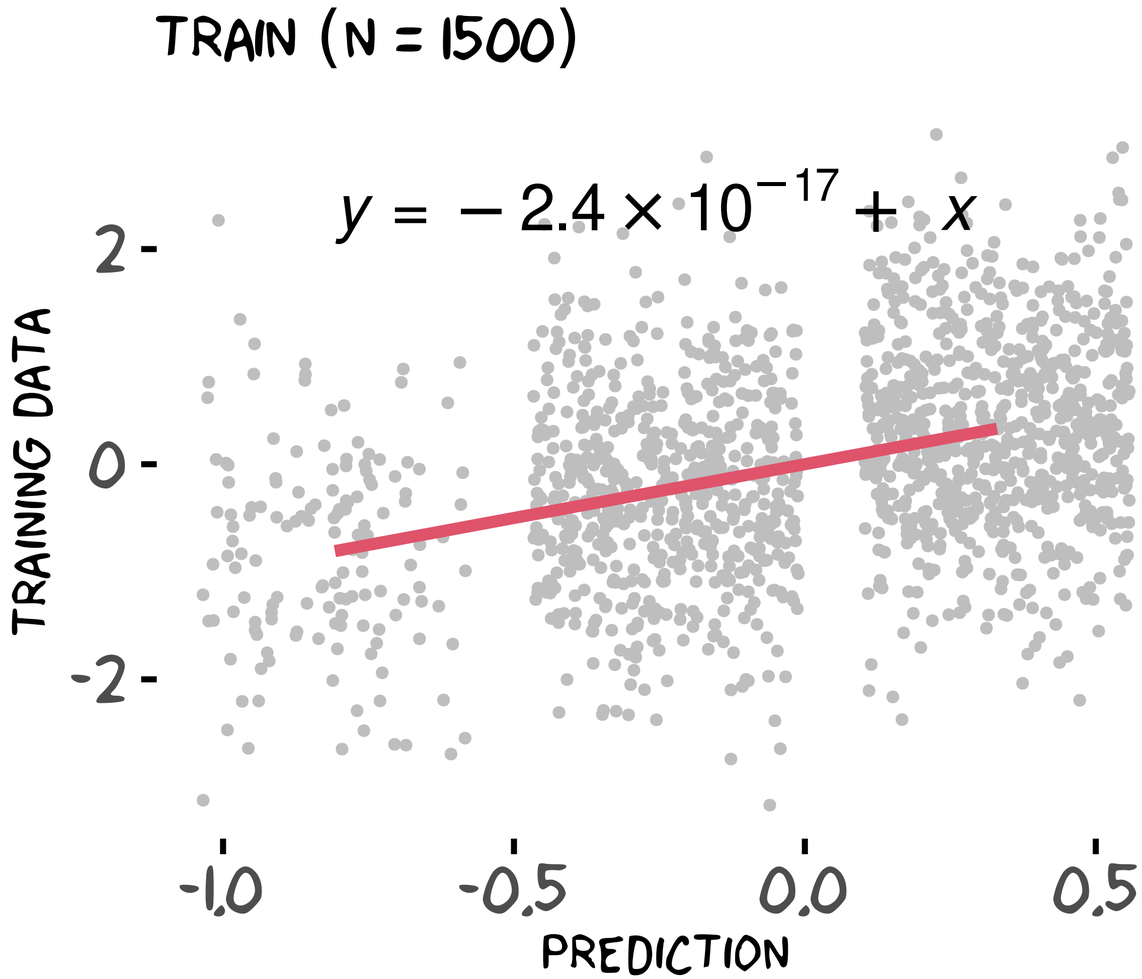

TRAIN (N = 1500)
$y = 1.9 \times 10^{-17} + x$
2
0
-2
Y
-0.5
0.0
0.5
Y.HAT

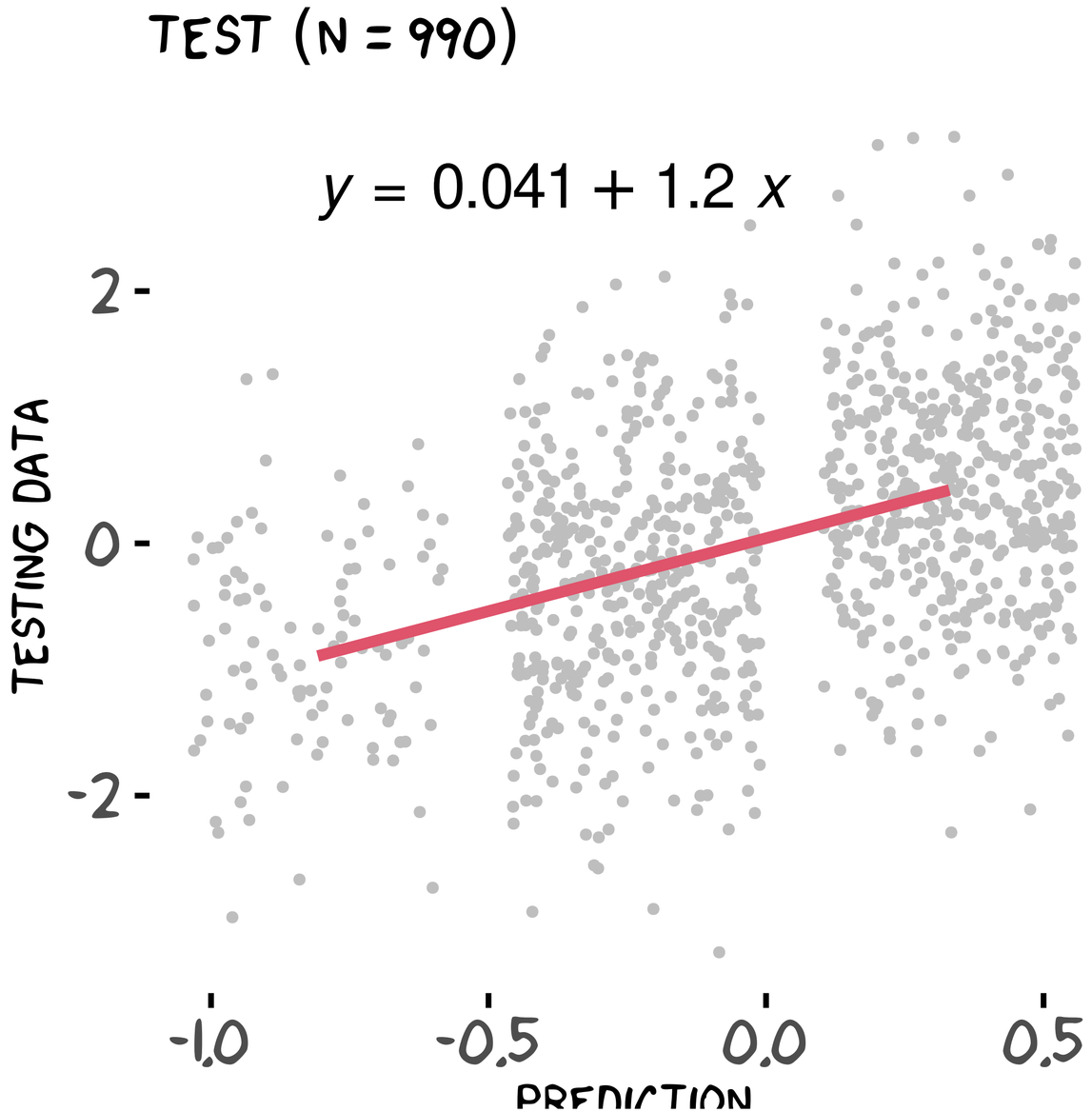

TEST (N = 990)
$y = 0.039 + 1.1\ x$
2
0
-2
Y
-0.5
0.0
0.5
Y.HAT

Do we need all of these GWAS variants?



Just one of the GWAS variants works fine





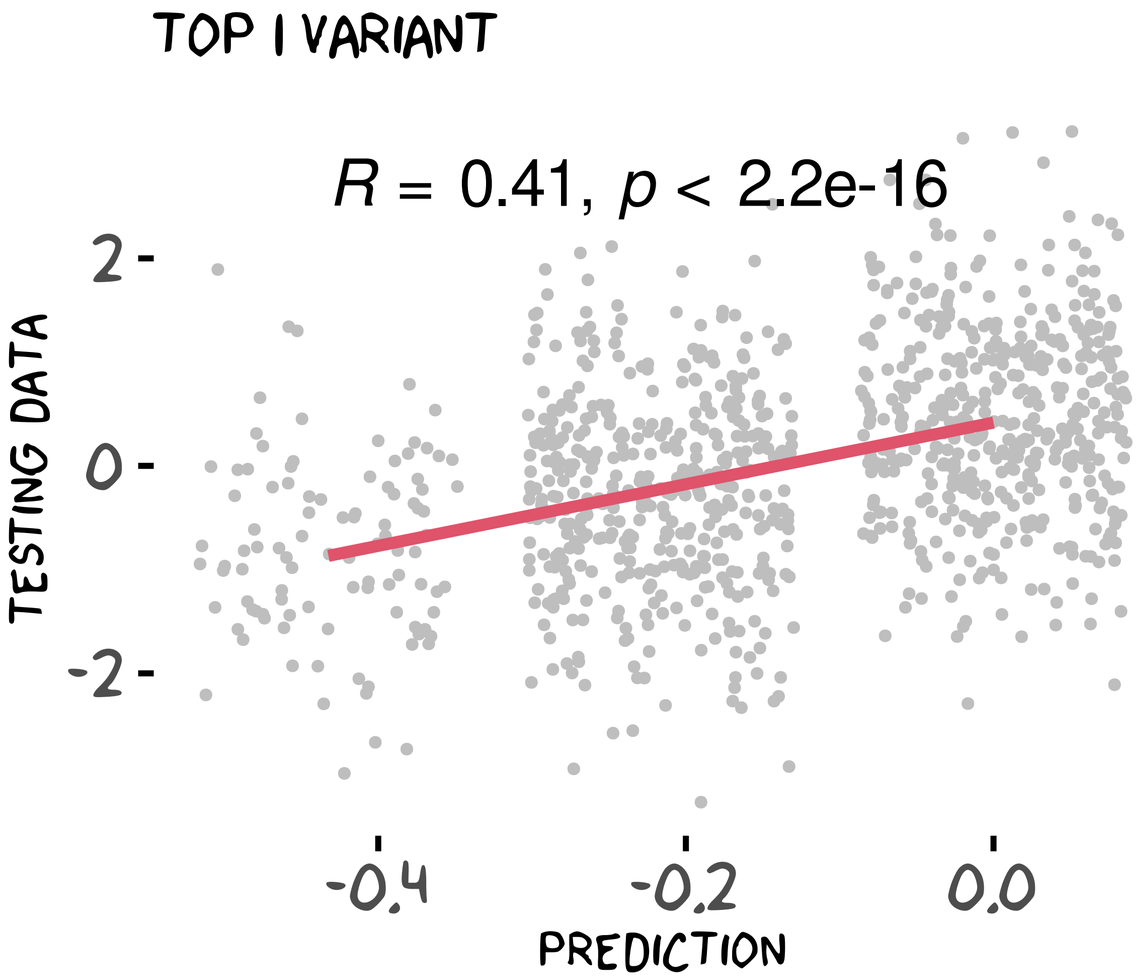
Can we simply add more variables by GWAS significance?

Let's do some experiments:

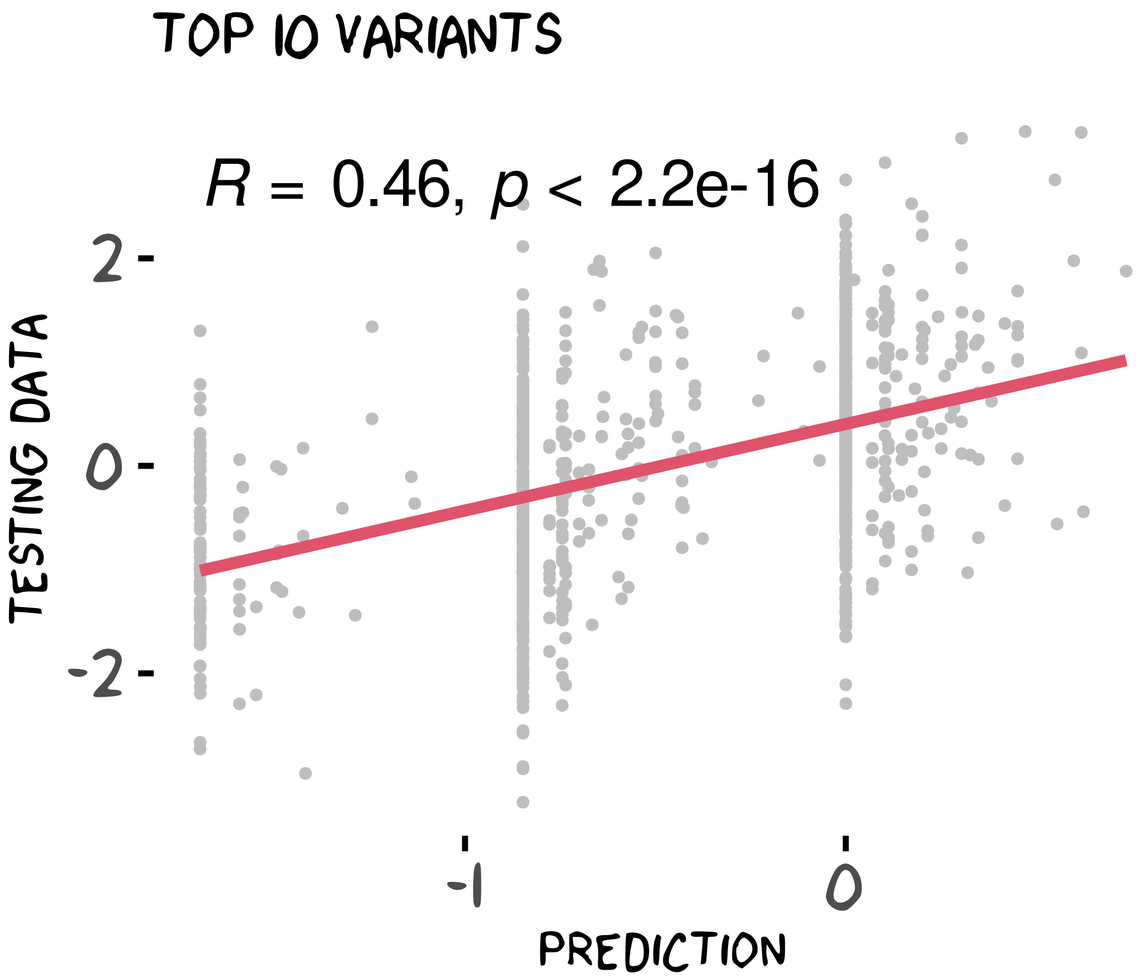
1. Pick top *K* variants (ordering by GWAS, $p_{(1)} < p_{(2)} < \ldots$)
2. Take GWAS effect sizes β (the mean difference between case vs. control)
3. Predict by a linear combination of these effects:

$$\sum_{j=1}^{K} X_{i(j)} \beta_{(j)}$$

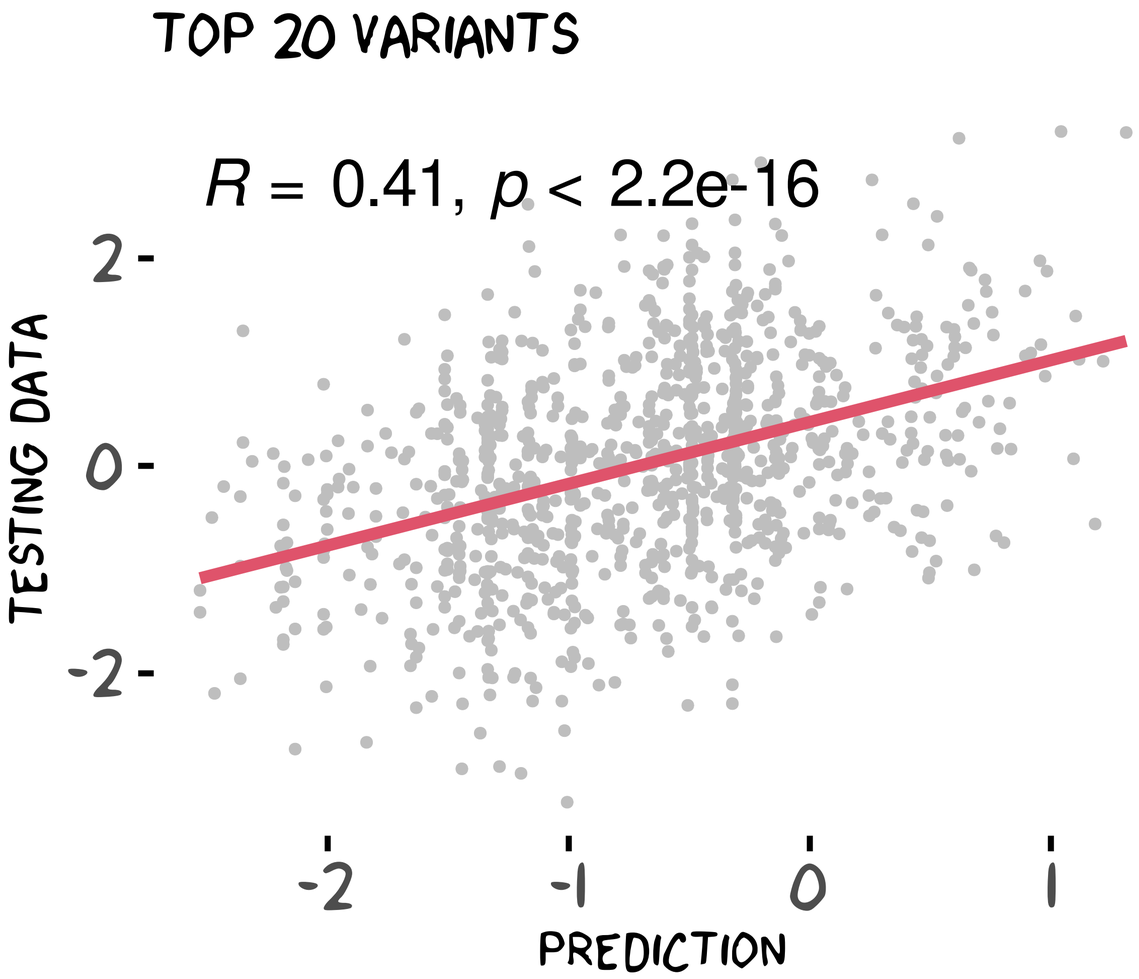
PGS is a variable selection problem



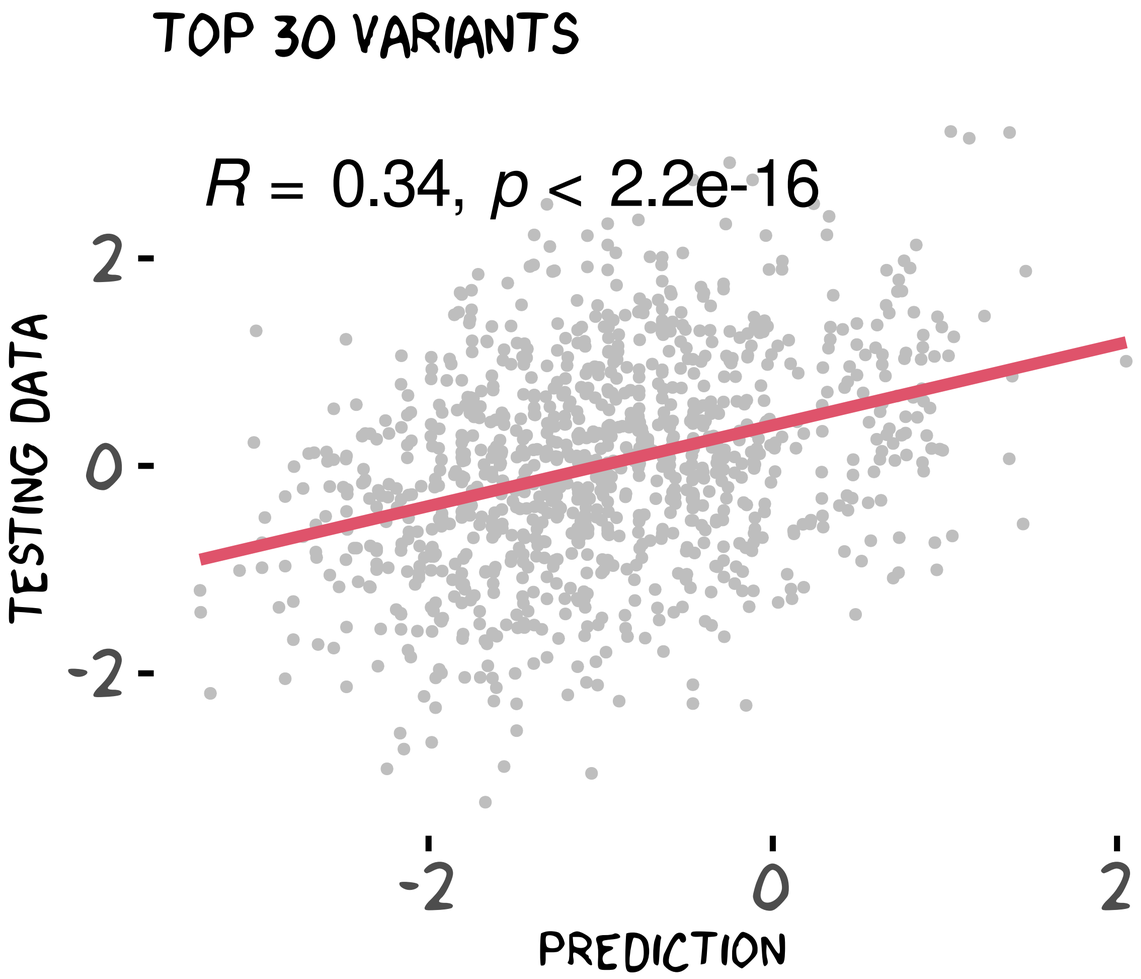
PGS is a variable selection problem



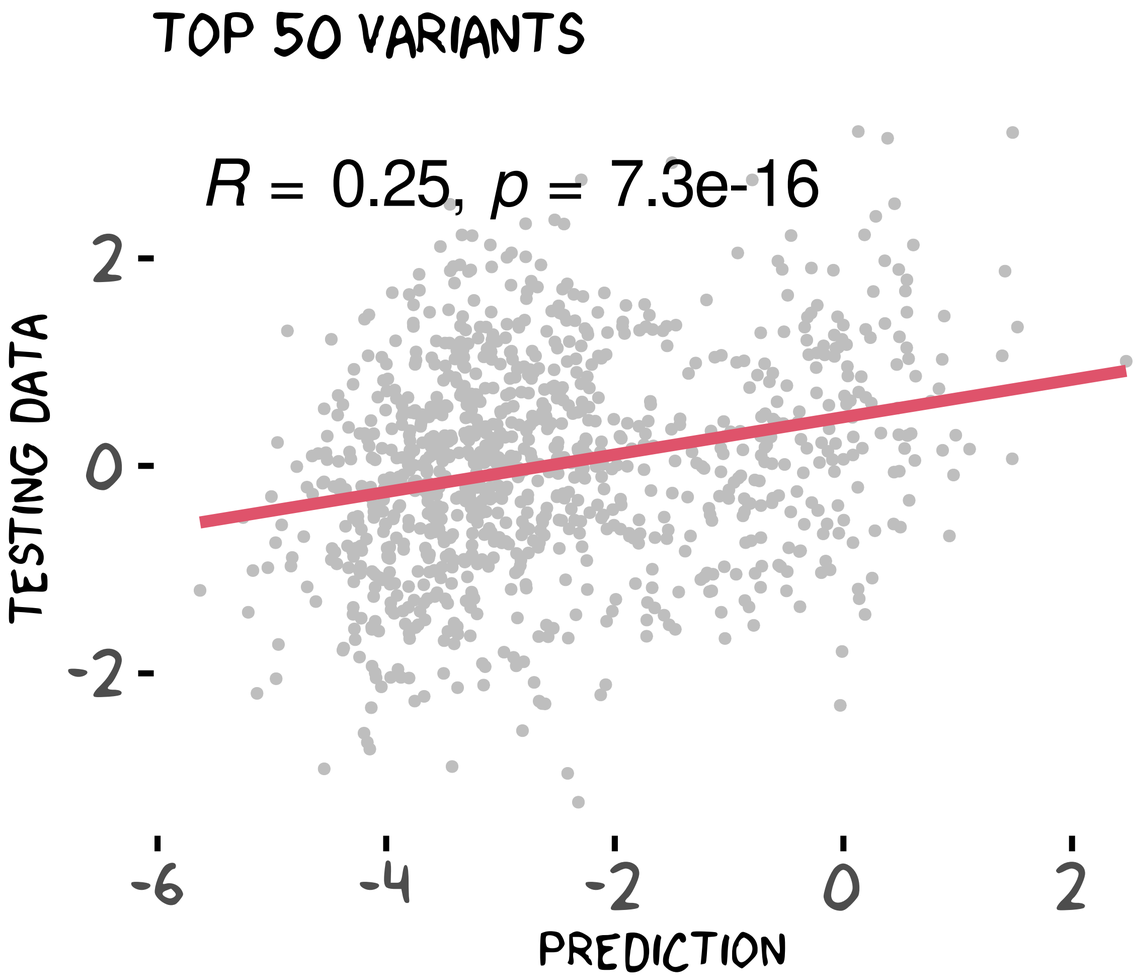
PGS is a variable selection problem

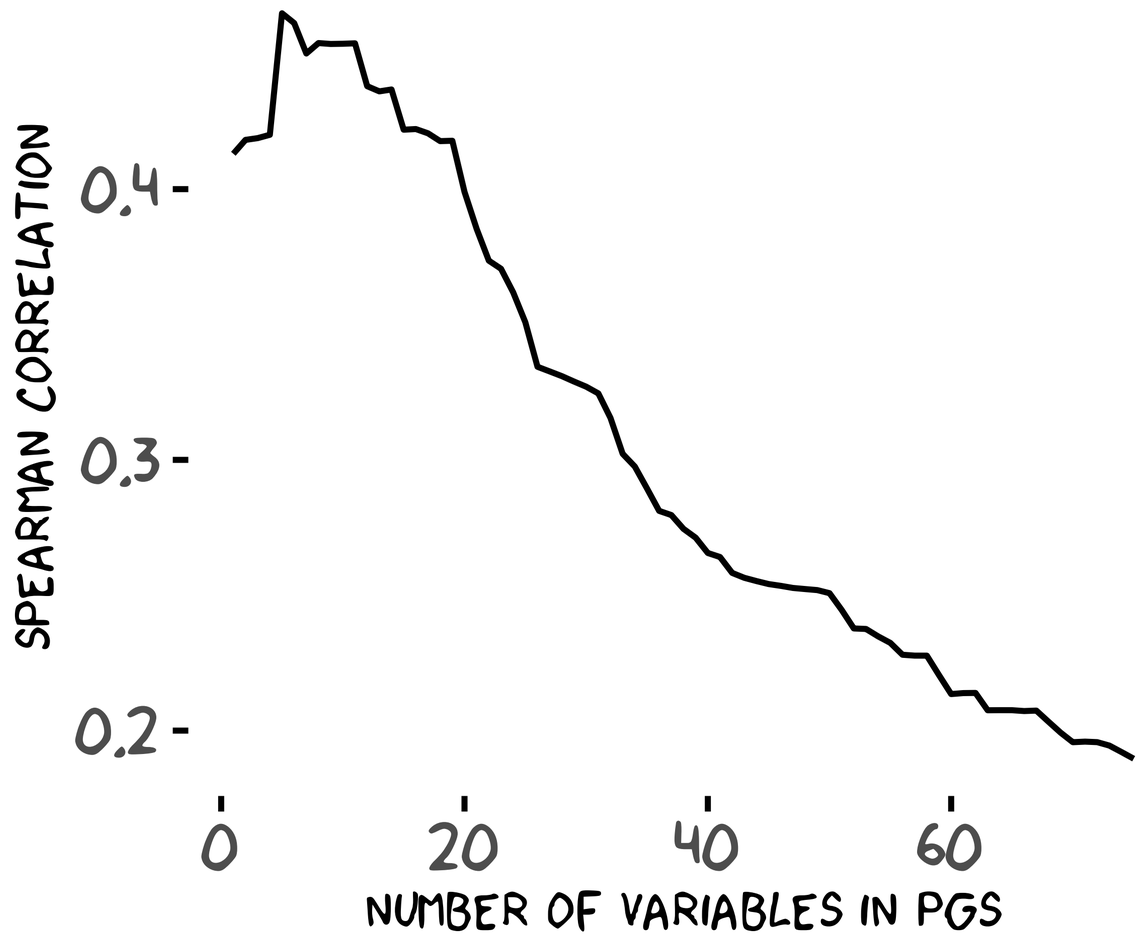


PGS is a variable selection problem



PGS is a variable selection problem

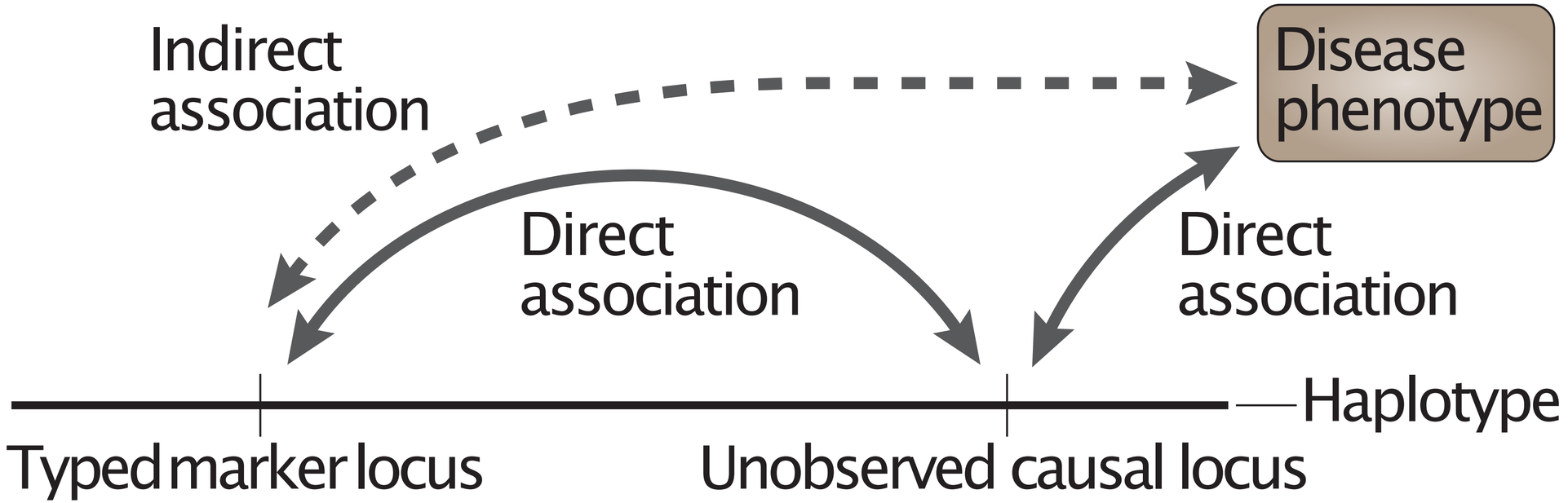




Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. What are the statistical challenges in PGS estimation?
4. Statistical fine-mapping to handle LD structures
5. Other topics

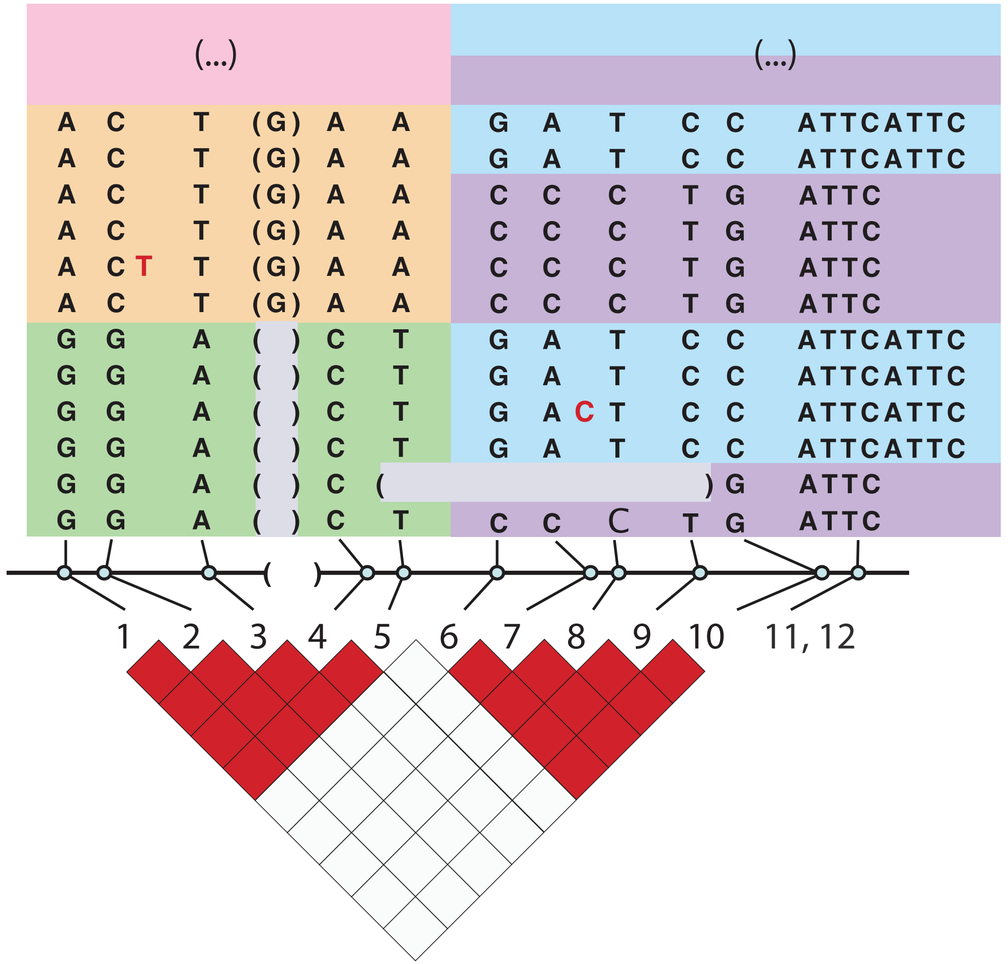
GWAS statistics only roughly tag causal variants



Balding *et al.* Nature Review Genetics (2006)

Covariance between variants obfuscates GWAS interpretation

- Linkage disequilibrium (LD) block
- The result of recombination events throughout generations



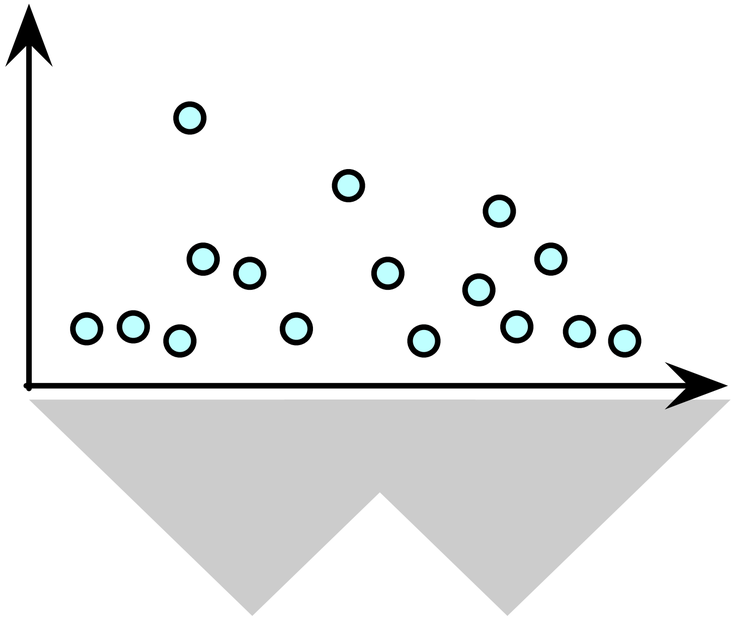
A typical setting of PGS training

PGS training

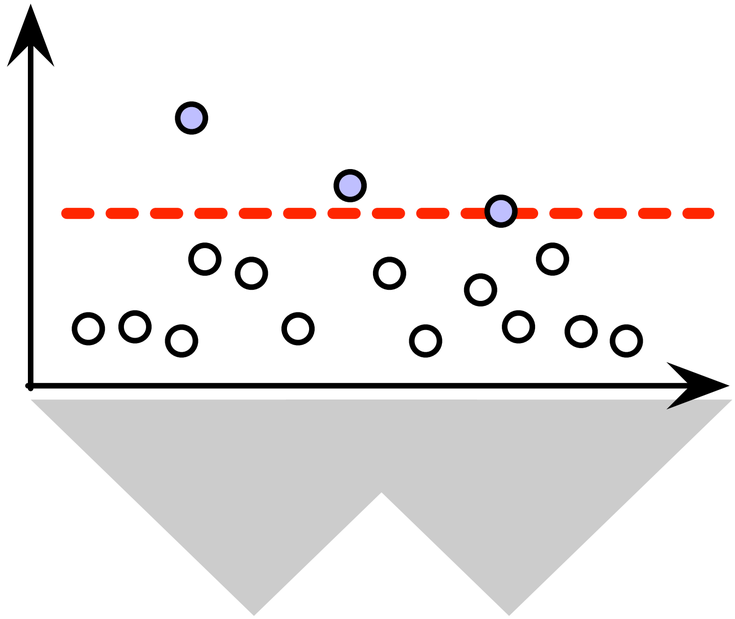
- Input:
  1. GWAS summary statistics (p-values, univariate effect sizes)
  2. $n \times p$ genotype matrix X, where $X_{ij} \in \{0, 1, 2\}$
  3. $n \times 1$ phenotype vector $\mathbf{y}$
- Output:
  - *Polygenic* effect sizes $\hat{\beta}_1, \ldots, \hat{\beta}_p$
- Objective:
  - We want to predict unseen $Y_i^\star$ by $f(X^\star; \hat{\beta})$ as accurately as possible
  - We want $\beta_j = 0$ as many as possible; only some $\beta_j \neq 0$.

Common strategies to deal with LD structures

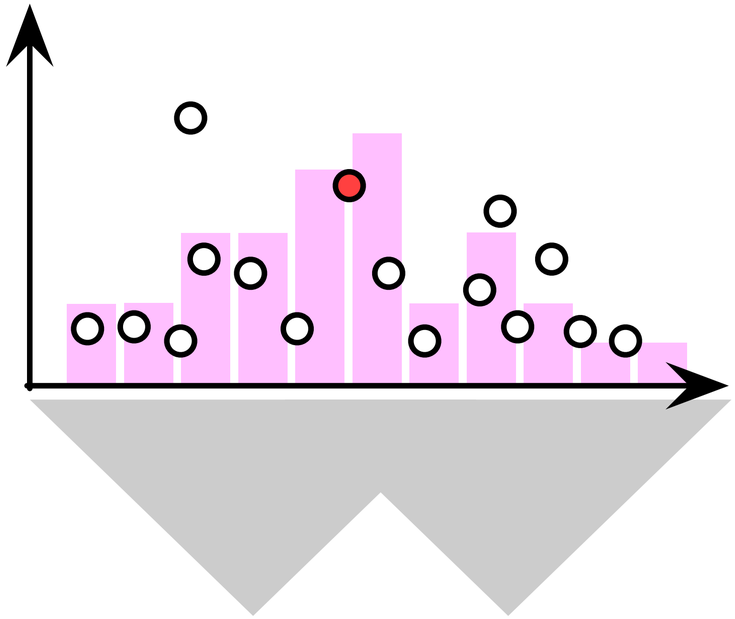
Strategy 1. Just use them all



Strategy 2. Pruning/clumping

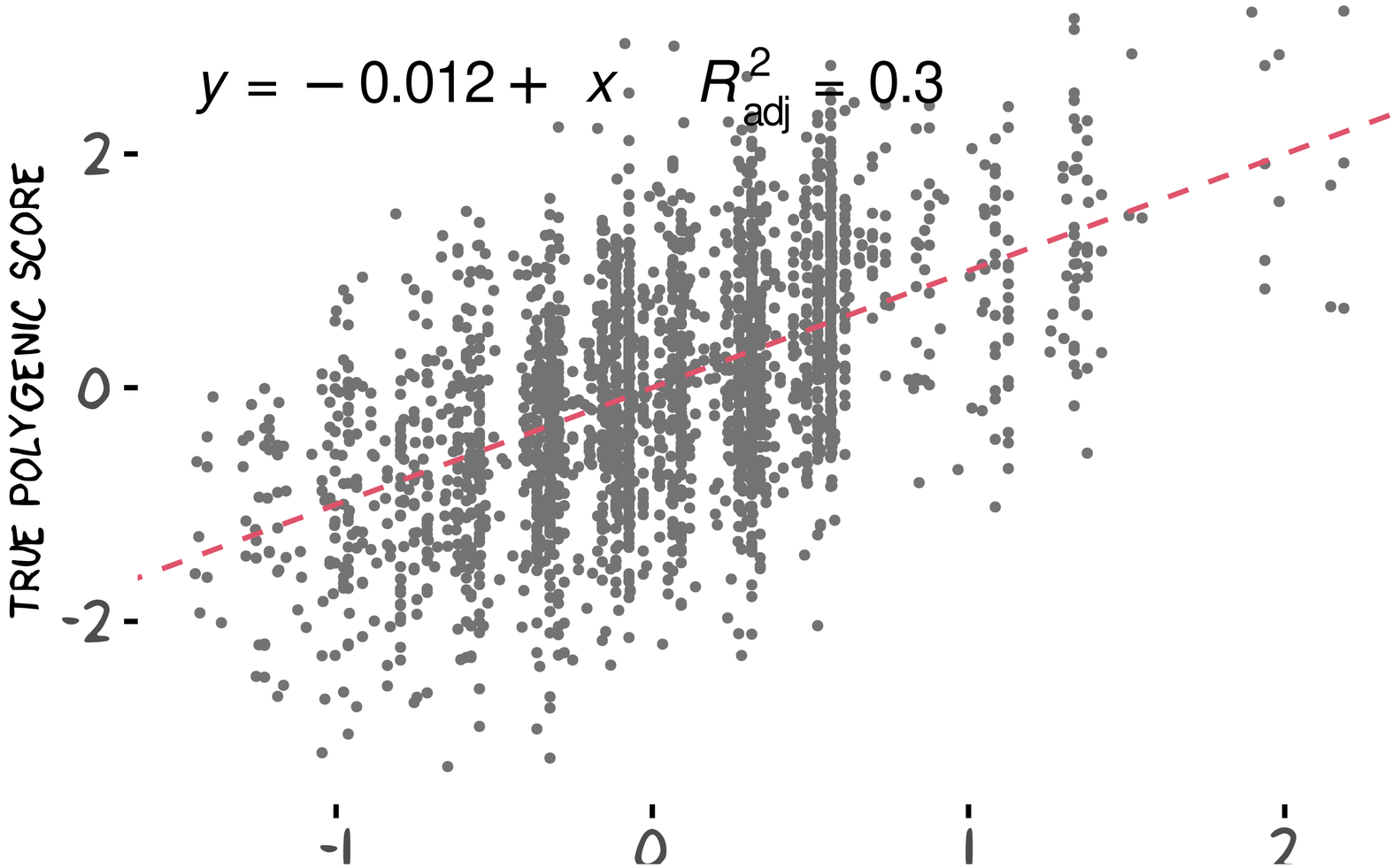


Strategy 3. **Fine-mapping**



True causal variables explain a large fraction of variation

```
.lm <- lm(.sim$y.q ~ .sim$x[, .sim$causal, drop = FALSE] - 1)
```



Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. What are the statistical challenges in PGS estimation?
4. Statistical fine-mapping to handle LD structures
5. Other topics

Classical statistics does not really help

- Classical variable selection by univariate (one-by-one) tests will not work for a $p \gg n$ regression problem
- Especially if we have col-linearity in the design matrix X
- also know as Linkage Disequilibrium (LD) in GWAS/PGS

Variable selection in high-dim. matrix ($n \ll p$)

Regression analysis = projecting the observed **y** vector on to column space of $\{\mathbf{x}_j : j \in [p]\}$,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_1 \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} + \cdots \beta_p \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix}.$$

Variable selection = column selection.

- Intuitive idea : choose the best combination of variables. $\rightarrow 2^p$ choices (even harder).

Variable selection in high-dim. matrix ($n \ll p$)

Regression analysis = projecting the observed **y** vector on to column space of $\{\mathbf{x}_j : j \in [p]\}$,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_1 \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} + \cdots \beta_p \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix}.$$

Variable selection = column selection.

- Intuitive idea : choose the best combination of variables. $\rightarrow 2^p$ choices (even harder).
- Alternative idea : make as many β_j's nearly zero values.

Variable selection in high-dim. matrix ($n \ll p$)

Regression analysis = projecting the observed **y** vector on to column space of $\{\mathbf{x}_j : j \in [p]\}$,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_1 \begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{pmatrix} + \cdots \beta_p \begin{pmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{np} \end{pmatrix}.$$

Variable selection = column selection.

- Intuitive idea : choose the best combination of variables. $\rightarrow 2^p$ choices (even harder).
- Alternative idea : make as many β_j's nearly zero values.
- What prior does: penalize $|\beta_j| > 0$ so that only the strong enough variables take non-zero values

Lasso, a linear regression with Laplace prior (L1)

Prior distribution

$$p(\theta) = \text{Laplace}(\theta|\lambda) \propto \exp\left(-\lambda\|\theta\|_1\right)$$

where

$$\|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|, \text{ L1-norm.}$$

Maximize

$$\ln p(\mathbf{y}|X,\theta) + \ln p(\theta|\lambda) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\theta)^2 - \lambda\|\theta\|_1$$

Minimize L_1-regularized error

`glmnet` solves this regularized optimization problem

Goal (by variable-by-variable updates):

$$\min_\beta \quad \overbrace{(\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta)}^{\text{RSS}} + \underbrace{\lambda\alpha\|\beta\|_1}_{\text{variable selection}} + \underbrace{\lambda(1-\alpha)\|\beta\|_2}_{\text{shrinkage}}$$

`glmnet` solves this regularized optimization problem

Goal (by variable-by-variable updates):

$$\min_\beta \quad \overbrace{(\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta)}^{\text{RSS}} + \underbrace{\lambda\alpha\|\beta\|_1}_{\text{variable selection}} + \underbrace{\lambda(1-\alpha)\|\beta\|_2}_{\text{shrinkage}}$$

For each β_j,

$$\hat{\beta}_j^{\text{glmnet}} \leftarrow \frac{S\left(\sum_{i=1}^n X_{ij}(y_i - \hat{y}_i^{(-j)}), \lambda\alpha\right)}{\sum_{i=1}^n X_{ij}^2 + \lambda(1-\alpha)}$$

Friedman *et al.*, Regularization Paths for Generalized Linear Models via Coordinate Descent (2010)

`glmnet` solves this regularized optimization problem

Goal (by variable-by-variable updates):

$$\min_\beta \quad \overbrace{(\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta)}^{\text{RSS}} + \underbrace{\lambda\alpha\|\beta\|_1}_{\text{variable selection}} + \underbrace{\lambda(1-\alpha)\|\beta\|_2}_{\text{shrinkage}}$$

For each β_j,

$$\hat{\beta}_j \leftarrow \frac{\overset{\text{threshold}}{S}\left(\sum_{i=1}^n X_{ij} \overbrace{(y_i - y_i^{(-j)})}^{\text{residual w/o the variable } \beta_j}, \lambda\alpha\right)}{\sum_{i=1}^n X_{ij}^2 + \underbrace{\lambda(1-\alpha)}_{\text{shrinkage}}}$$

Running `glmnet` for polygenic risk prediction of OGD

```
glm.cv.out <- glmnet::cv.glmnet(X, Y, nfolds = 5, alpha = 1)
lambda.cv <- glm.cv.out$lambda.min
glm.out <- glmnet::glmnet(x = X, y = Y, lambda = lambda.cv, alpha = 1)
```

```
head(glm.out$beta)
```

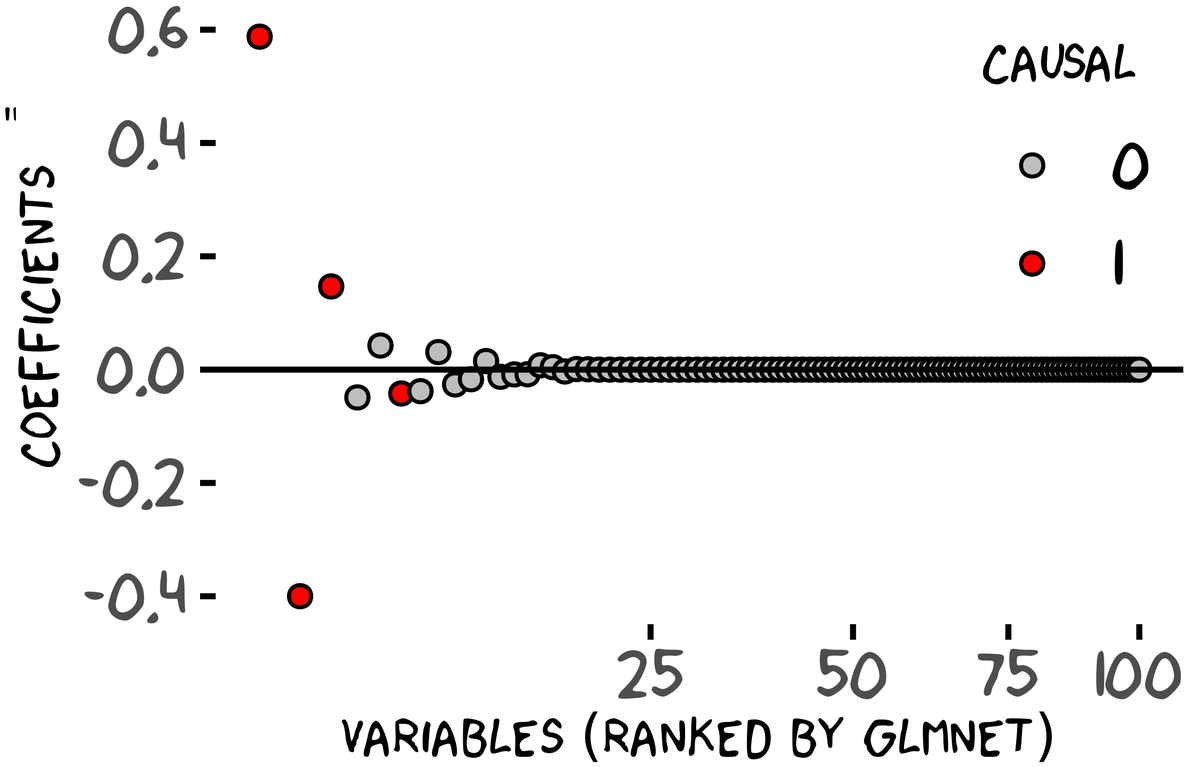
```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##        s0
## 566875  .
## 752721  .
## 754105  .
## 754182  .
## 754334  .
## 756604  .
```

Running `glmnet` for polygenic risk prediction of OGD

```
glm.cv.out <- glmnet::cv.glmnet(X, Y, nfolds = 5, alpha = 1)
lambda.cv <- glm.cv.out$lambda.min
glm.out <- glmnet::glmnet(x = X, y = Y, lambda = lambda.cv, alpha = 1)
```

```
head(glm.out$beta)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##        s0
## 566875  .
## 752721  .
## 754105  .
## 754182  .
## 754334  .
## 756604  .
```



Sum of Single Effect (SuSiE) regression

```
library(susieR)
susie.out <- susie(X, Y)
```

$$\mathbf{y} = \sum_{l=1}^{L} \underbrace{\sum_{j} \mathbf{x}_j \quad \alpha_j^{(l)} \quad \beta_j^{(l)}}_{\text{layer-by-layer}} + \epsilon$$

probabilistic selection single variant effect

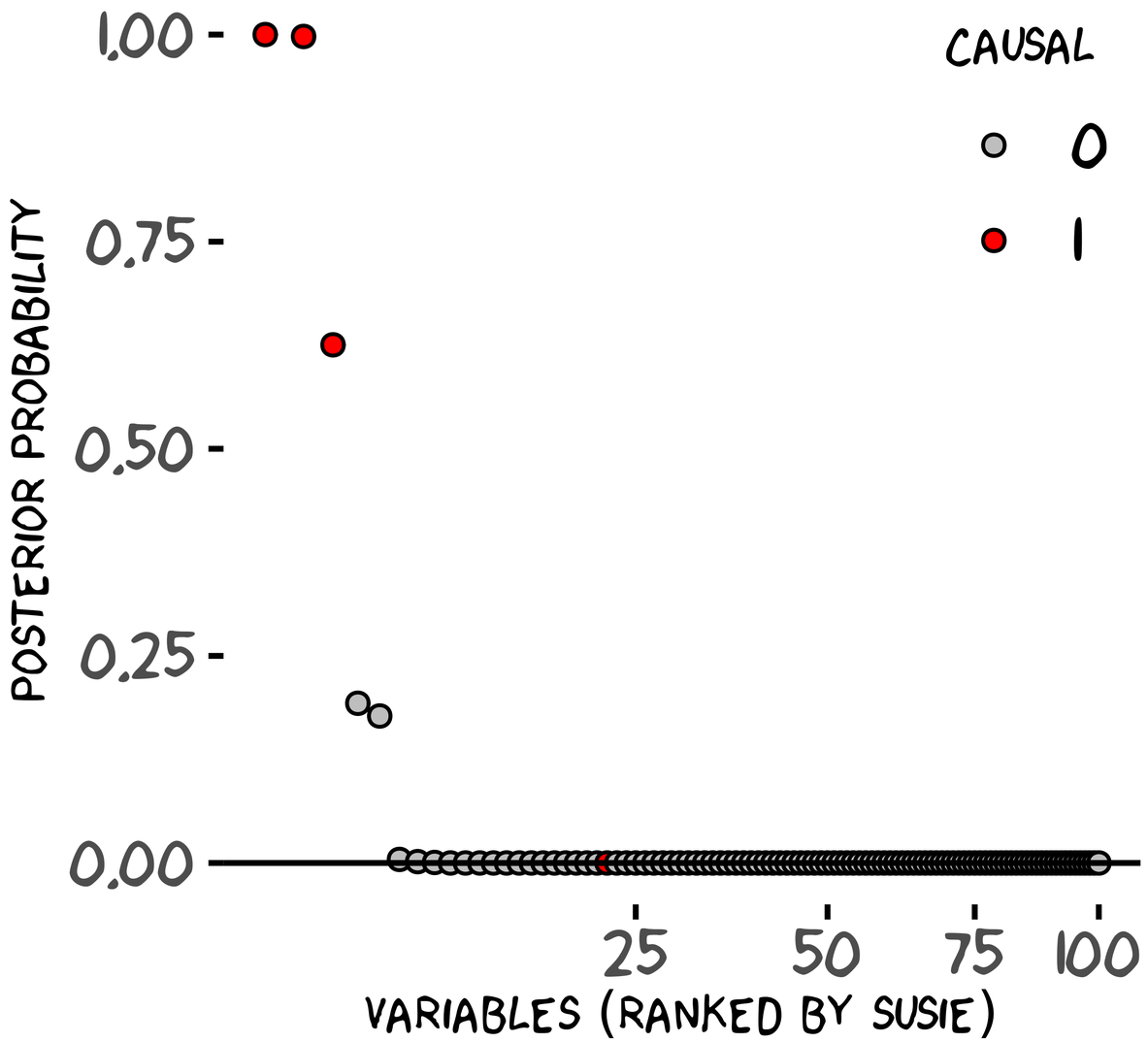
where $\sum_{j=1}^{p} \alpha_j^{(l)} = 1$ for each layer l.

Wang .. Stephens, *Journal of the Royal Statistical Society* (2020)

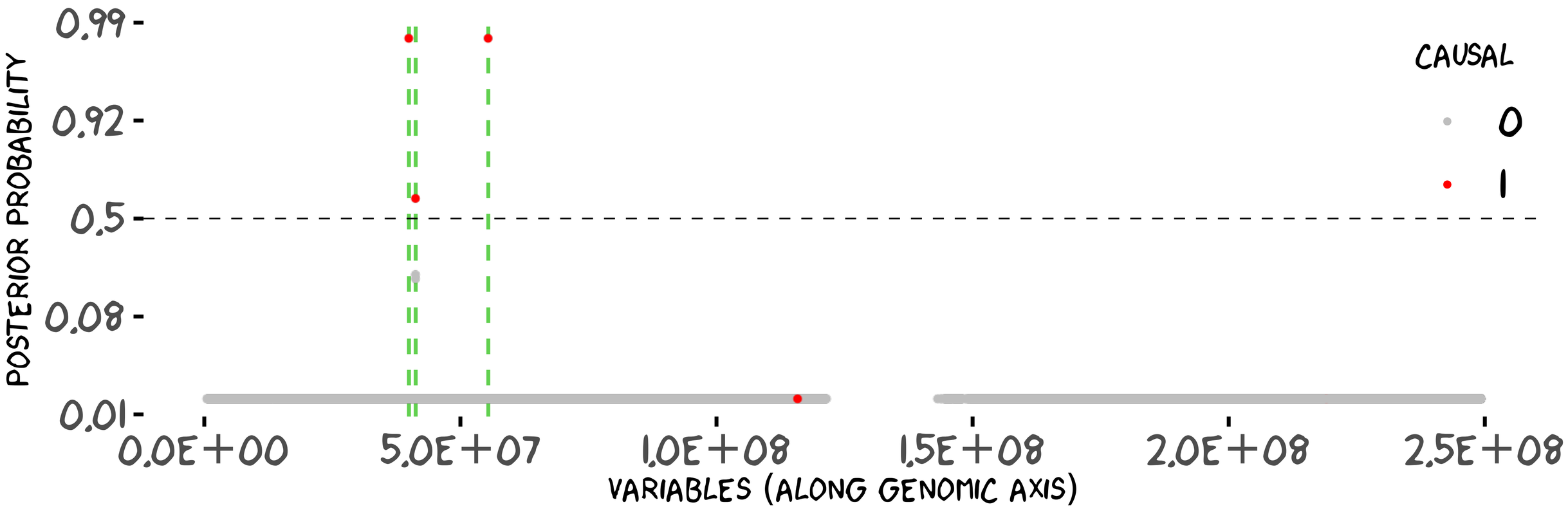
Ideas behind SuSiE

1. Estimating univariate effects is easy (GWAS)
2. Many variants can show similar effects (LD)
3. Let’s weight variants probabilistically
4. Regress out the probabilistically reweighted effects
5. Repeat 1-4.

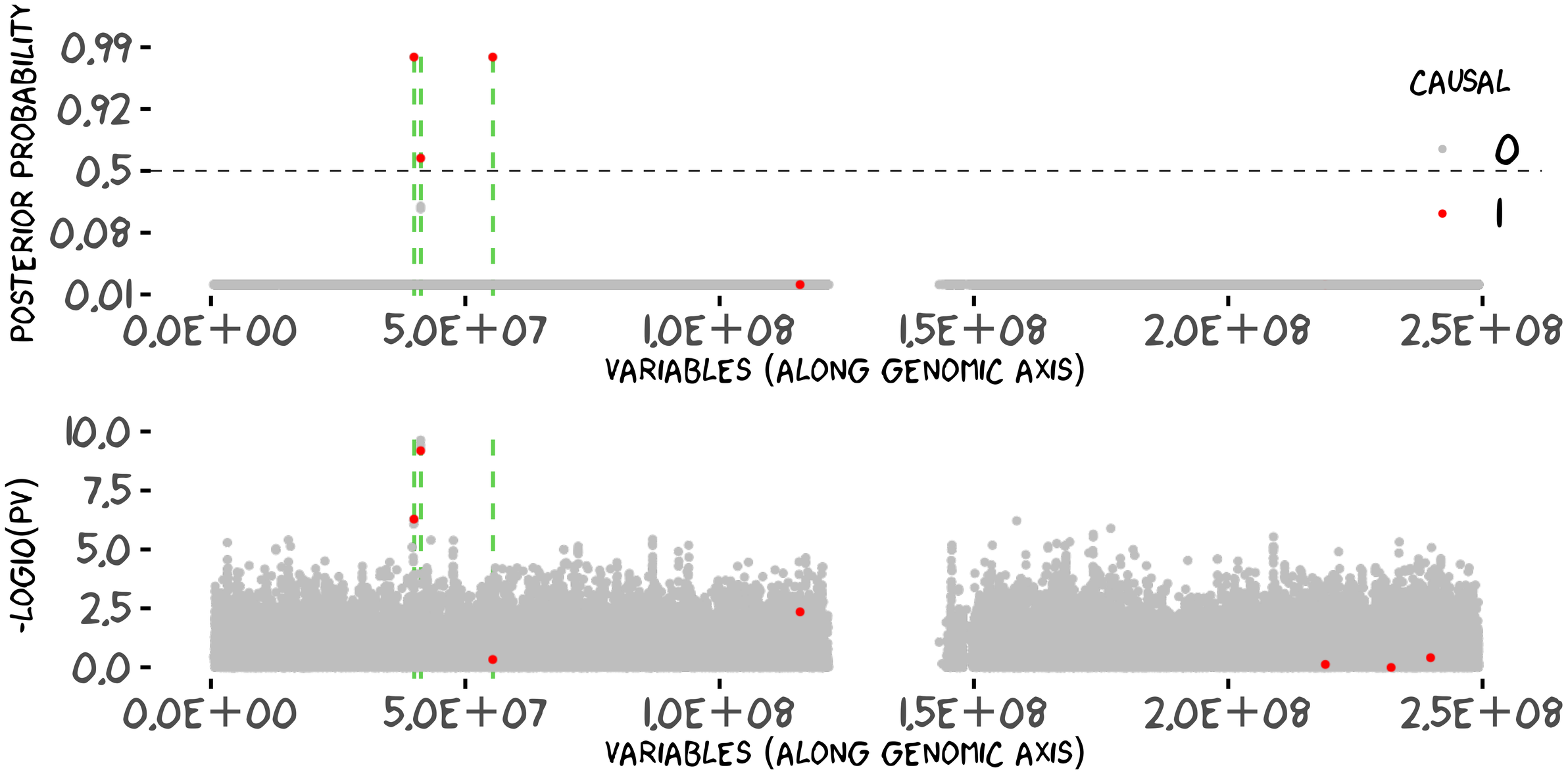
SuSiE identifies top causal variants



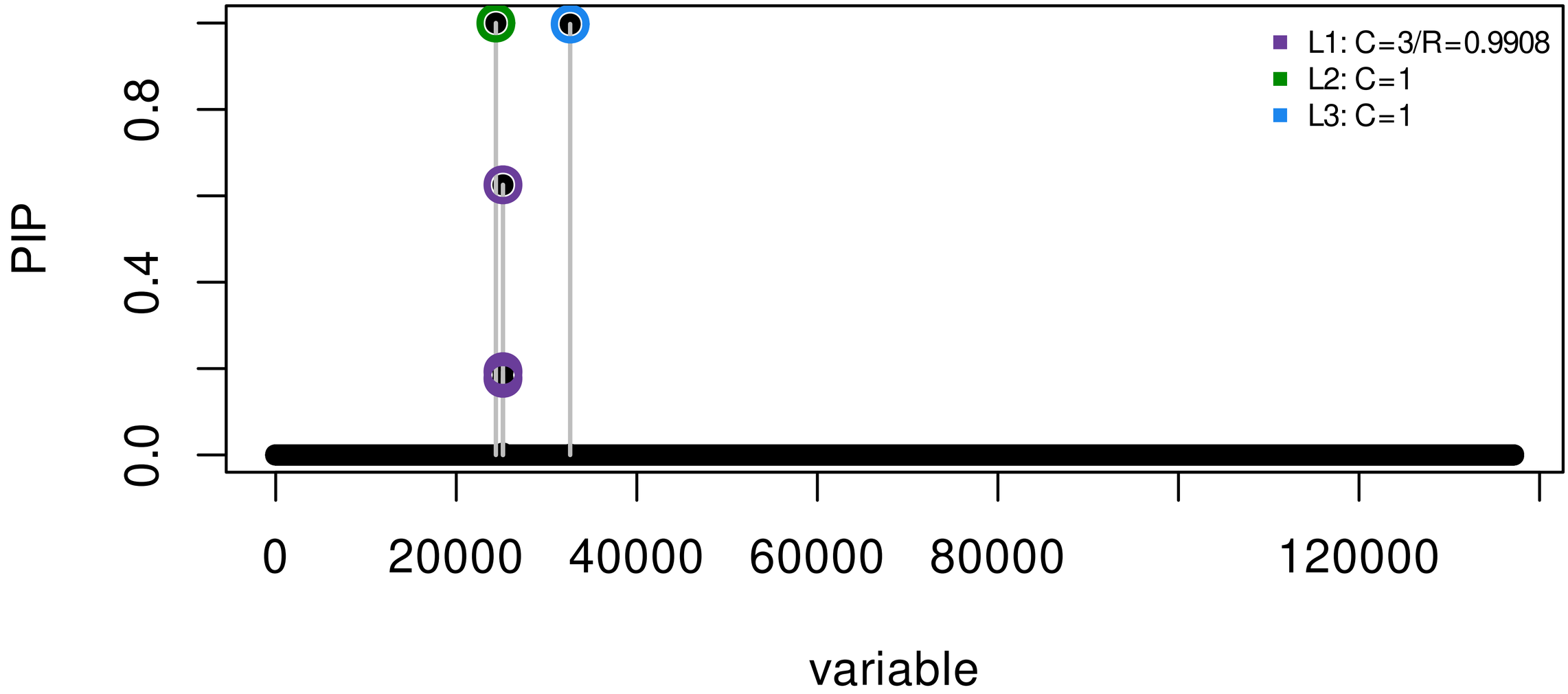
SuSiE can avoid the col-linearity (LD) problem



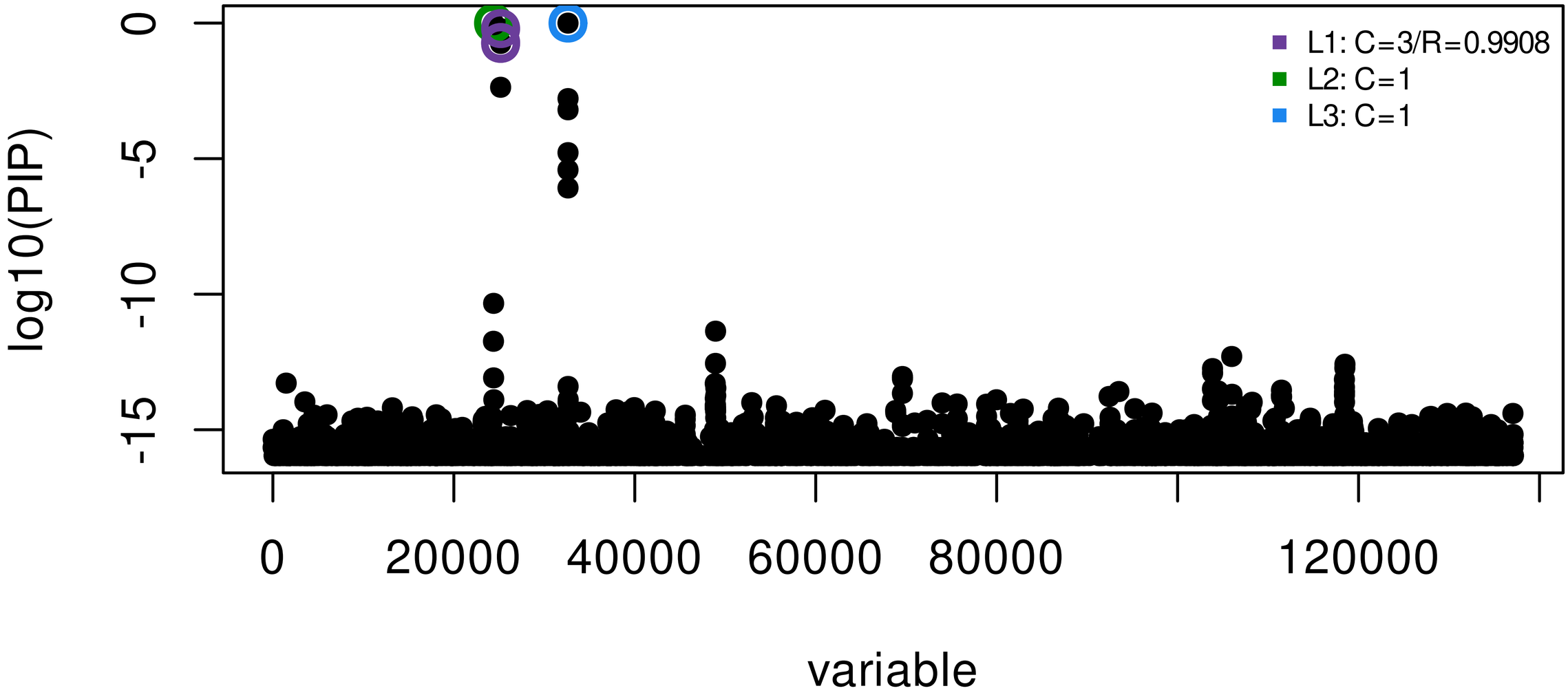
SuSiE can avoid the col-linearity (LD) problem



SuSiE can avoid the col-linearity (LD) problem

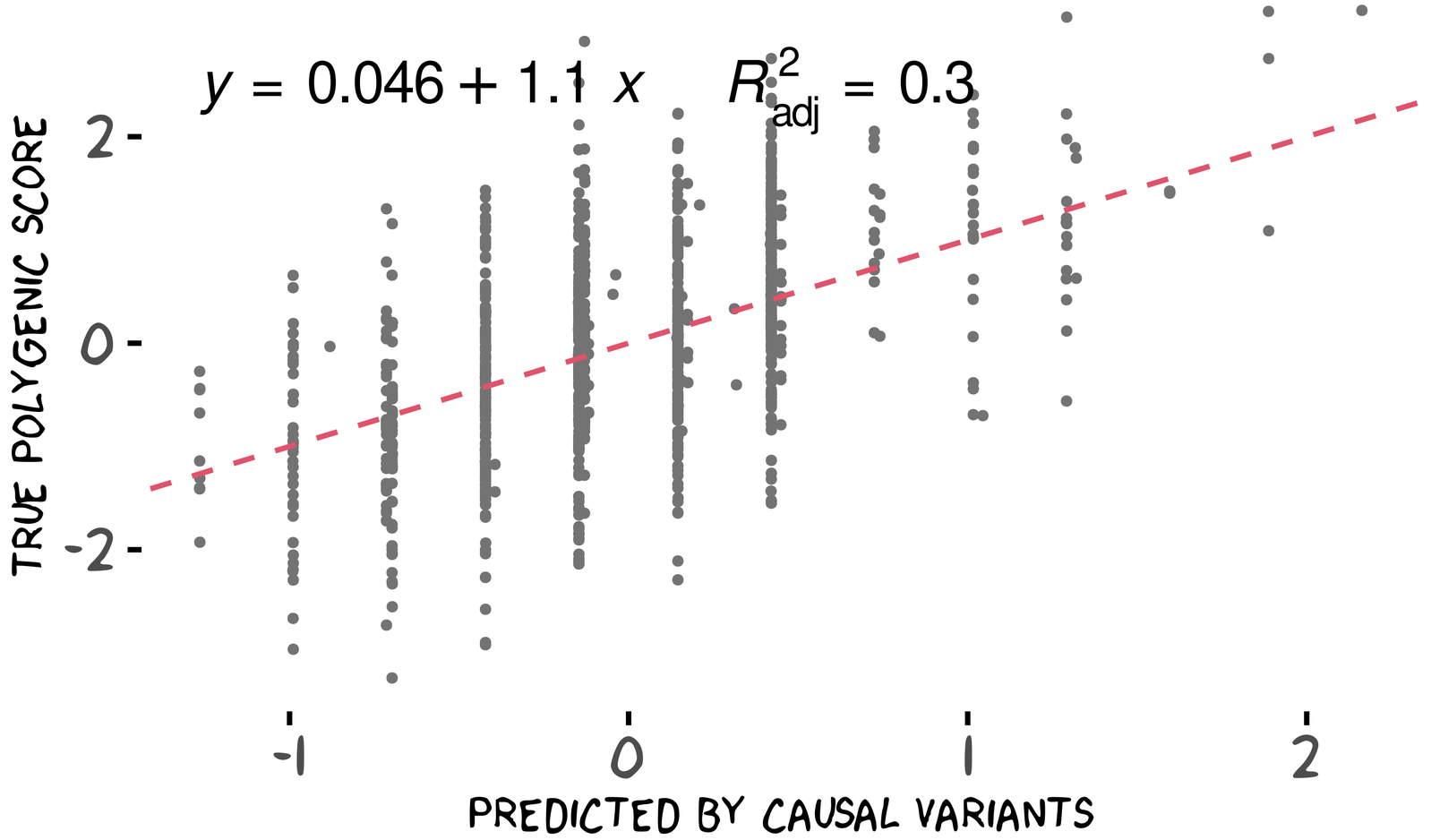


SuSiE can avoid the col-linearity (LD) problem

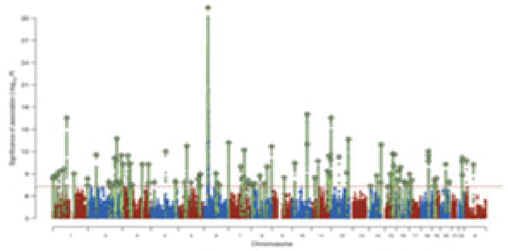


SuSiE+PGS accurately predict the held-out data

```
y.hat <- predict.susie(susie.out, x.test)
```



A general workflow of PGS estimation



Base data

Independent samples

Target data

Data

- Summary statistics
- Betas/ORs weights in PRS calculation

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

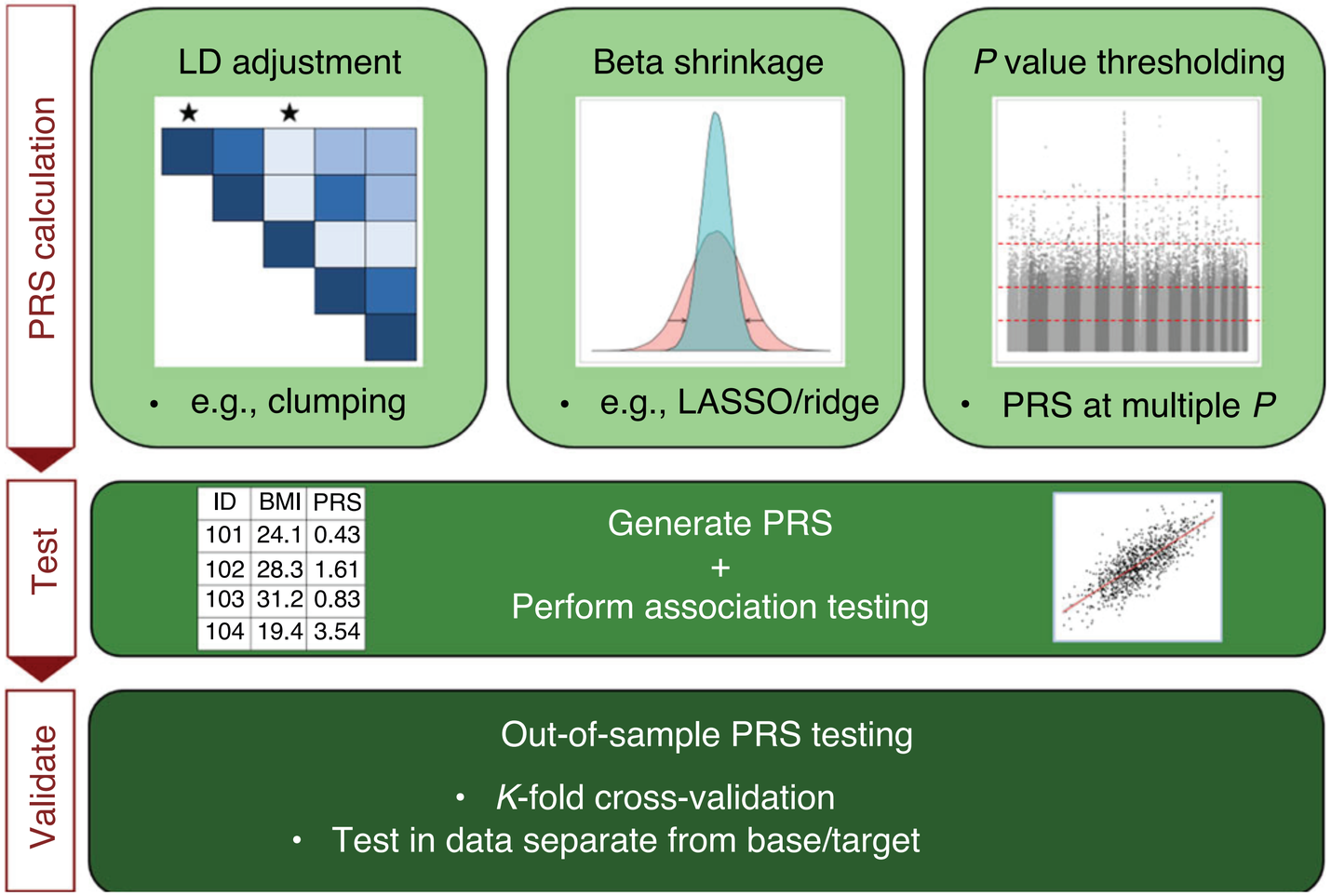
Processing

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

Choi .. O'Reilly, Nature Protocol, (2020)

A general workflow of PGS estimation



Choi .. O'Reilly, Nature Protocol, (2020)

A proposed workflow of PGS estimation

1. Take GWAS summary statistics and target population genotype X
2. Run SuSiE[2] chromosome by chromosome to select variables
3. Predict PGS per chromosome and aggregate them

[2] SuSiE can run with summary statistics only

Advanced topics that we might consider next lecture

- Summary-based GWAS, post-GWAS analysis
- Mendelian Randomization and other causal inference techniques

Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. What are the statistical challenges in PGS estimation?
4. Statistical fine-mapping to handle LD structures
5. Other topics

Training PGS in multiple ancestry groups

Article

Polygenic scoring accuracy varies across the genetic ancestry continuum

https://doi.org/10.1038/s41586-023-06079-4

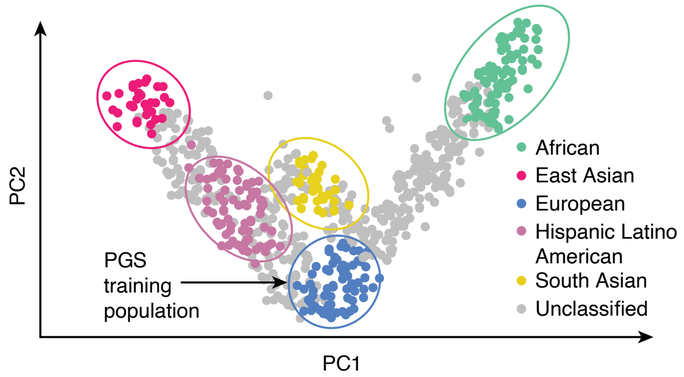
Received: 28 September 2022

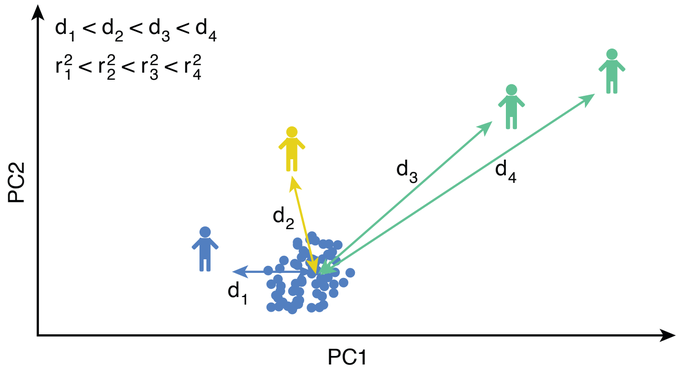
Yi Ding[1]✉, Kangcheng Hou[1], Ziqi Xu[2], Aditya Pimplaskar[1], Ella Petter[2], Kristin Boulier[1], Florian Privé[3], Bjarni J. Vilhjálmsson[3,4,5], Loes M. Olde Loohuis[6,7] & Bogdan Pasaniuc[1,7,8,9,10]✉

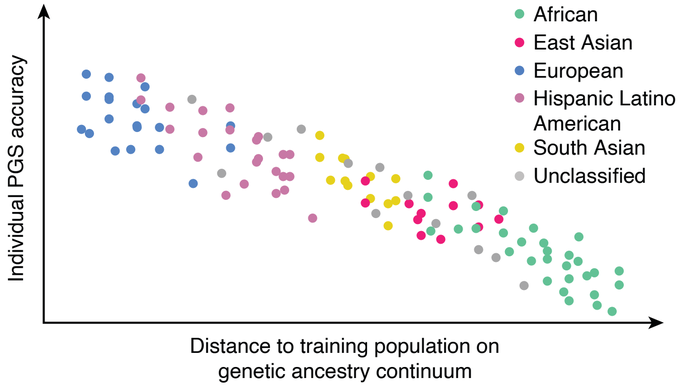
PGS in EUR ≠ PGS in EAS

Ding .. Pasaniuc, Nature (2023)

PGS accuracy rapidly decays across ancestry groups

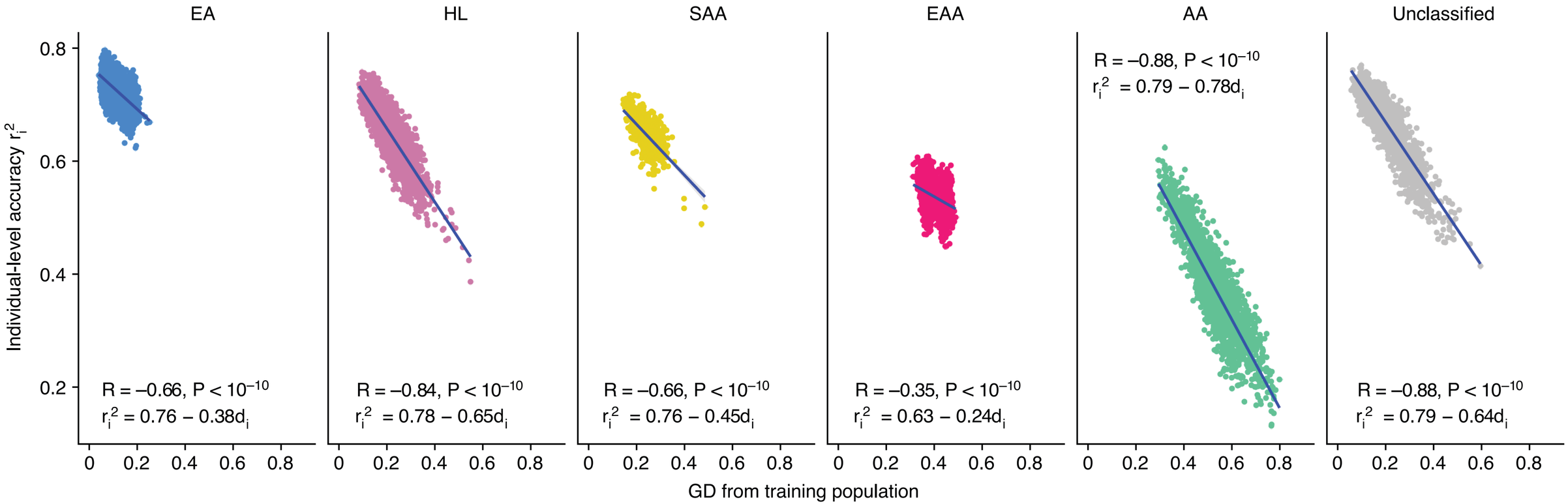






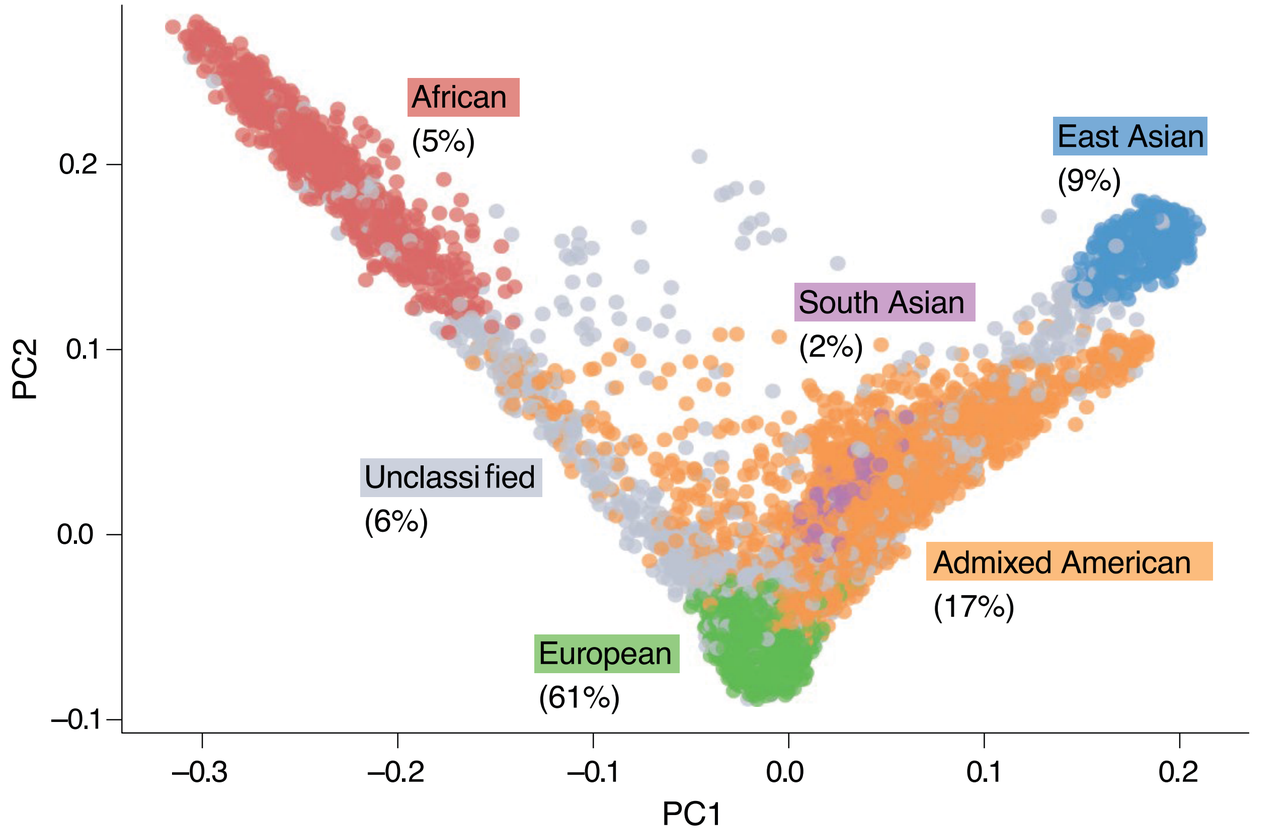
Ding .. Pasaniuc, Nature (2023)

PGS accuracy rapidly decays across ancestry groups



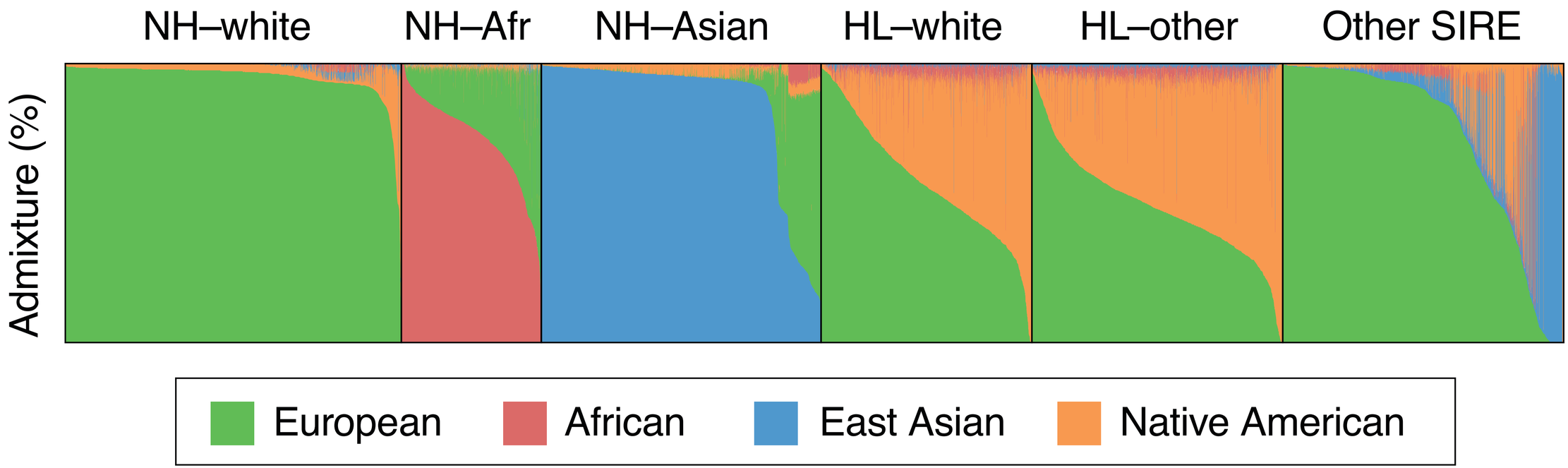
Ding .. Pasaniuc, Nature (2023)

Accurate PGS across many populations is challenging



Kachuri et al. Nature Review Genetics (2023)

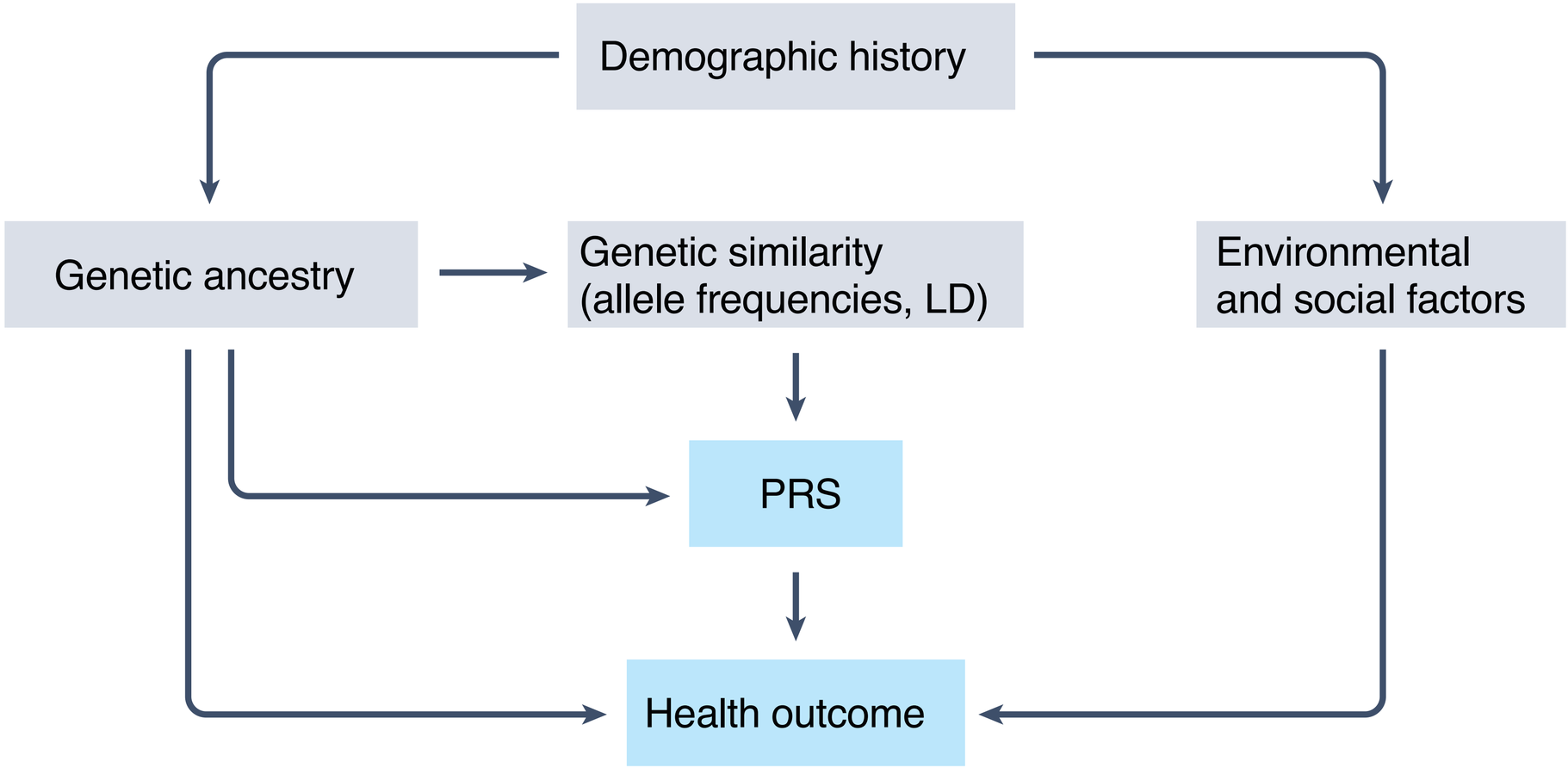
Accurate PGS across many populations is challenging



Kachuri et al. Nature Review Genetics (2023)

NH = non-Hispanic, HL = Hispanic Latino

Accurate PGS across many populations is challenging



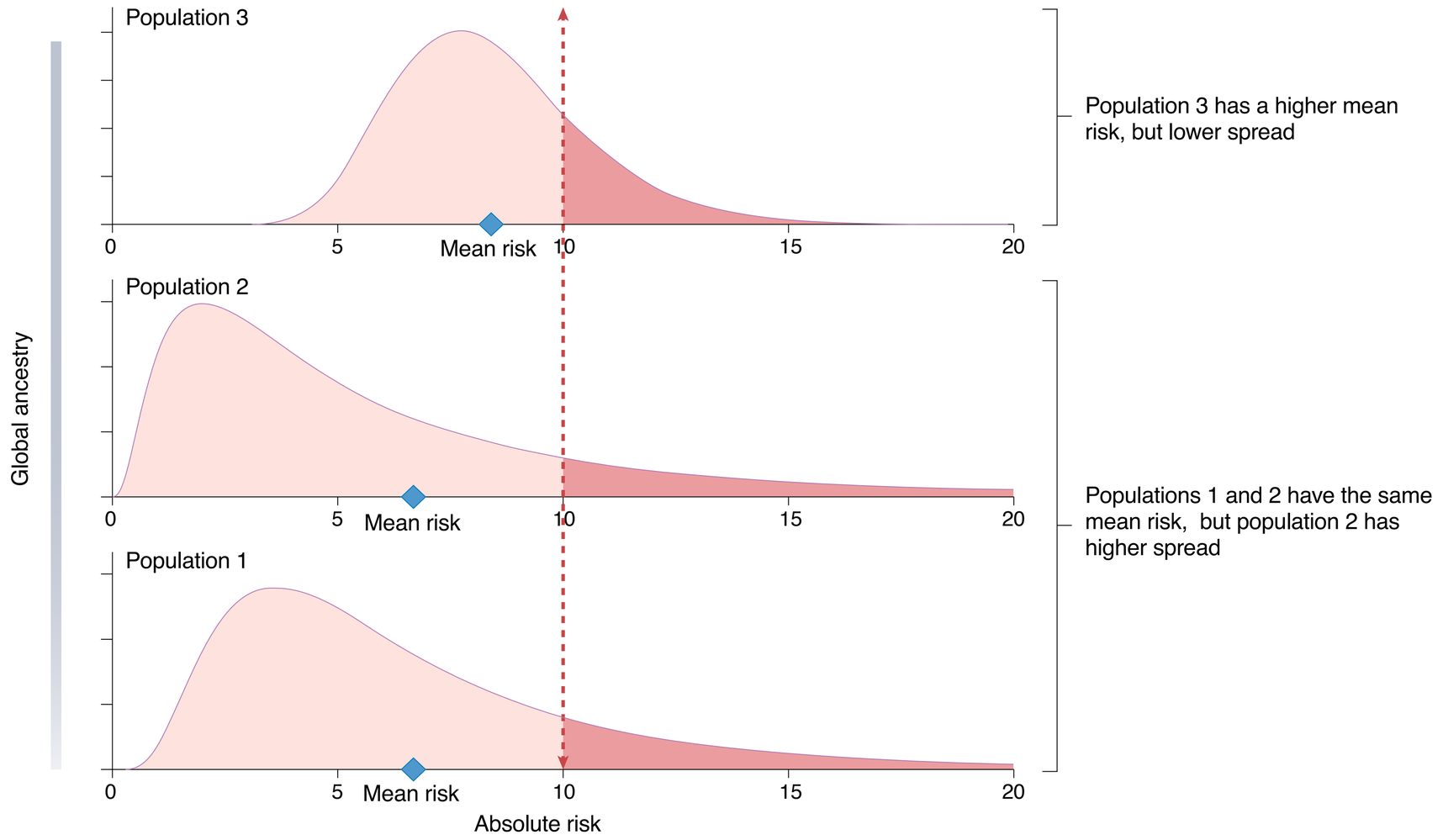
Kachuri et al. Nature Review Genetics (2023)

Accurate PGS requires fine-tuning within pop

High-risk group depends on the mean and spread of risk distribution

Differences in mean risk may or may not be explained by known risk factors and pre-existing conditions

Spread within a population depends on the distribution of known risk factors



Kachuri et al. Nature Review Genetics (2023)

Today's lecture

1. Why do we want to build a polygenic score model?
2. What is a polygenic score model?
3. How to estimate a reliable PGS model by variable selection

PGS Catalog: You may not need to train everything from the scratch

https://www.pgscatalog.org/