Demystifying Polygenic Scores

Workshop

BioinfoHub at the Center for Molecular Medicine (CMM)

Clara Albiñana

27/03/2025

About me

Clara Albiñana, PhD

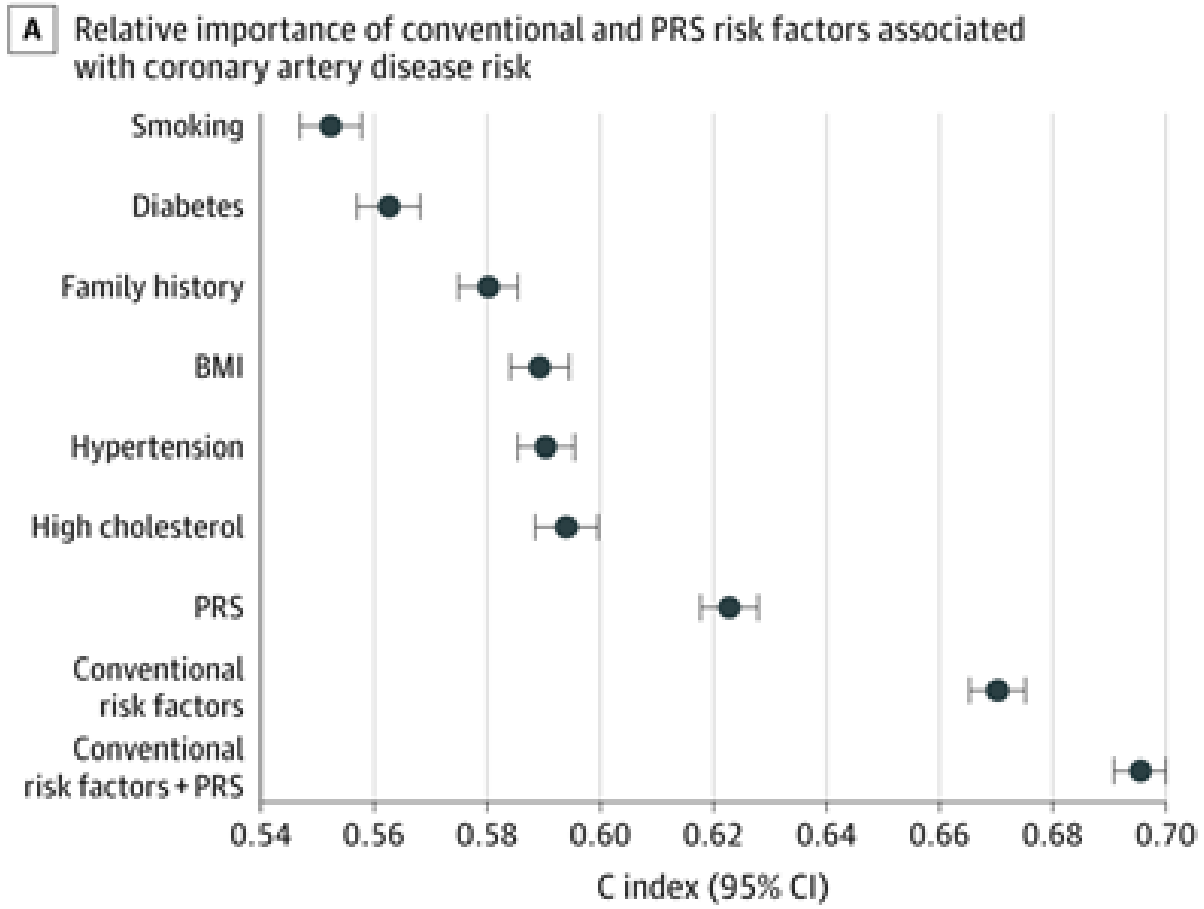
Postdoctoral Researcher at the Big Data Institute, University of Oxford

Aarhus University

- LDpred (Vilhjalmsson et al., 2015)
- LDpred2 (Privé et al., 2020)
- MetaPGS (Albiñana et al. 2021) and MultiPGS (Albiñana et al. 2023)

Table of contents

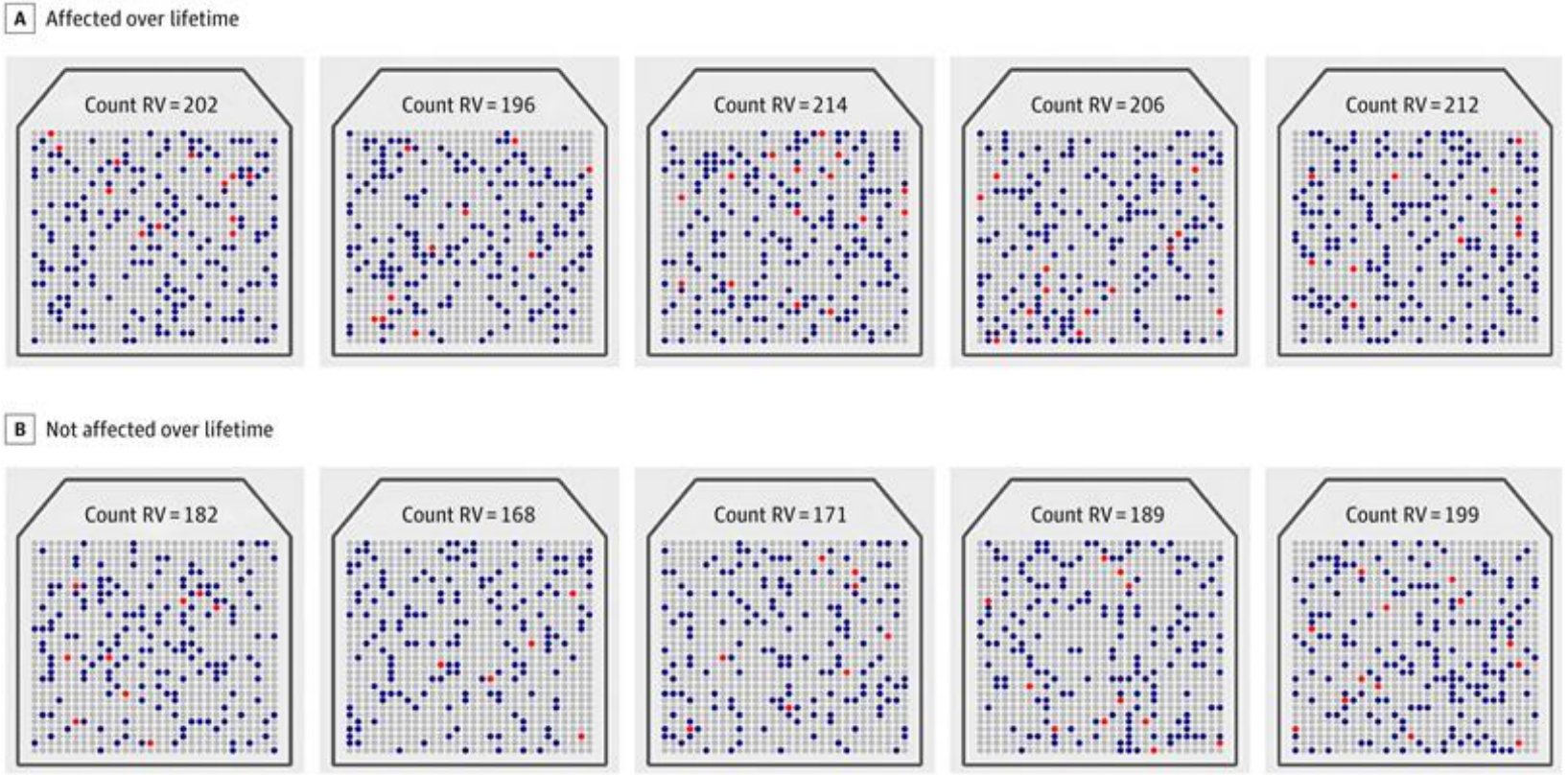
Genetic risk is only one of multiple factors that cause diseases



From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. Wray et al. 2020 JAMA Psychiatry

What is a PGS?

10 individuals with 900 common DNA locations that contribute to risk of disease



	0 copies of risk allele
	1 copy of risk allele
	2 copies of risk allele

From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. Wray et al. 2020 JAMA Psychiatry

Key concepts

1) Not all genetic variants contribute to disease risk -> identify causal variants

2) Variants do not contribute equally to disease risk -> estimate weights

$$PGS_i = \sum_{j=1}^{M} \hat{\beta}_j \, G_{ij}$$

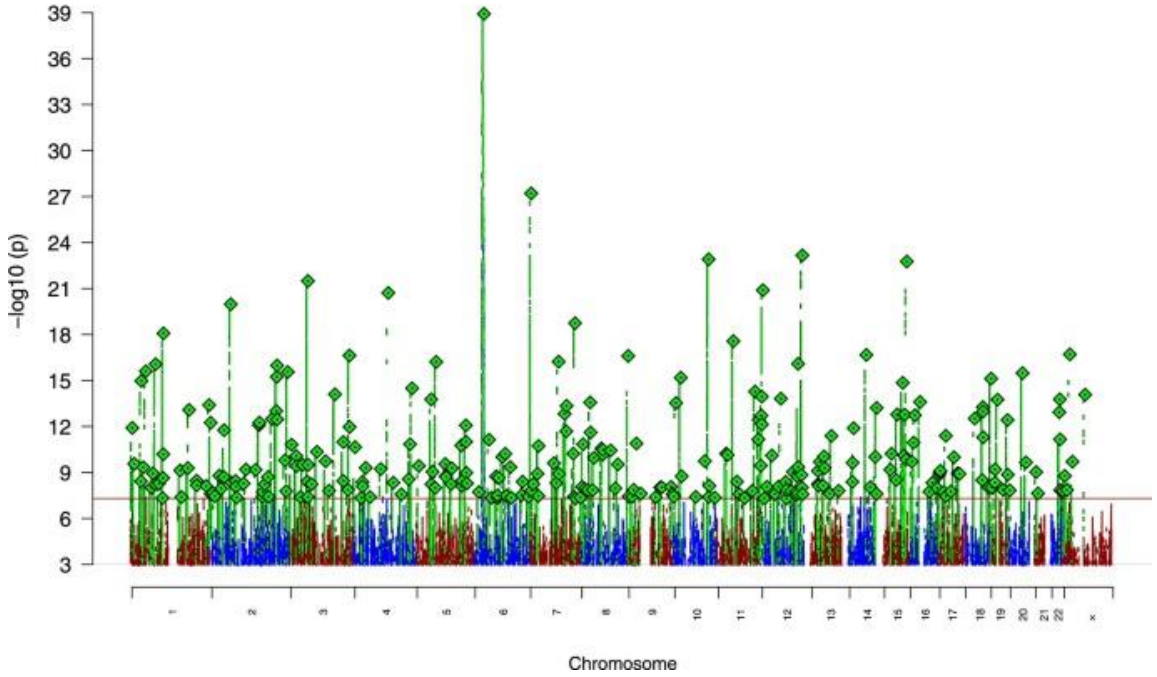
The optimal selection of variants (M) and the weights associated with them (β) is an active area of research.

Simplest approach - Pvalue thresholding

Weights can be effect sizes from a genome-wide association study (GWAS)

SCHIZOPHRENIA
76,755 individuals with schizophrenia and 243,649 control
Trubetskoy et al. 2022

$$SCZ\ PGS_i = \sum_{j=1}^{M} \hat{\beta}_j\ G_{ij}$$

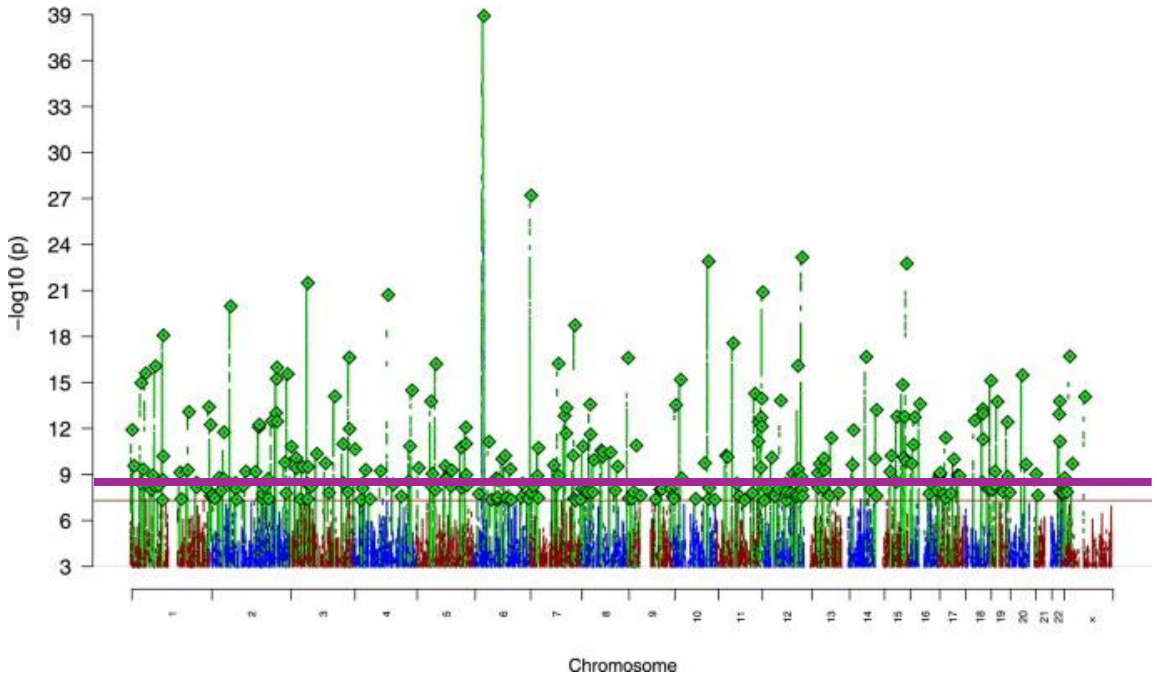


Simplest approach - Pvalue thresholding

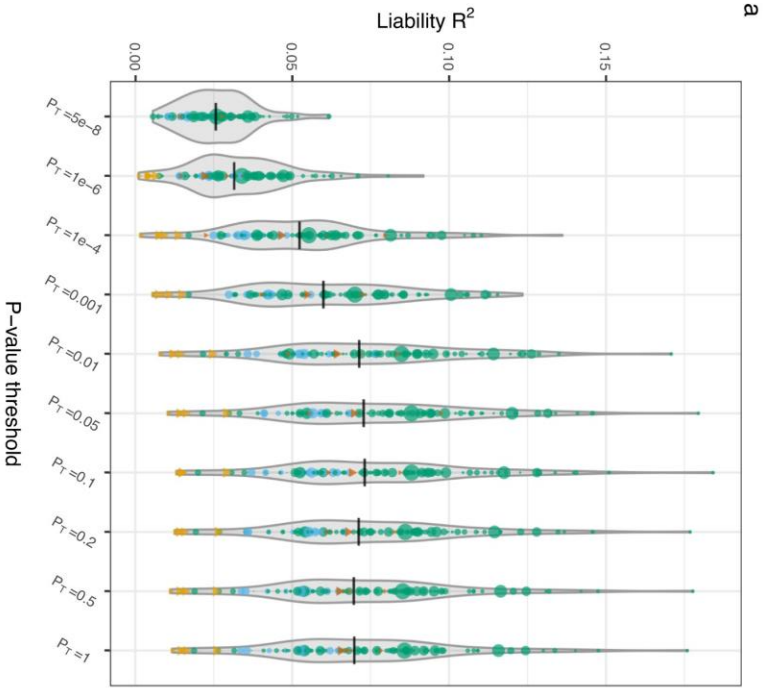
Weights can be effect sizes from a genome-wide association study (GWAS)

SCHIZOPHRENIA
76,755 individuals with schizophrenia and 243,649 control
Trubetskoy et al. 2022

$$SCZ\ PGS_i = \sum_{j=1}^{Mthreshold} \hat{\beta}_j\ G_{ij}$$



pvalue threshold
5e-8
1e-6
1e-4
1e-3
0.01
0.05
0.1
0.2
0.5
1



Complex trait genetics

Mendelian trait

Polygenic trait

1 causal genetic variant
100% penetrance

Ear lobe attachment
Hitchhiker's thumb
Cystic fibrosis

Thousands of causal genetic variants
Partial penetrance

Mental disorders
Behavioral traits
Height

Complex trait genetics

Mendelian trait **Polygenic trait**

1 causal genetic variant
100% penetrance

Ear lobe attachment
Hitchhiker's thumb
Cystic fibrosis

Everything in between

Immune disorders
Blood biomarker levels
etc.

Thousands of causal genetic variants
Partial penetrance

Mental disorders
Behavioral traits
Height

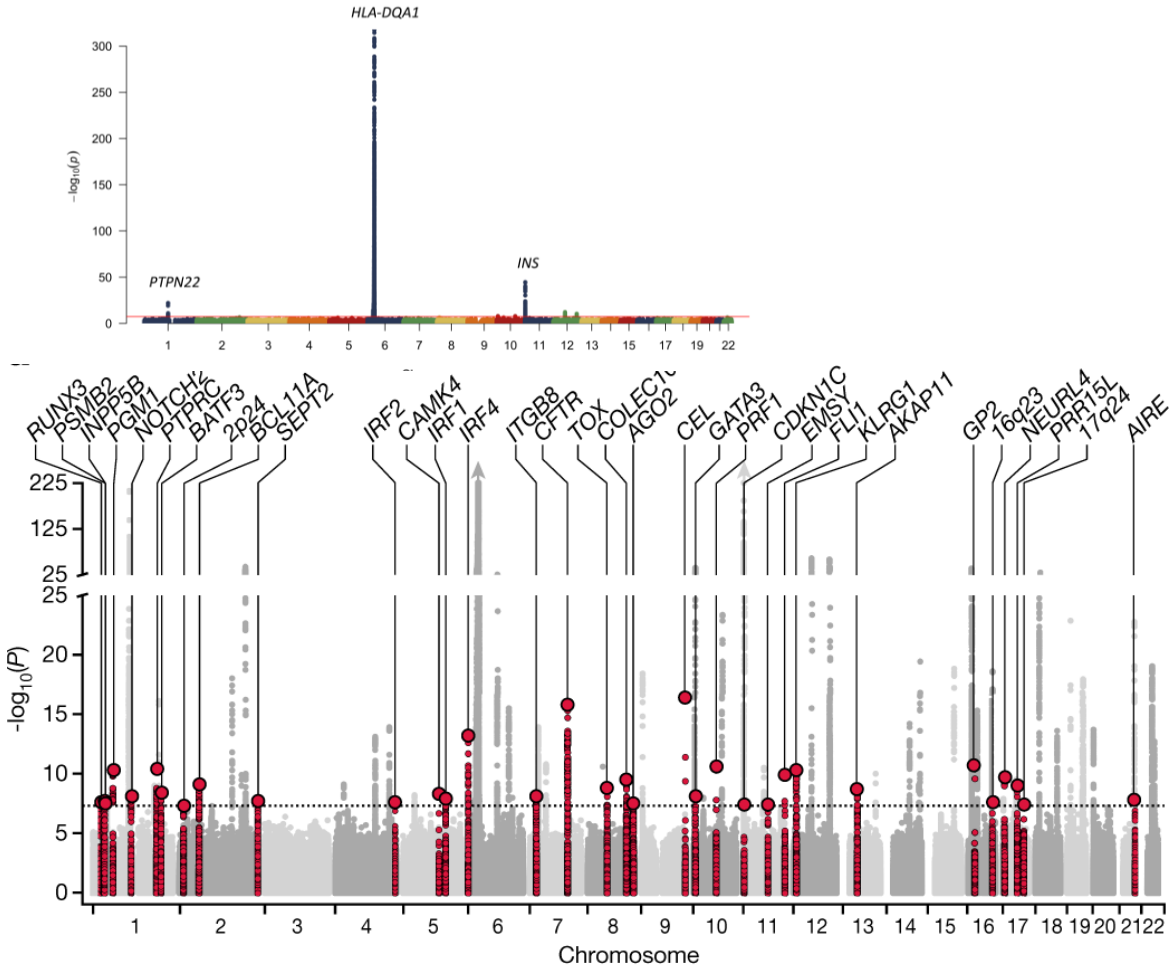
Most traits show a combination of few variants with strong effect and a polygenic background

Genetic architecture of type 1 diabetes vs. Schizophrenia

TYPE 1 DIABETES

18,942 patients with T1D and 501,638 control

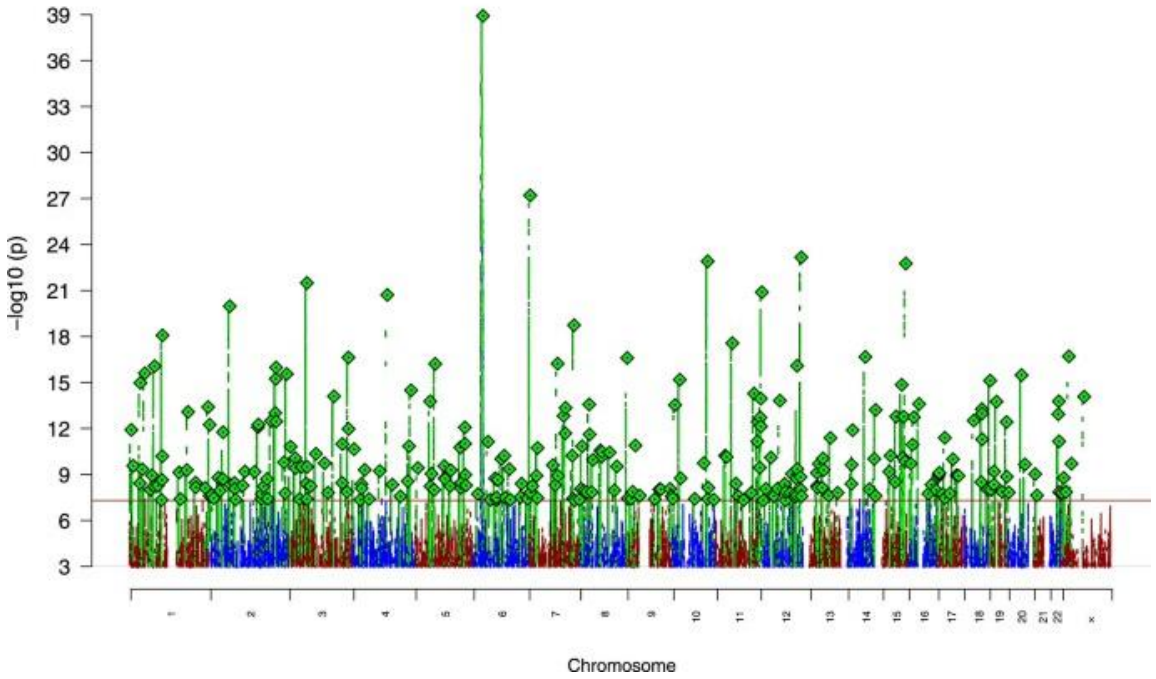
Chiou et al. 2021



SCHIZOPHRENIA

76,755 individuals with schizophrenia and 243,649 control

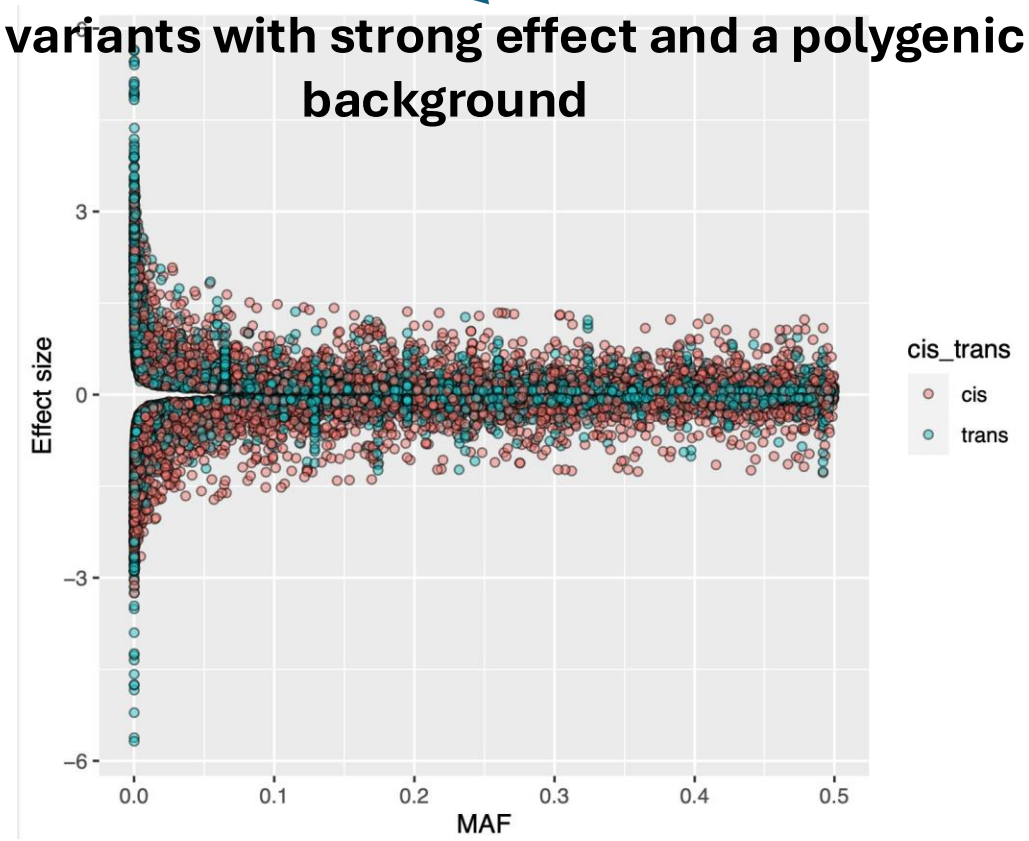
Trubetskoy et al. 2022



Relationship between allele frequency and effect size

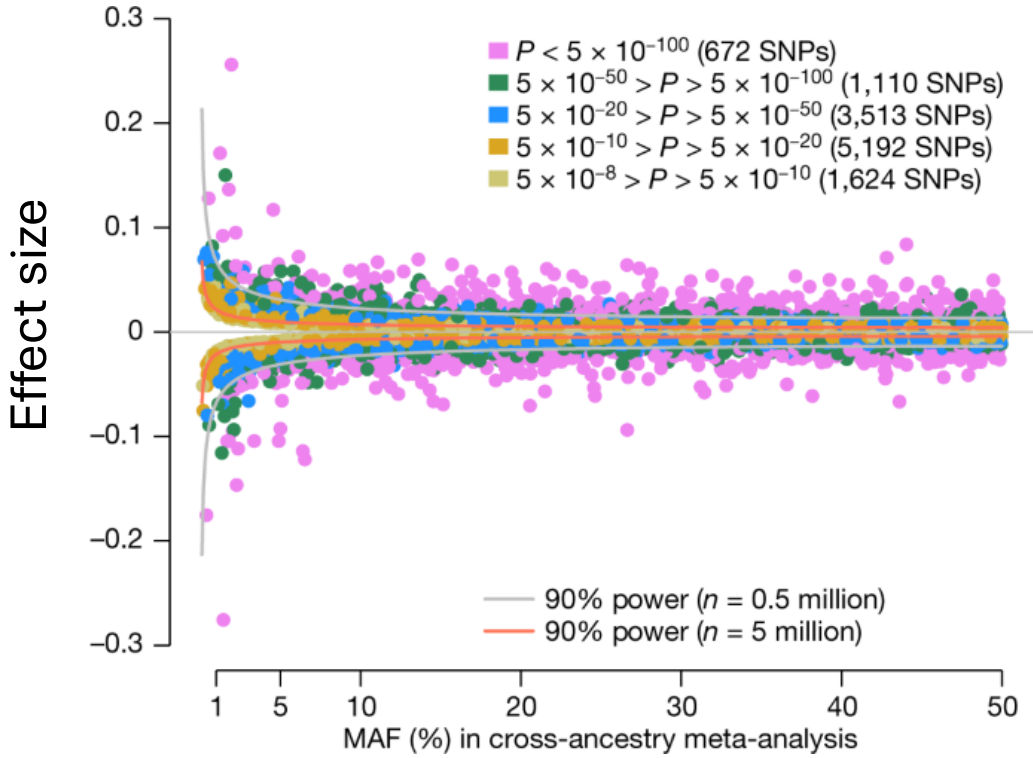
Mendelian trait **Polygenic trait**

A few variants with strong effect and a polygenic background

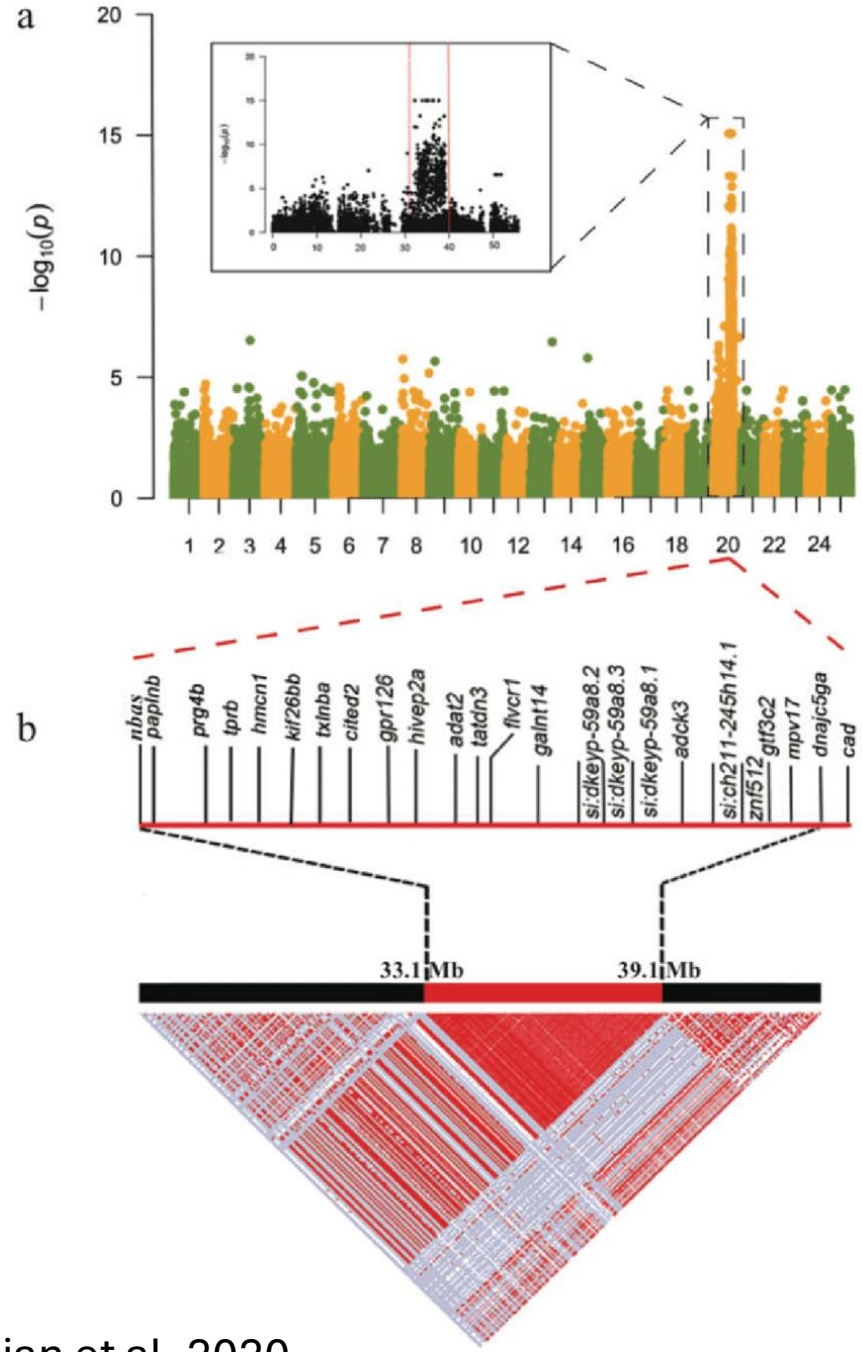


Protein levels in blood Ferkingstad et al. 2021

Most variants have very small effects



Height Yengo et al. 2022



Bian et al. 2020

Removing other type of noise - Clumping

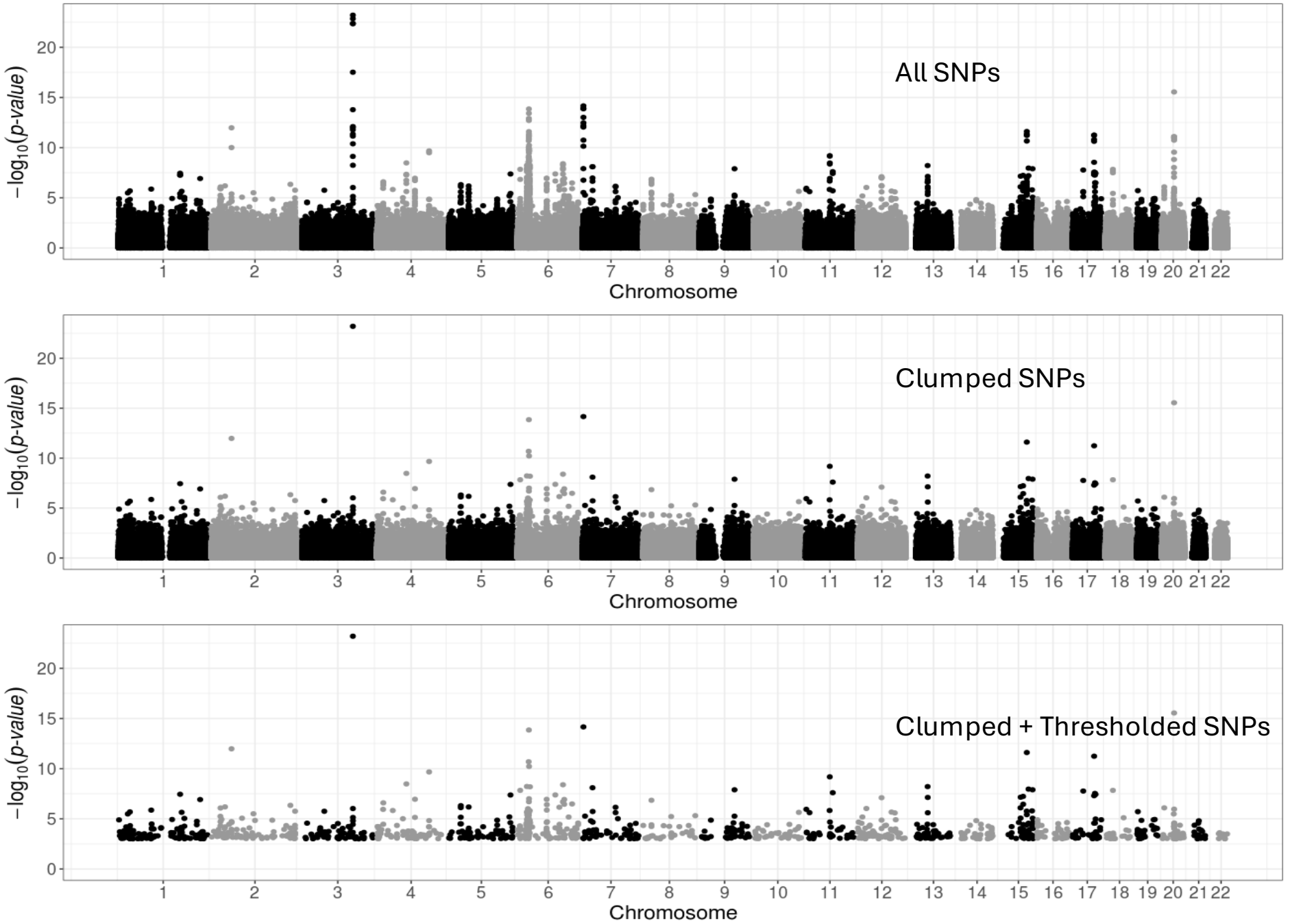
Linkage Disequilibrium (LD) implies that causal genetic variants are indistinguishable from nearby correlated variants

Clumping algorithms:

1. Define a window of SNPs, e.g. 1Mbp
2. Calculate correlation between all SNPs in window
3. Group ones with correlation greater than, e.g. 0.1
4. Rank the clumped SNPs by lowest p-value
5. Select lowest p-value SNP to represent group
6. Do until no more SNPs left

Clumping + Thresholding

$$PGS_i = \sum_{j=1}^{M} \hat{\beta}_j \, G_{ij}$$



LDpred ~ Modelling LD

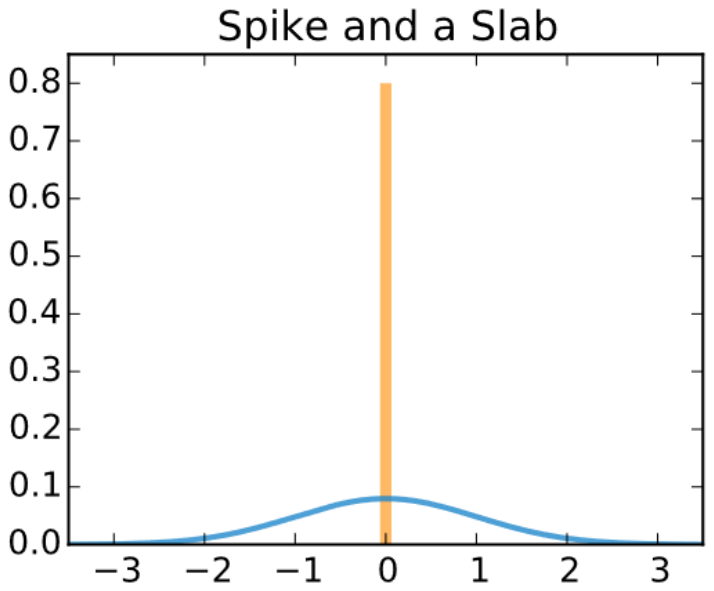
Clumping and thresholding is a bit of a “brute force” method.

Bayesian approach that modelS the polygenicity and LD by assuming prior:

$$\beta_j \sim \begin{cases} N\left(0, \frac{h^2_{snp}}{Mp}\right) & with\ probability\ p \\ 0 & with\ probability\ 1 - p \end{cases}$$

- Where h^2_{snp} is the heritability captured by the SNPs,
- $\boldsymbol{M}$ is the number of SNPs,
- p the proportion of causal SNPs.
- This prior is commonly referred to as the
- spike-and-slab prior

- Note: p = 1 is the infinitesimal model



Practical

STEP1: SIMULATE THE PHENOTYPE from real genotype data.

Create TRAINING and TEST sets

STEP2: GWAS

STEP3: Compare PGS methods for prediction

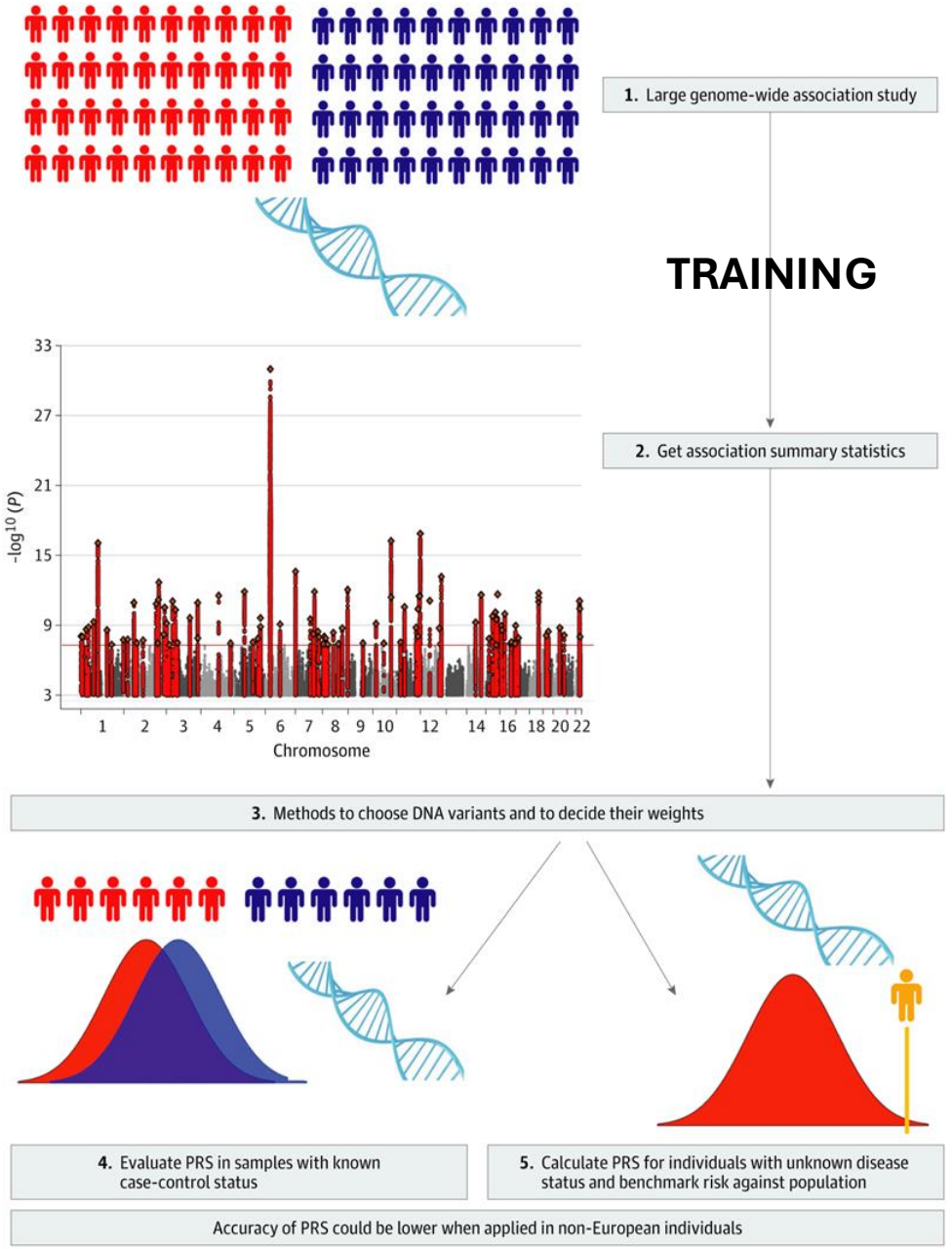
ALL SNPS

Thresholding

Clumping

Clumping + Thresholding

LDpred



Wray et al. 2020 JAMA Psychiatry