HELMHOLTZ
MUNICH

Genetic architecture of complex traits and Polygenicity

Ozvan Bocher and Ana Arruda
December 6, 2022

Agenda

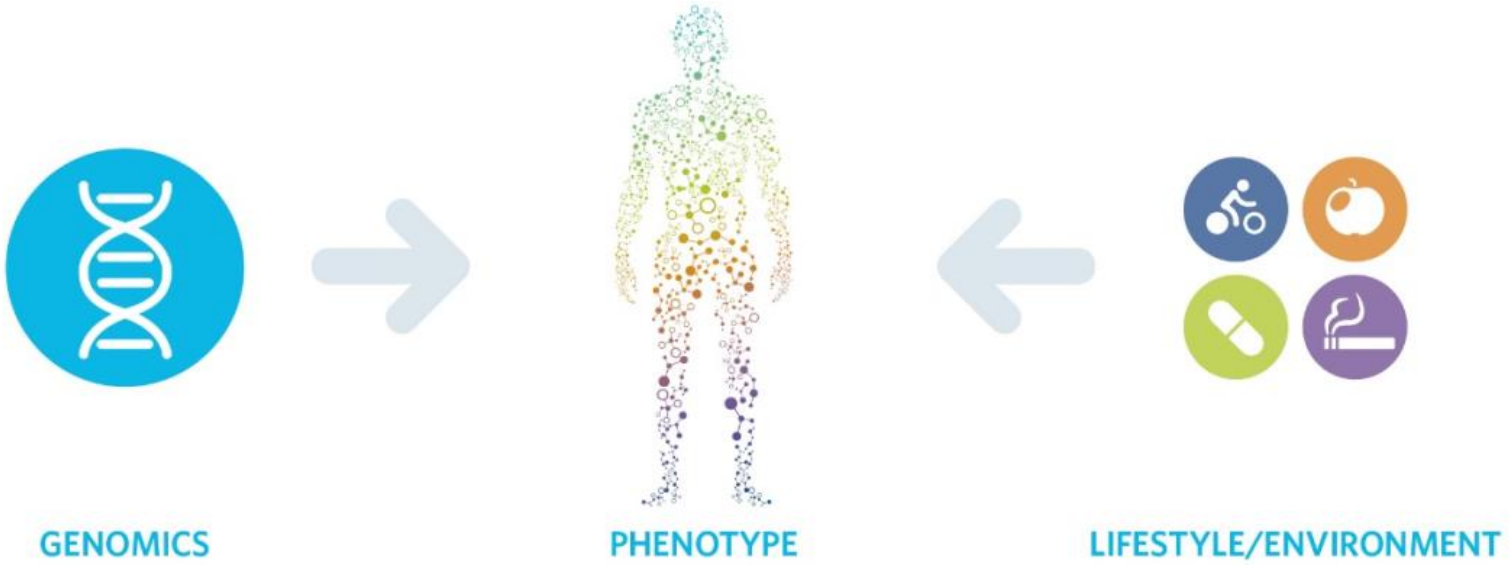
1. Genetics overview
2. Linkage disequilibrium
3. Polygenic Scores

1

Complex traits and genetic overview

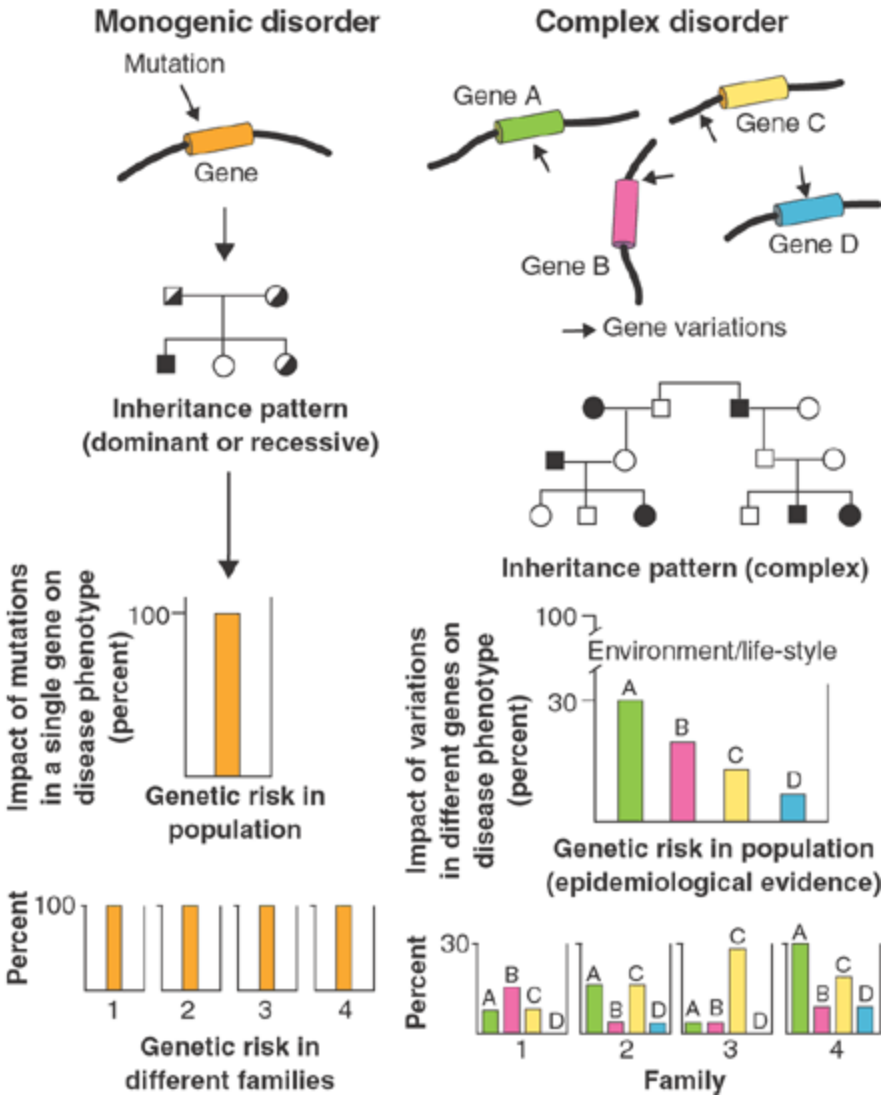
Complex traits

- Complex traits = interaction between (often many) genetic and environmental factors



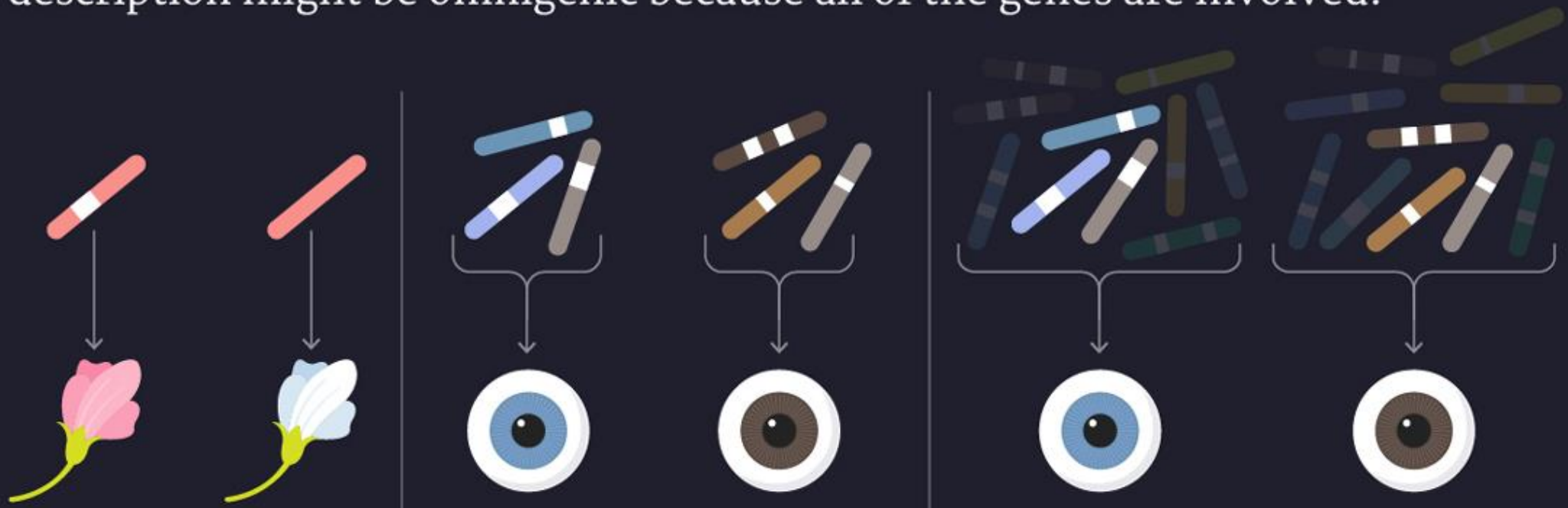
- Examples: body shape, type 2 diabetes, Alzheimer's disease…
- Complex diseases tend to be common
  → Tool of choice = GWAS

Complex traits



How Many Genes Are at Work?

Simple traits may be controlled by just one gene (monogenic). More complex traits are usually considered polygenic, but a new theory suggests that a better description might be omnigenic because all of the genes are involved.



Monogenic

A single gene gives rise to a trait.

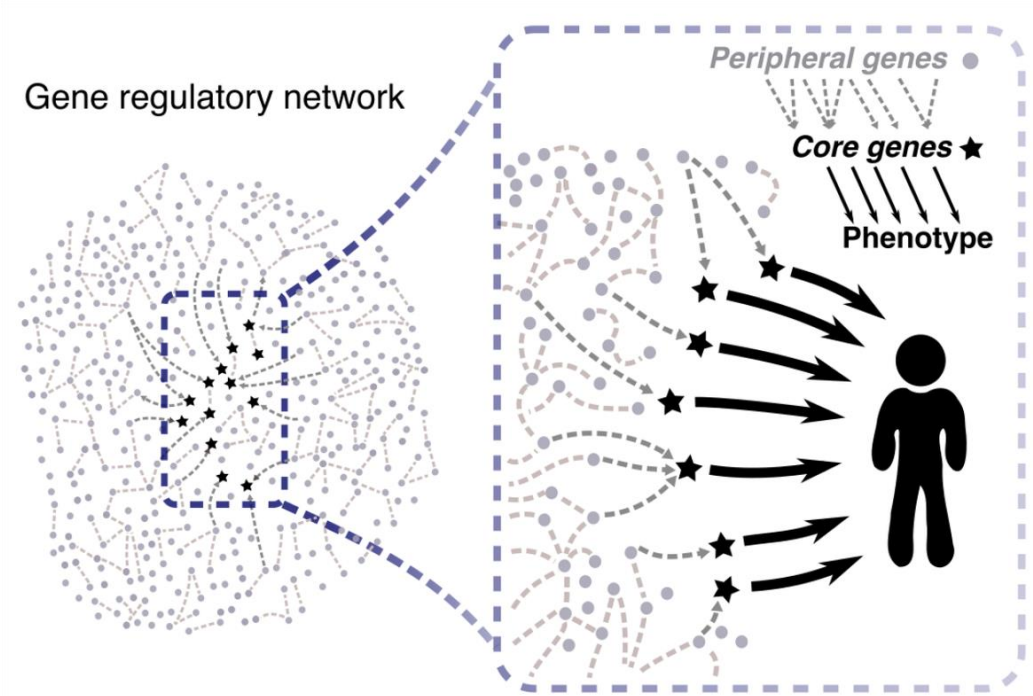
Polygenic

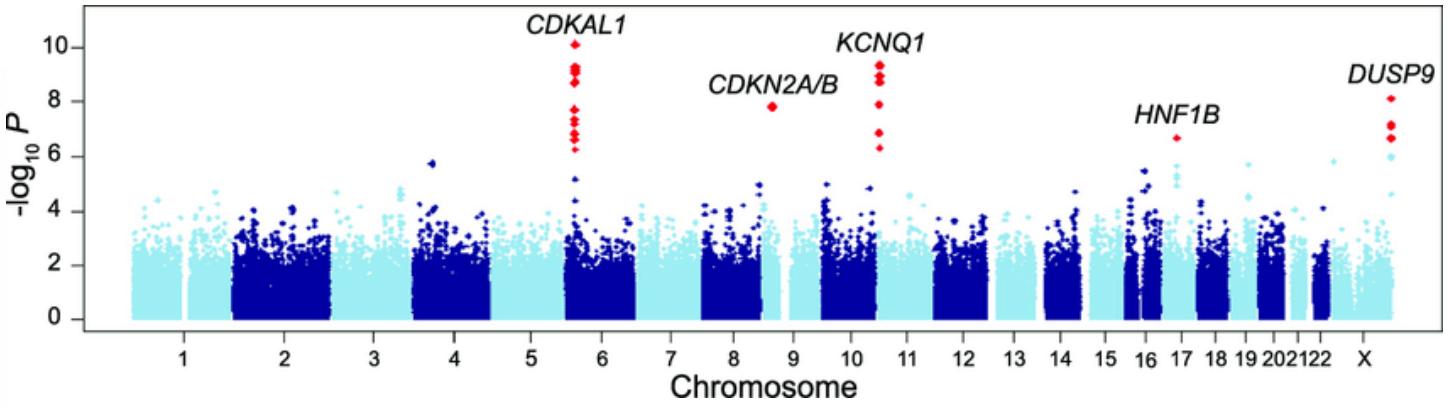
A handful of genes jointly give rise to a trait.

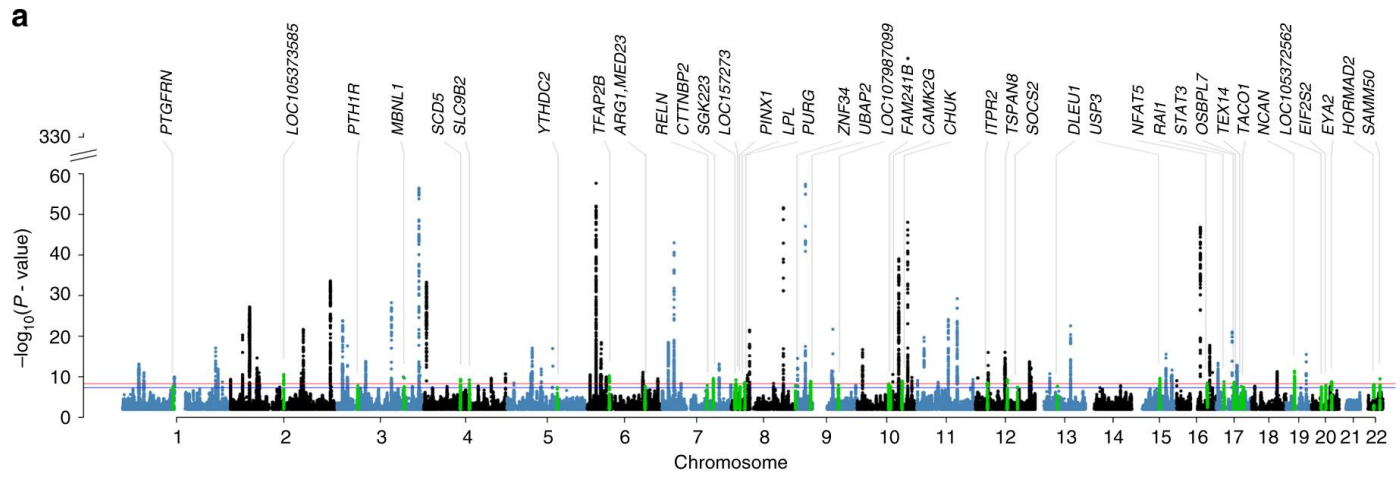
Omnigenic

A few core genes are essential but all the genes are involved.

Omnigenic vs Polygenic model







Genome-Wide Association Studies (GWAS)

- Is there an association between the **phenotype** (disease, continuous trait) and the **genotype** ?

$$\boldsymbol{phenotype} \sim \boldsymbol{\beta} \times \boldsymbol{genotype} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix} \qquad \begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix} \qquad \begin{bmatrix} 1 \\ \vdots \\ 2 \end{bmatrix}$$

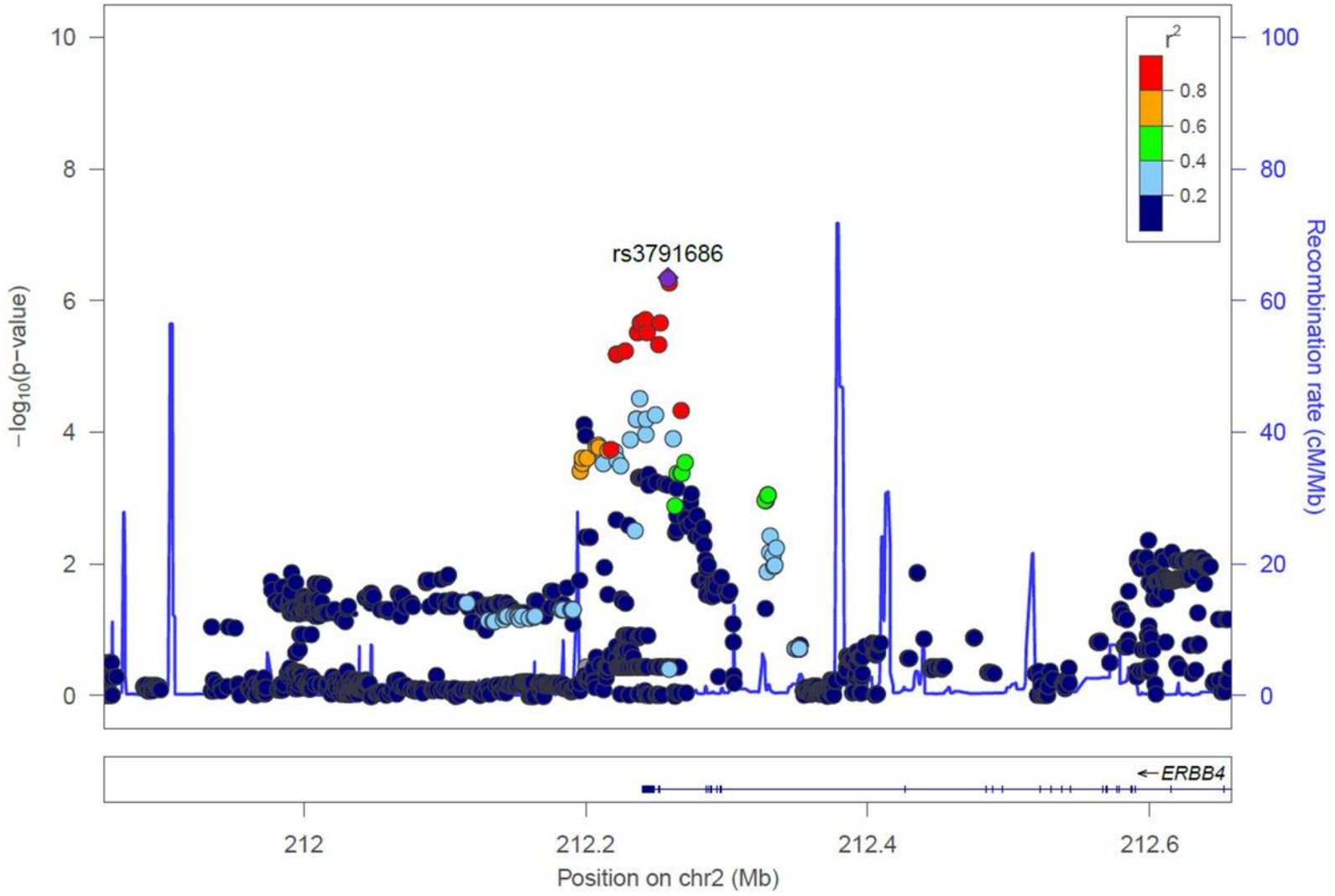
$= \{0,1\}$ (case-control)
$\in \mathbb{R}$ (quantitative) $\sim \mathcal{N}(0,1)$

$= \{0,1,2\}$ (genotype, directly typed)
$\in [0,2]$ (dosage, imputed)

$$\begin{bmatrix} 0.965 \\ \vdots \\ 1.816 \end{bmatrix}$$

- For each variant, association test → if $p \leq 5 \cdot 10^{-8}$: variant significantly associated
- Estimation of the effect of the variants: β or Odds Ratio (OR)

GWAS – regional association plot



2

Linkage disequilibrium

Linkage disequilibrium (LD)

- **Non-random** association of alleles at different loci

  → Correlation between the genetic variants
  → Different between the populations

- Mechanisms include selection, genomic recombination, genetic drift…
- Co-occurrence of alleles:

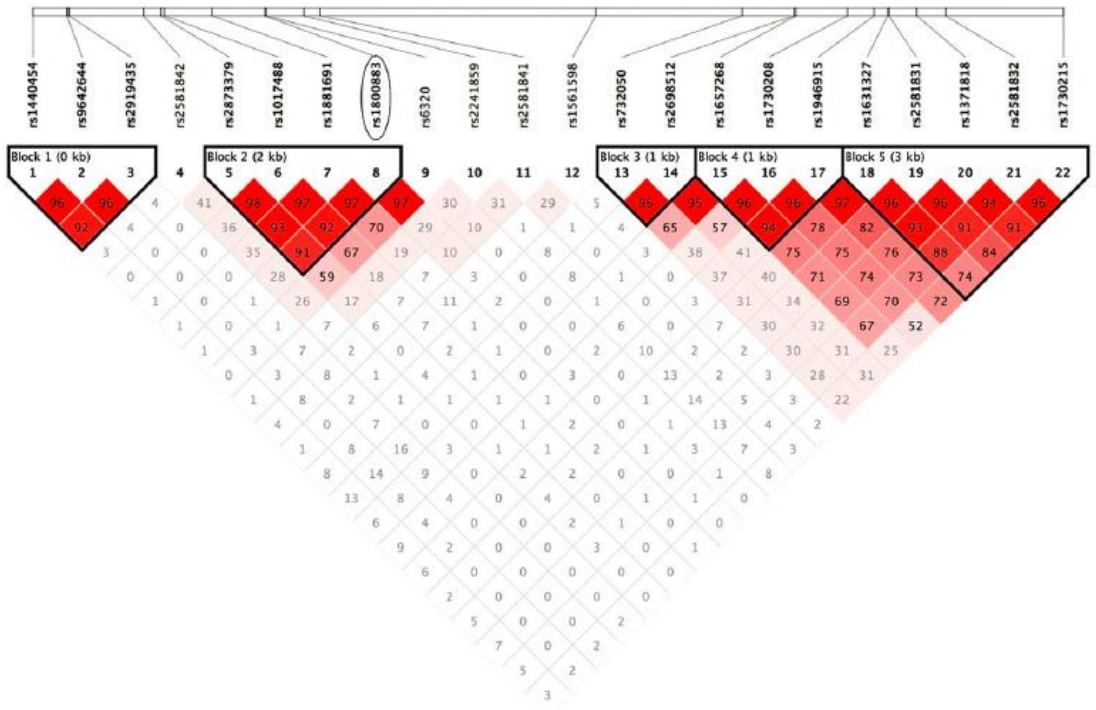
$$D_{AB} = p_{AB} - p_A p_B$$

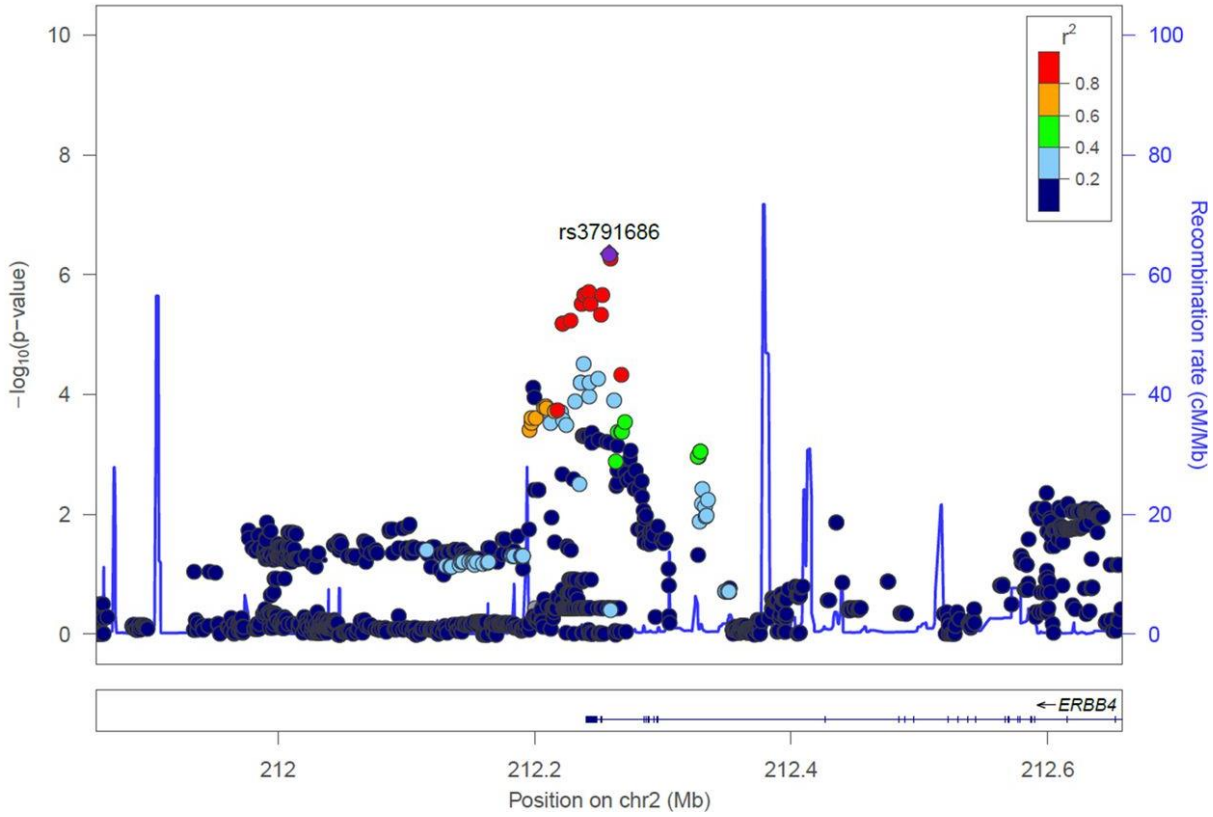
  → D = coefficient of linkage disequilibrium
  → If D = 0: linkage equilibrium

- $\boldsymbol{r^2}$: squared coefficient of correlation

$$r^2 = \frac{D_{AB}}{p_A \cdot (1 - p_A) \cdot p_B \cdot (1 - p_B)}$$

LD plots





→ Find the variants in high LD
→ Identify blocks of LD

HELMHOLTZ MUNICH

Guan et al. Sci Rep. 2016

LD pruning

- Need to be taken into account in genetic studies

  → Number of significant variants ≠ number of independent signals
  → Particularly important in Polygenic Scores
  → Different pattern between populations

- Solution = Pruning

  → Select a SNP in a genomic window and remove the correlated SNPs according to r^2
  → Keep only the 'independent' SNPs
  → Can be performed using Plink (example in practical)

3

Polygenic scores

3.1

Introduction

Polygenic scores

- **Polygenic model**: many genetics variants influence complex traits

  → T2D: 338 association signals (*Mahajan et al. Nat Genet. 2022*)
  → Height: >12,000 independent variants (*Yengo et al. Nat. 2022*)

- Try to predict quantitative traits (polygenic scores) and disease risk (polygenic risk scores) based on genetics
- Use estimates from **GWAS**
- With larger sample size for GWAS summary stats → increased predictive power of PGS
- PGS are constructed under an additive model: each copy of the effect allele increases the risk

Polygenic scores (PGS)

Polygenic score for individual *i*

Total number of SNPs included in PGS

$$PGS_i = \sum_{j=1}^{N_{snps}} G_{ij} * \beta_j$$

Genotype at SNP *j* for individual *i*

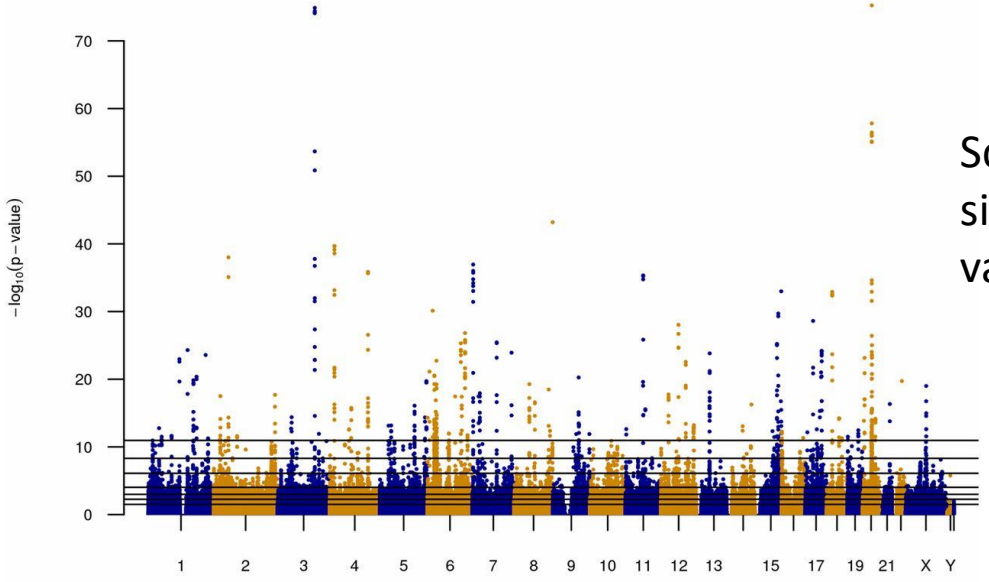
- Coded as 0, 1, 2: counting number of risk allele
- Additive model

Effect of variant *j* on trait

- Estimated in GWAS

- **Scores = sum of effects of many variants** → normally distributed in the populations

Polygenic scores (PGS)

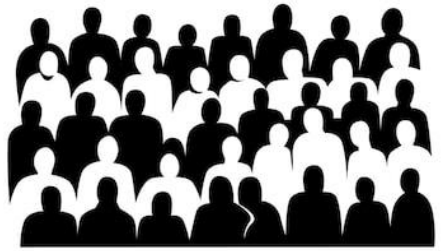


Large GWAS

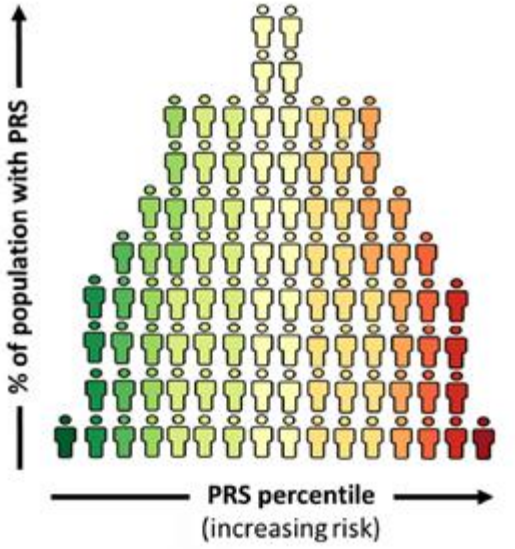
Scores for significant variants

Genetic data

$$PHS_x = \sum_i^n X_i \beta_i$$

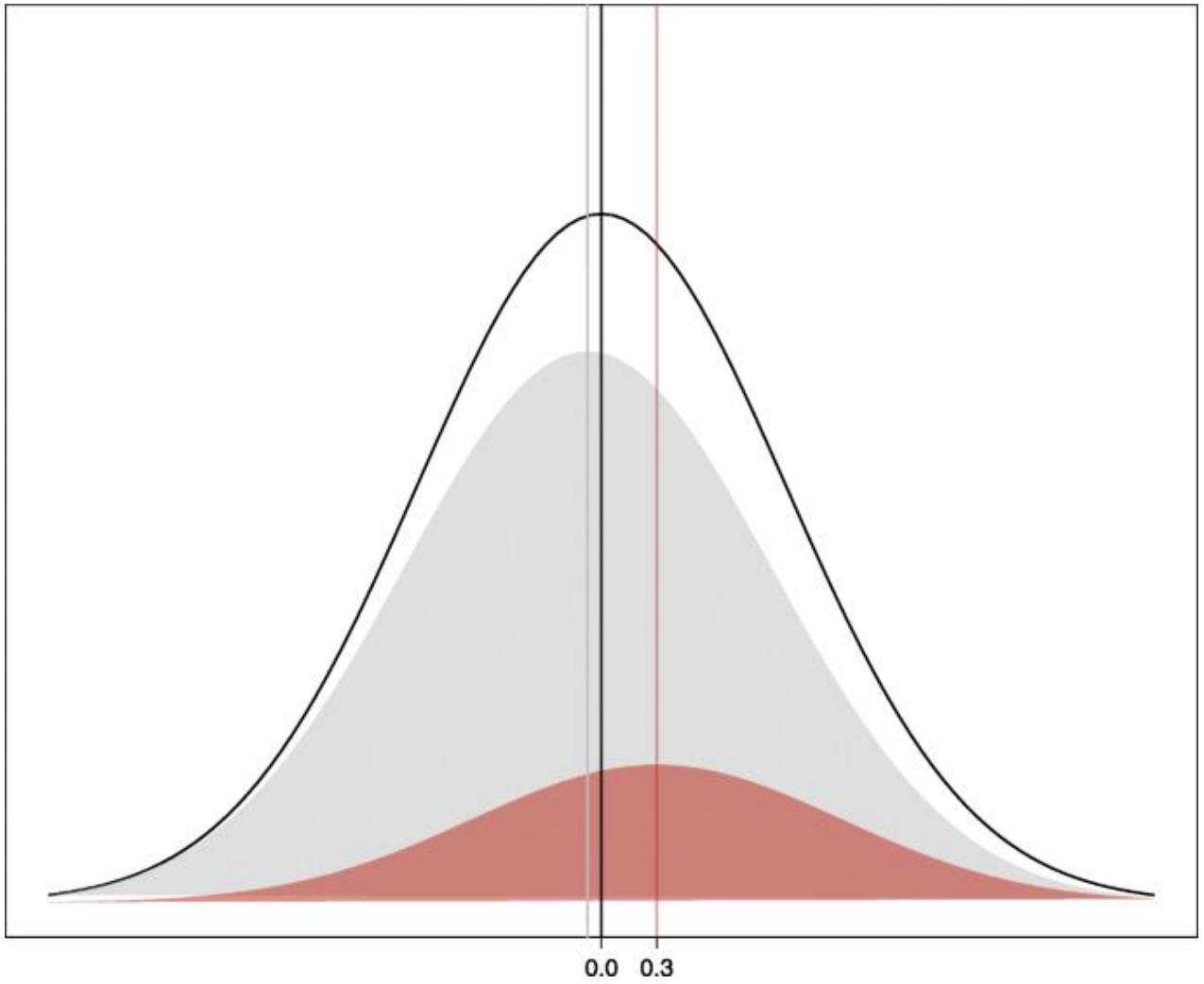


External cohort



Polygenic scores (PGS)

- Grey: score distribution of controls
- Red: score distribution of cases
- Overall mean = 0 (standardized score distribution)
- Amount of shift = discriminative power of PGS

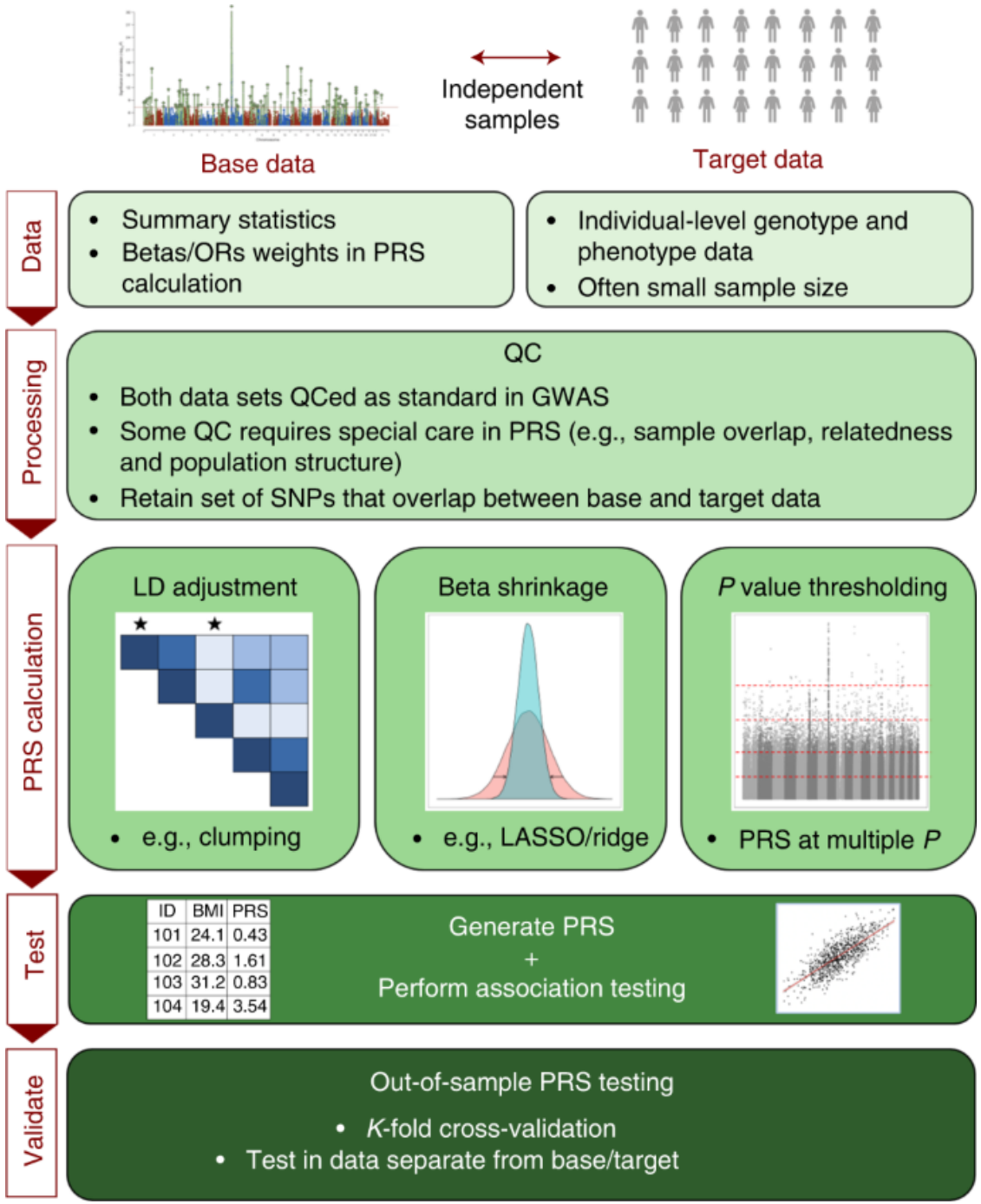


3

Polygenic scores

3.2

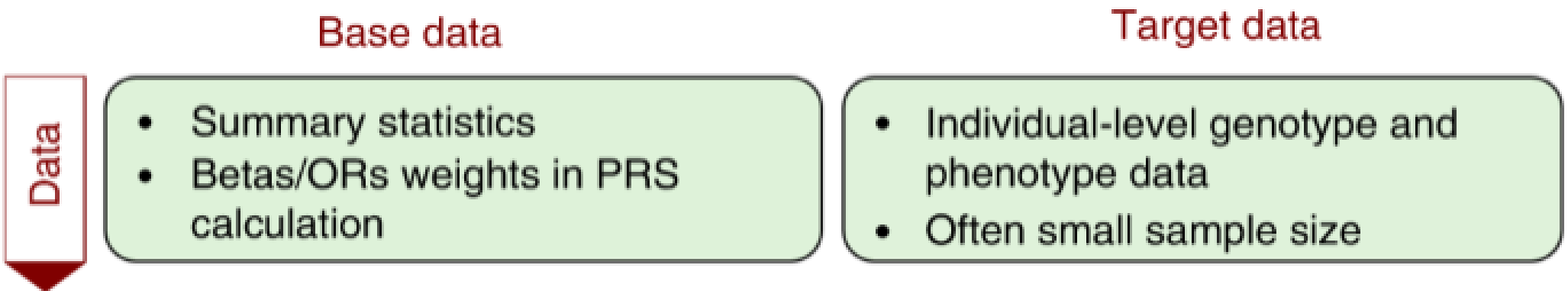
Construction



Choi et al. Nat. prot. 2020

HELMHOLTZ MUNICH

Input data

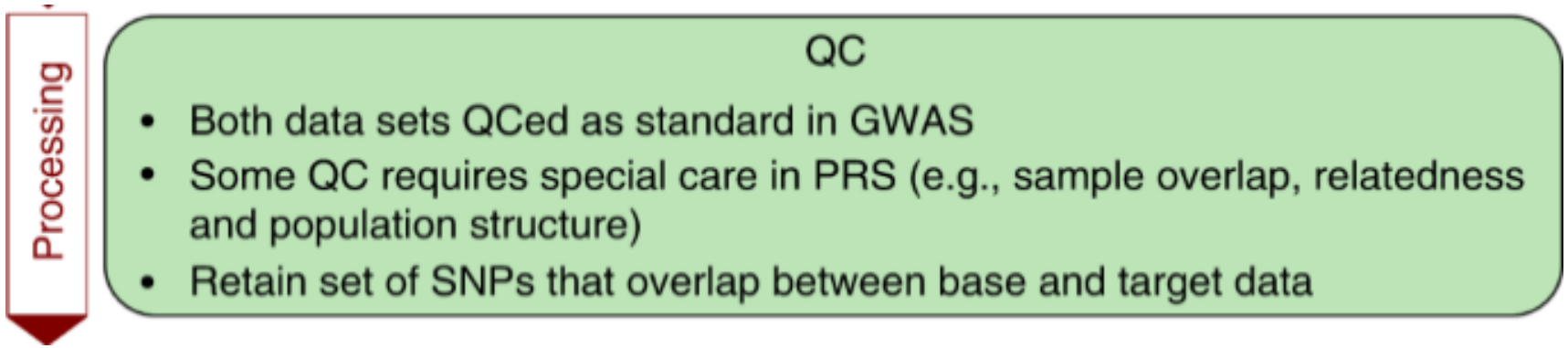


- **Base data** = sample used to estimate parameters for the PGS
  - → Effect sizes of the variants: OR or β
  - → Standard errors
  - → P-values
  - → Obtained from summary statistics, often from large published GWAS (GWAS catalog)

- **Target data** = sample where we will apply the PRS
  - → Individual genotype and phenotype data
  - → Often small sample sizes

- Goal: apply on real patients

Data processing



- Apply the same quality control as standard GWAS
- Only SNPs in base AND target samples
- Sample overlap

  → An overlap between base and target sample could lead to inflation: ‘overfitting’

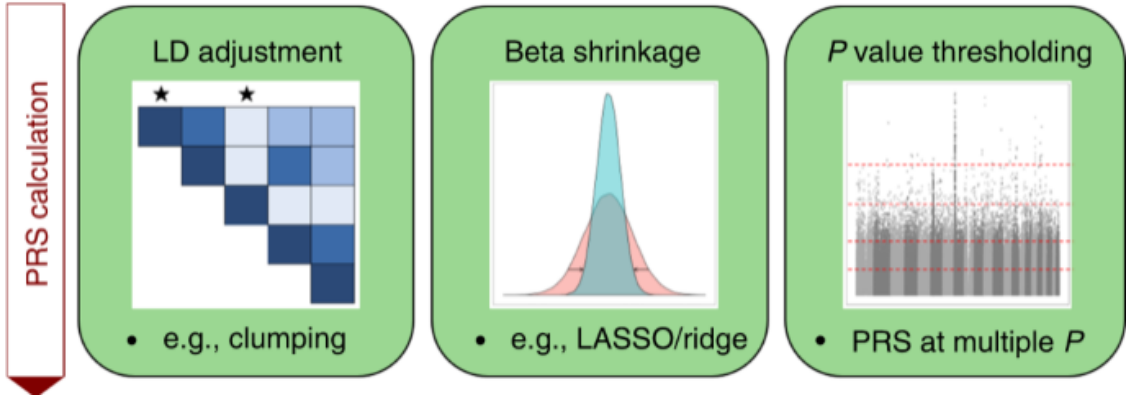
- Homogeneity between base and target samples

  → Hypothesis = samples have the same underlying genetic architecture
  → Also suppose homogeneity in environment

- Population structure

  → Match the ancestry between base and target samples

Selection of variants



- Historically: genetic predictions based on independent variants providing more risk

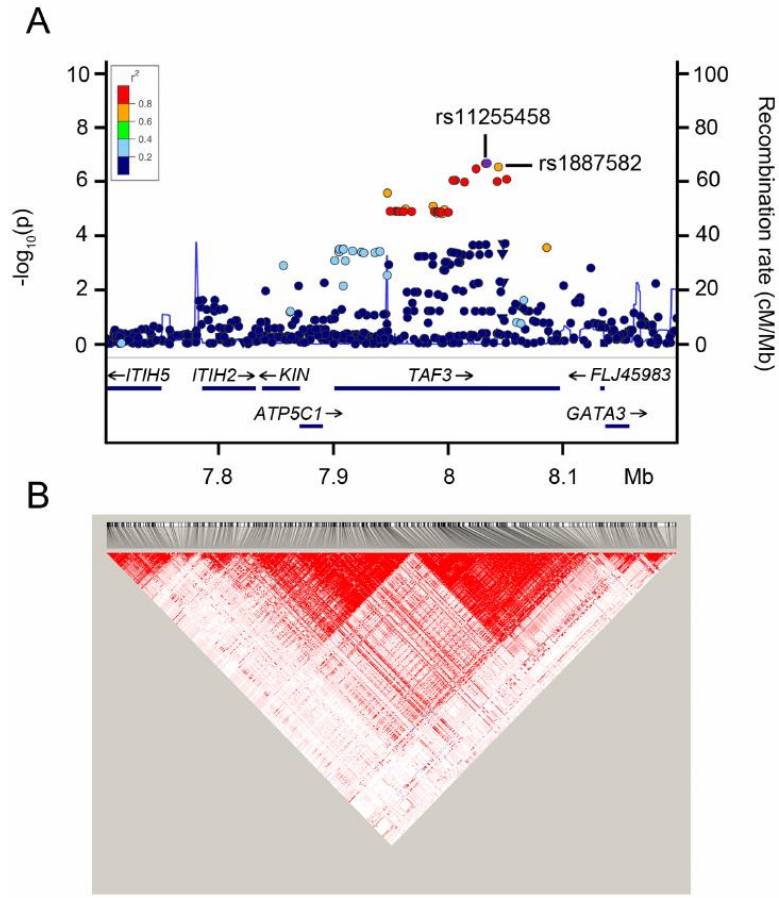
  → Challenging in omnigenic and polygenic models
  → With more power, more peaks appear

- Take all variants into PGS ?

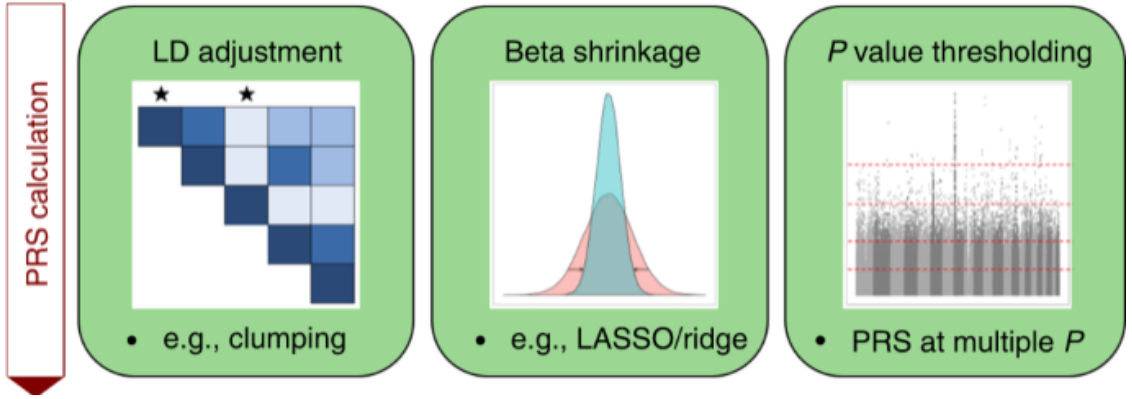
  → LD issue

- Selection of variants influencing complex traits

  → Common practice: select **independent variants**
    → Clumping: pruning with a p-value thresholding
    → No overweighting of high-LD blocks
  → Beta shrinkage



Clumping + Thresholding (C+T)



P-value aware pruning

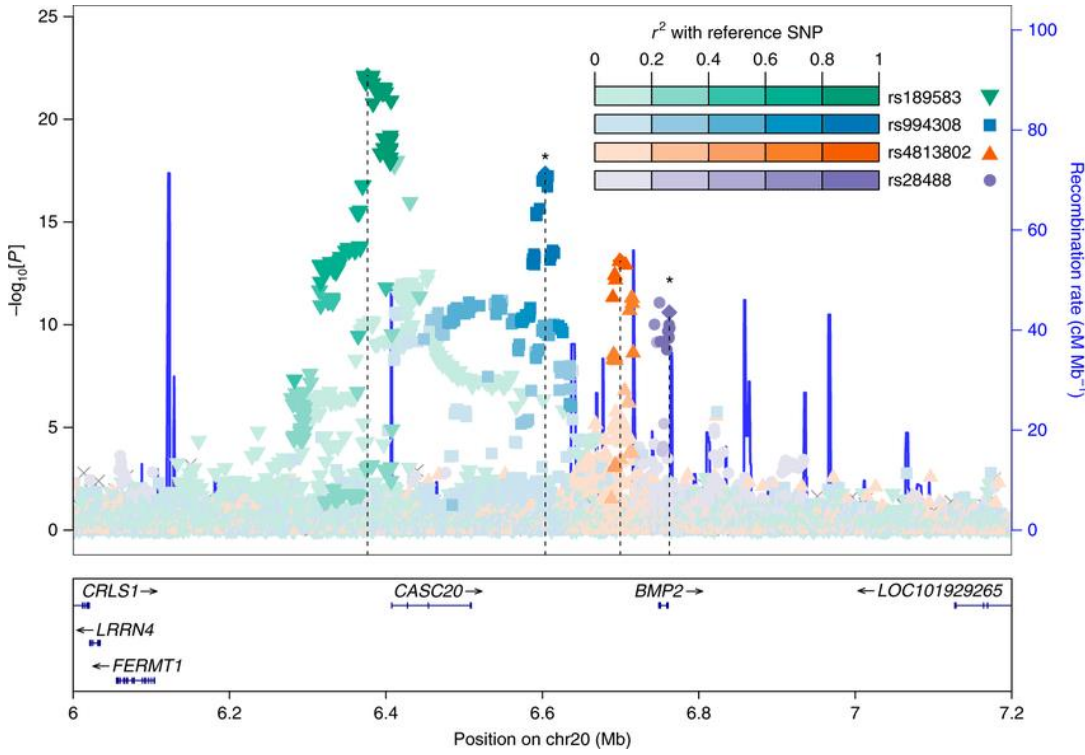
1. Select SNPs with a p-value<threshold
2. Select top associated variant
3. Remove all variants in LD with this SNP

YES

Any other significant variant left in the block ?

NO

Set of pruned and significantly associated variants



HELMHOLTZ MUNICH

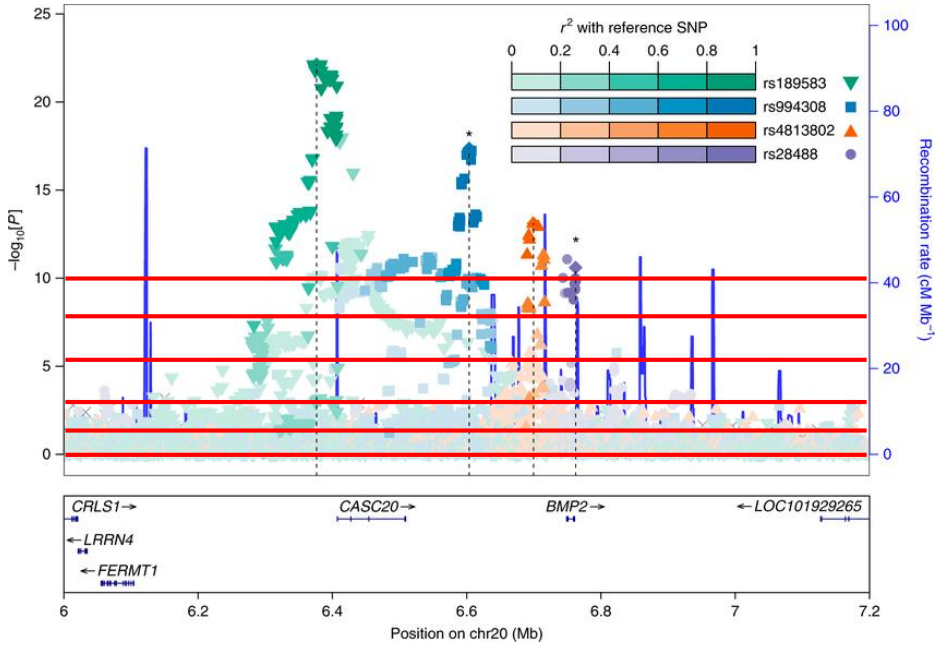
Clumping + Thresholding (C+T)

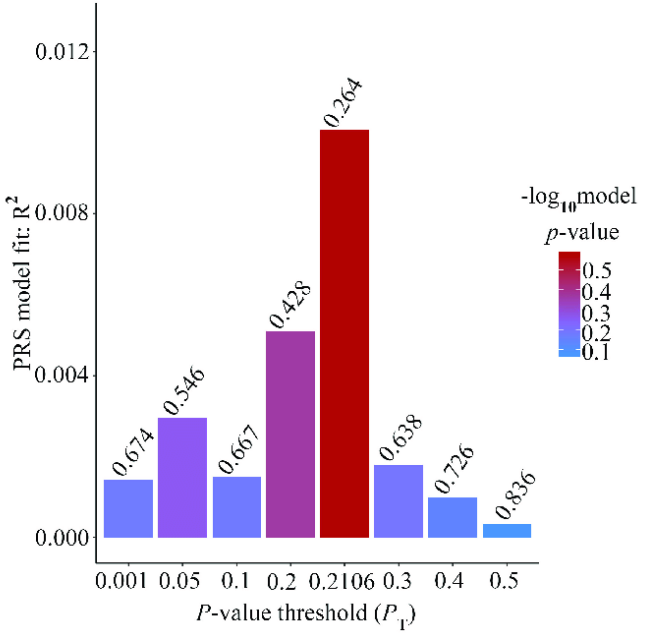
- Which significance threshold to use ?

  → Optimal threshold depends on the trait

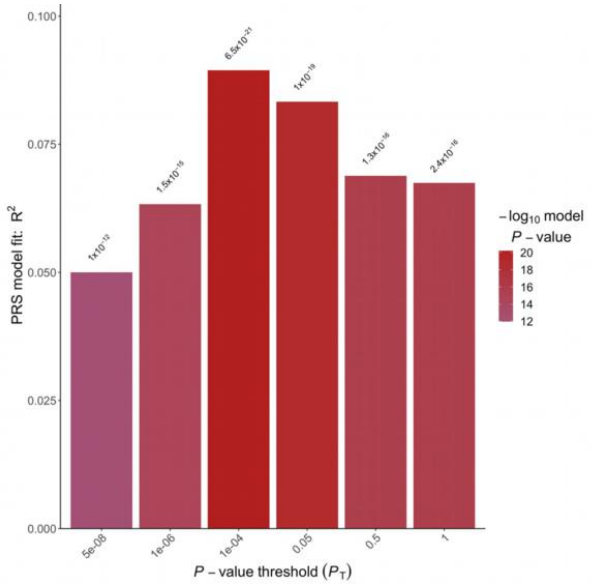
- Unknown beforehand

  → Try multiple values with validation
  → Integrated into software, e.g. PRSice



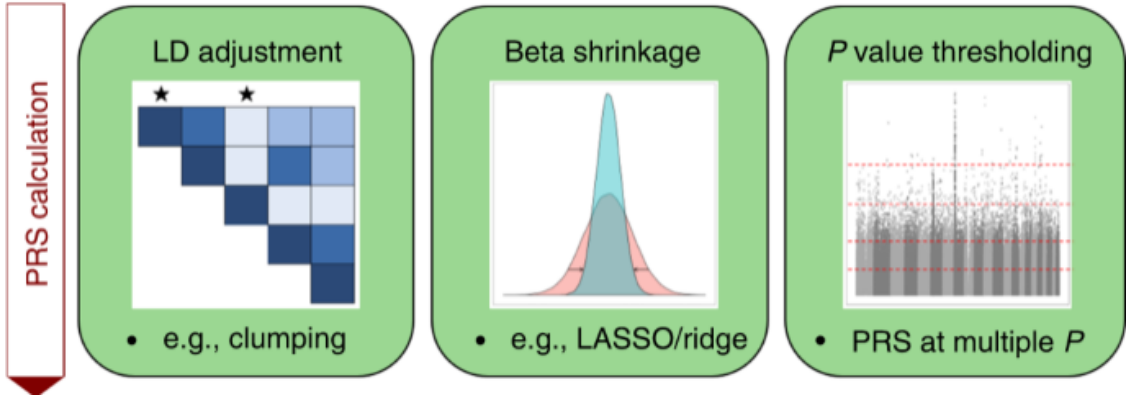


Wang et al. Frontiers in Genetics, July 2019



Maj et al. Frontiers in Cardiovascular Medicine, Feb 2022

Limitations of C+T



- Potential removal of secondary signals
- Based on the p-values but not the **effect sizes**

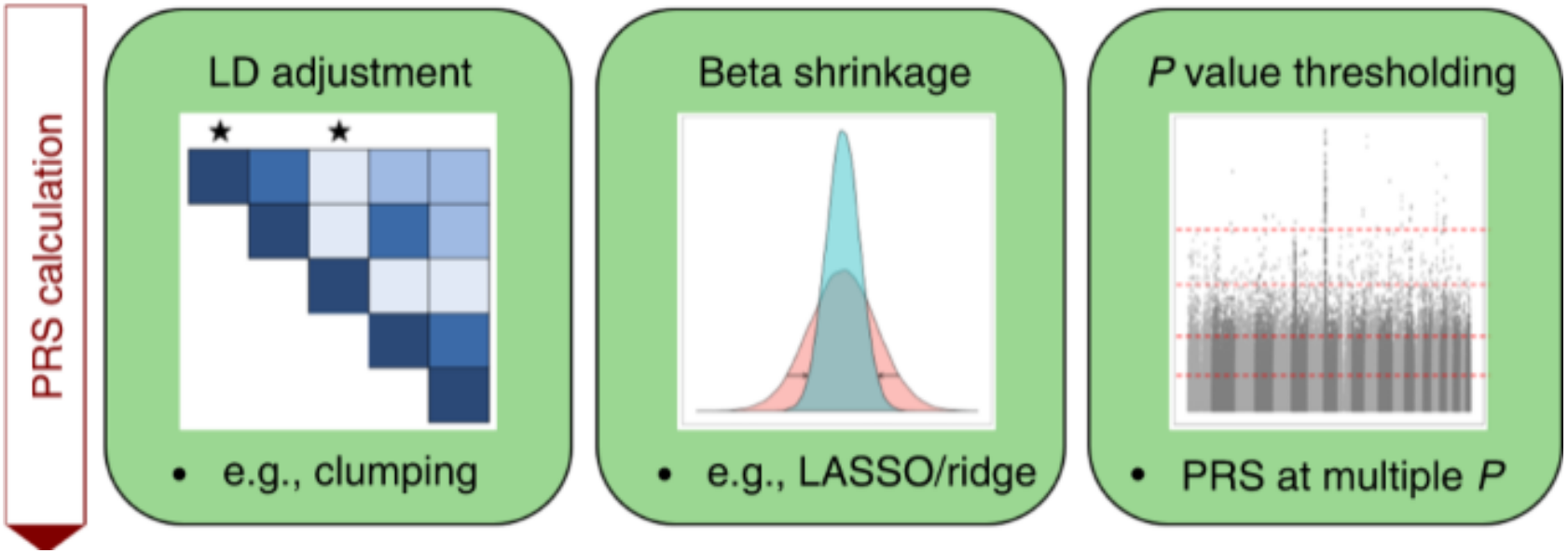
  → The p-value is related to the power of the study
  → Can miss low-effect variants in small sample sizes

- Sample size = still a limiting factors for improved methods
- Example of software: Plink, PRSice

- Ideal model = 'whole-genome' model

  → Account for LD
  → Perform "shrinkage" estimation for association coefficients

Bayesian methods



- C+T: find subset of variants that best describe the trait of interest
- Bayesian methods: find optimal transformation of the vector of effect sizes to best represent the trait

$$PRS = \sum_{m=1}^{M} E\{\beta_m | Data\} G_m = \sum_{m=1}^{M} \widehat{\beta_m}$$

Bayesian methods

- Models the distribution of shrunk/re-weighted effect sizes
- Uses:
  → prior that reflects the genetic architecture (e.g. all SNPs have non-zero weight)
  → genome-wide LD matrix to weigh variants

→ Shrinkage method that produces **scaled weights genome-wide**

- Downsides: too many hyperparameters → harder to interpret

- Examples of software:
  - Ldpred: Vilhjalmsson, 2015
  - SBayesR: Ge et al, 2019
  - PRS-CS: Zeng et al, 2017

3

Polygenic scores

3.3

Application

Applying PRS

$$PGS_i = \sum_{j=1}^{N_{snps}} G_{ij} * \beta_j$$

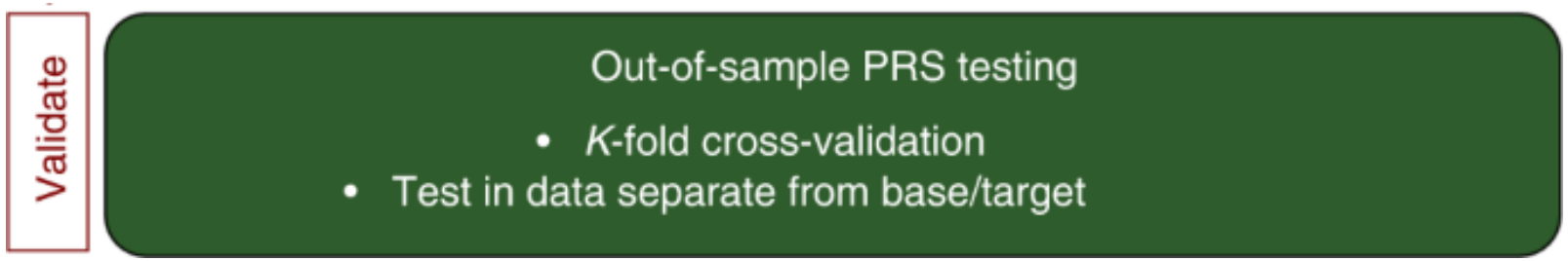
- Alleles need to be matched between base and target samples

  → An effect size (OR or β) is always associated to an allele

- Currently, PGS applied in target (independent sample) mainly for validation
- Future = application in the general population

  → Predict complex traits: prevention, monitoring, …
  → Patient stratification

Validation of PGS – independent sample



- Values to assess the prediction of PGS:

  → R2: amount of phenotypic variance explained by PGS (continuous traits) and pseudo-R2 for binary traits

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

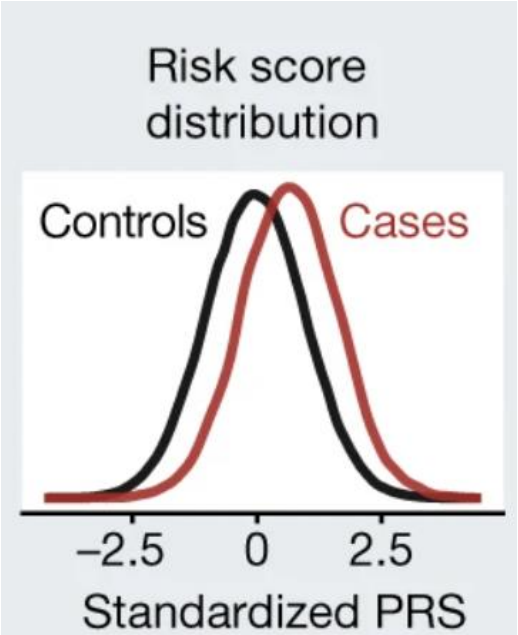
Variability in dependent variable not predicted by the model

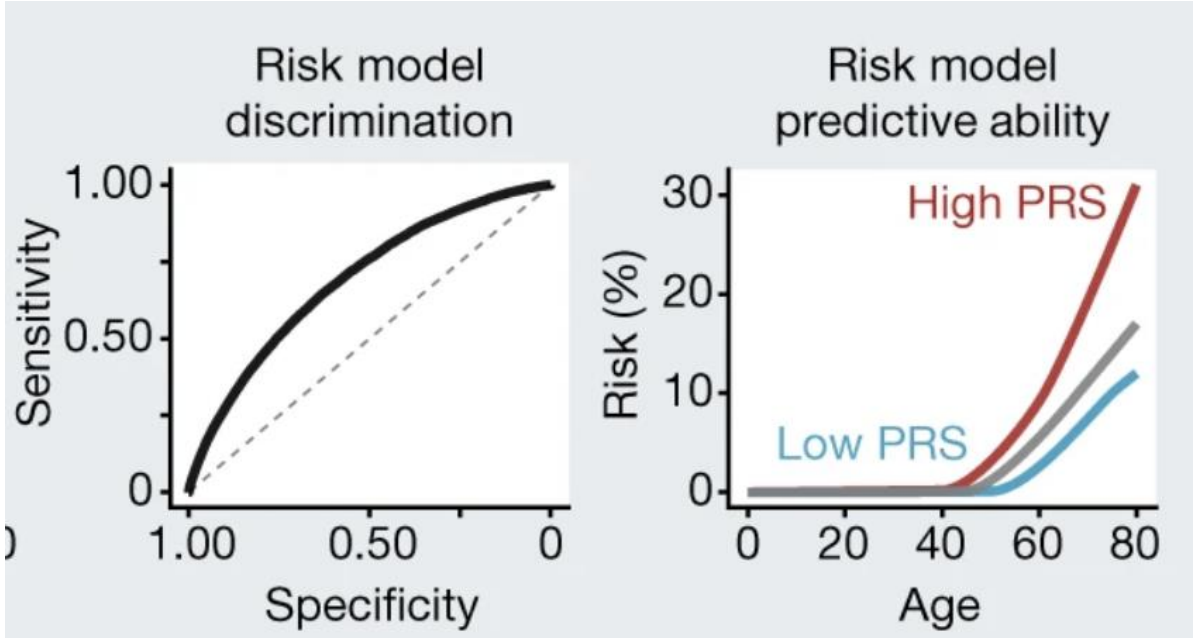
Variability in dependent variable

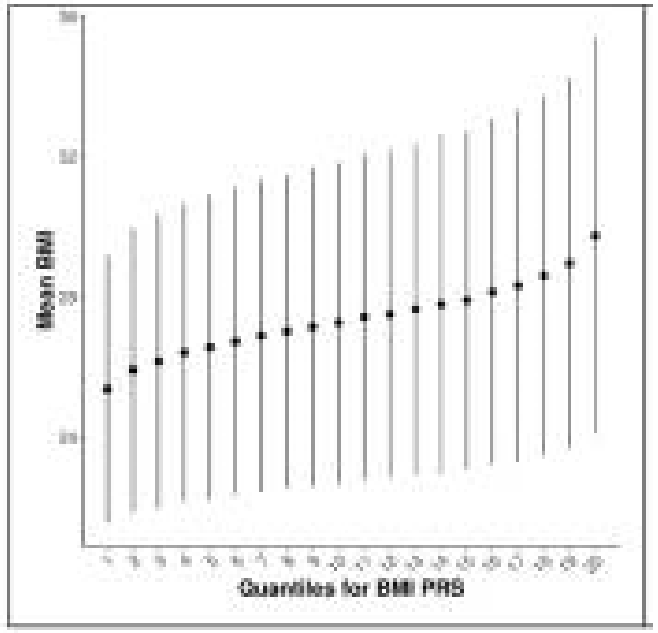
  → Odds ratio between strata
  → Area under the curve…

- Visualization techniques

Validation of PGS - visualization







- **ROC curves**: Measure of discrimination in disease prediction
- **Incidence plots**: changes in OR in each quantile compared to the reference
- **Quantile plots**: changes in OR in each quantile compared to the reference

Wand et al. Nat. 2021
Choi et al. Nat. prot. 2020

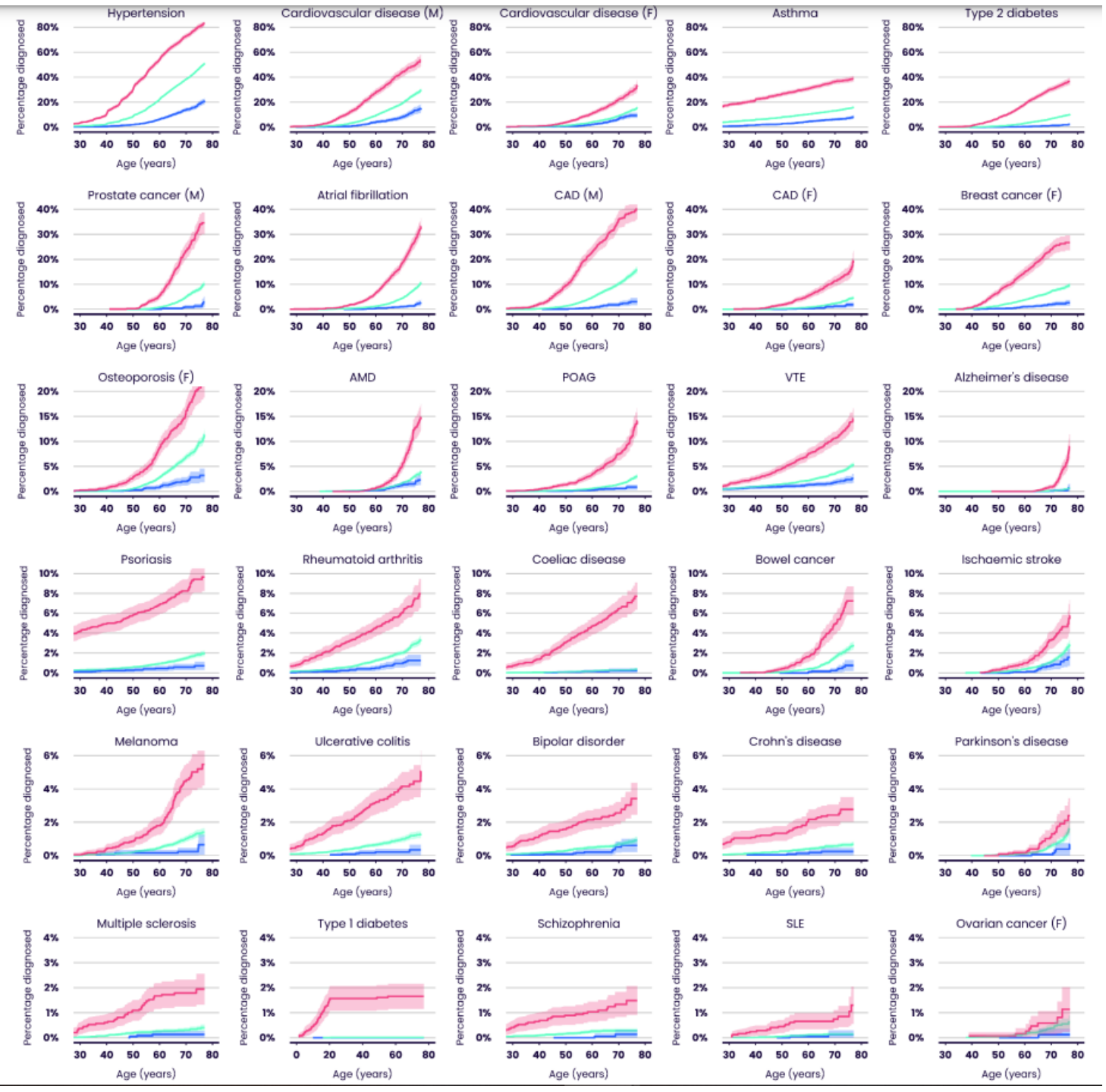


Figure 1. **Cumulative incidence plots illustrating the predictive performance of the UK Biobank PRS Release for 28 diseases in European ancestry individuals (Enhanced Set)**. Each plot shows

Thompson et al. 2022

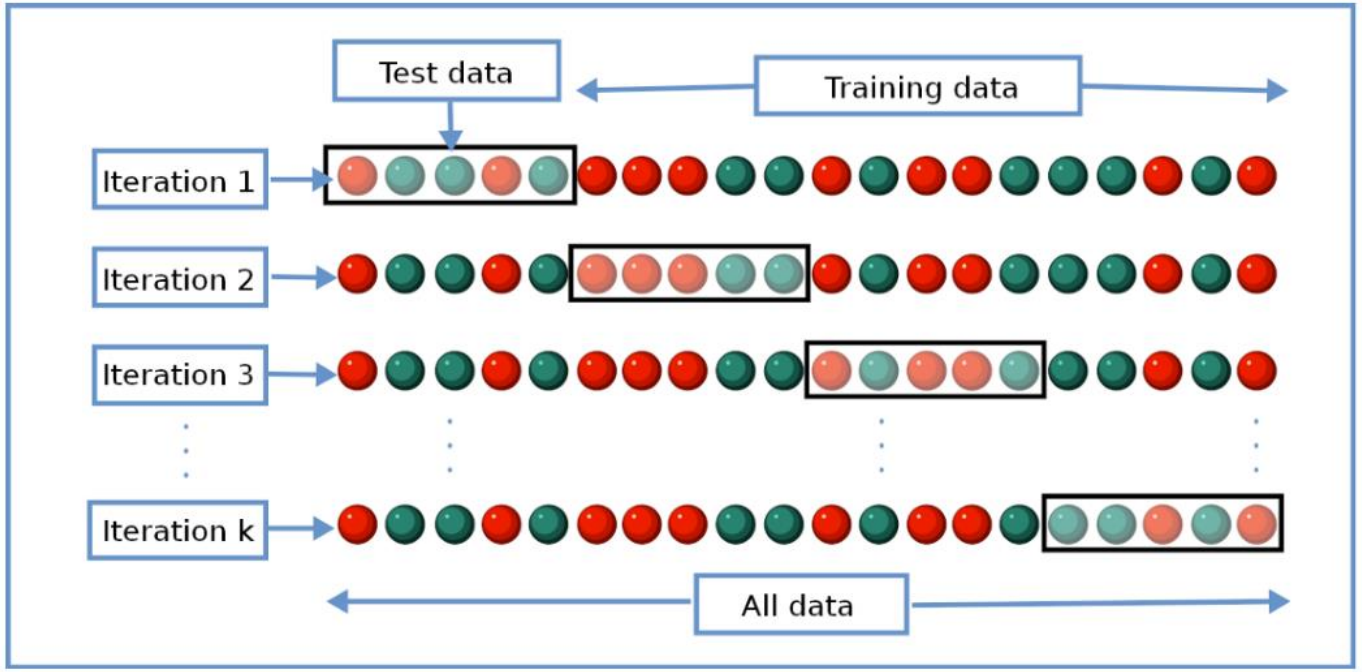
Validation of PGS

Validate

Out-of-sample PRS testing
- *K*-fold cross-validation
- Test in data separate from base/target

- K-fold cross-validation

  → When no independent dataset available
  → Divide the sample in training and validation data
  → Repeat multiple times



3

Polygenic scores

3.4

Limits

Limitations of PGS

- Ultimate goal = prediction of complex traits based on genetics

  → Patient stratification
  → Preventive medicine

- PRS relies on assumptions:

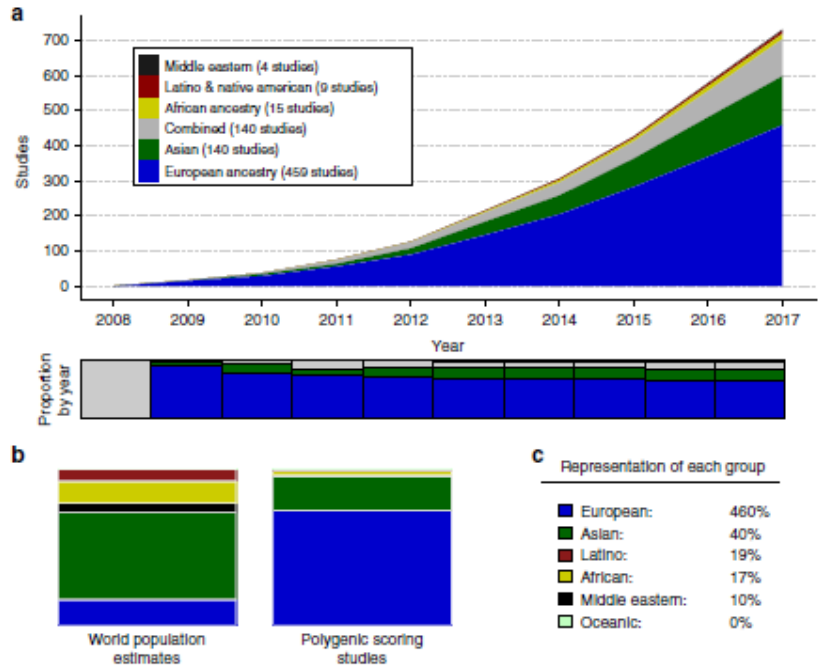
  → No environmental factors considered
  → Genetic associations = genetic causation
  → Homogeneity in discovery and testing samples

- Currently:

  - Limited clinical use: classical risk factors perform better
  - Only focus on common variants
  - Low transferability when deviation from original GWAS cohort: environment, ancestry …

Trans-ancestry PGS

- Currently, PGS mainly derived from European populations
- Poor transferability to non-European populations due to differences in:
  - → Allele frequencies
  - → LD
  - → Effect sizes
  - → Environmental factors
- Non-European PGS are limited due to small sample sizes
- Trans-ancestry PGS = active area of research
  - → New methods developed
  - → Decrease health disparities



Duncan et al. Nat. Comm. 2019

Mahajan et al. Nat. Genet. 2022

Overview

- Complex diseases underlined by a polygenic architecture

- LD induces correlation between SNPs and needs to be accounted for

- PGS aim at predicting complex traits based on genetic variants identified in GWAS

  → Choice of base and target samples
  → Selection of SNPs in PGS
  → Trans-ancestry PGS

Any questions ?

Thank you.