HELMHOLTZ MUNICH

Polygenic scores

Ana Arruda and Ozvan Bocher
14th of June 2023

Agenda

1. Human genetics recap
2. GWAS recap
3. Complex traits
4. Polygenic scores

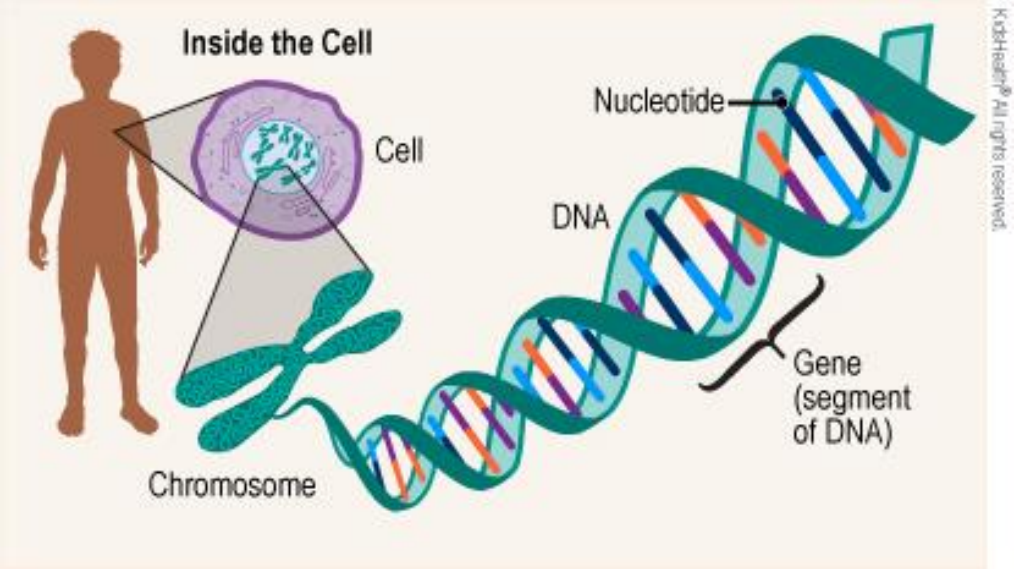
1

Human genetics recap

What is a **gene**?

A gene is a **sequence of nucleotides** in DNA or RNA that **encodes the synthesis of a gene product**, either RNA or protein.

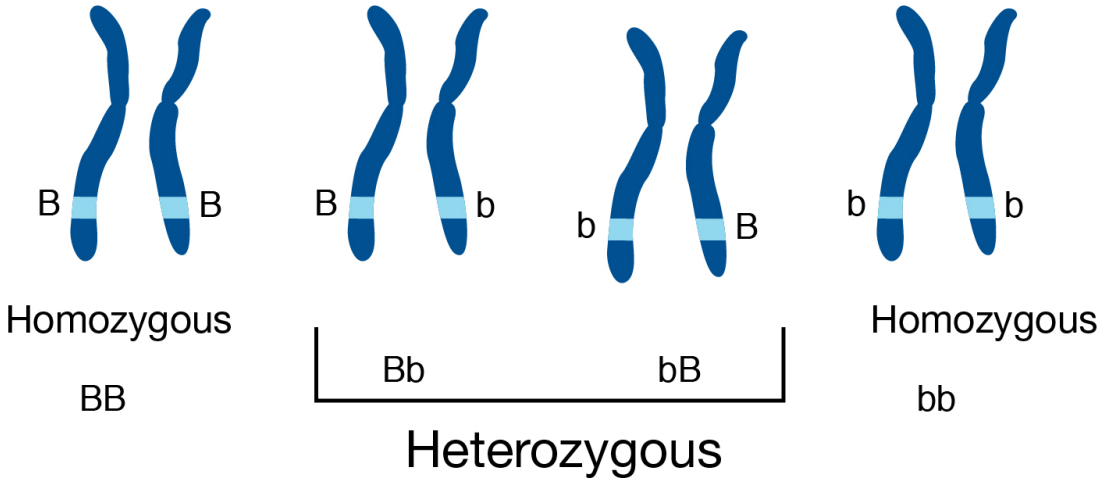
A genome region that includes all of the sequence elements necessary to **encode a functional transcript** and specifies a trait.



What is an **allele**?

Allele: different forms of the same gene that determines an organism's phenotype. It is represented by letters.

Humans are **diploid organisms**, which means that they have **two alleles at each genetic position**, or locus, with one allele inherited from each parent.



Allele b count			
BB	bB	Bb	bb
0	1	1	2

What is a **genotype**?
What is a **phenotype**?

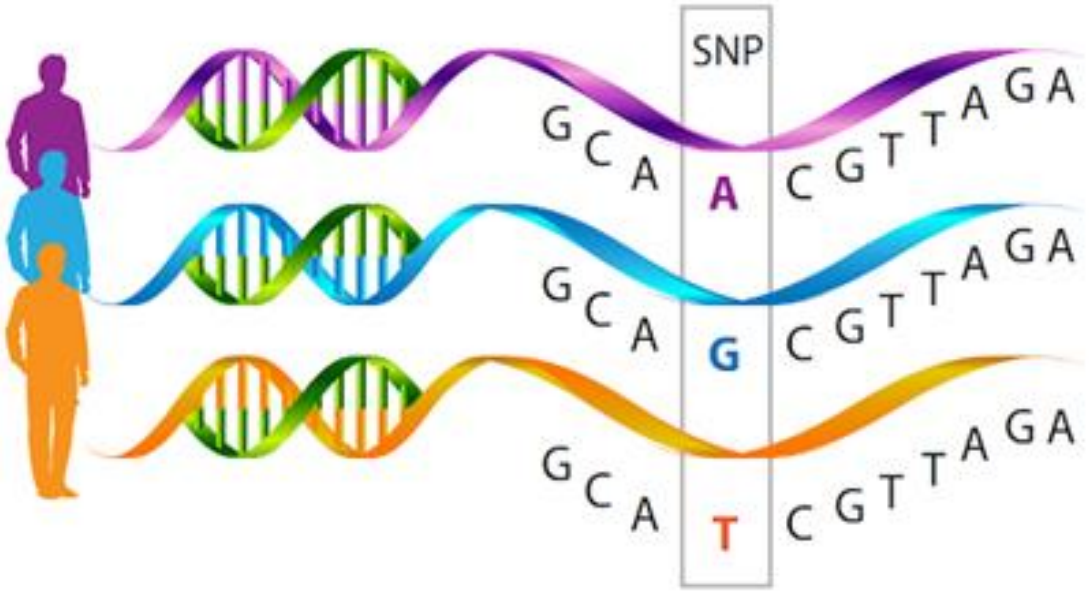
Genotype vs Phenotype

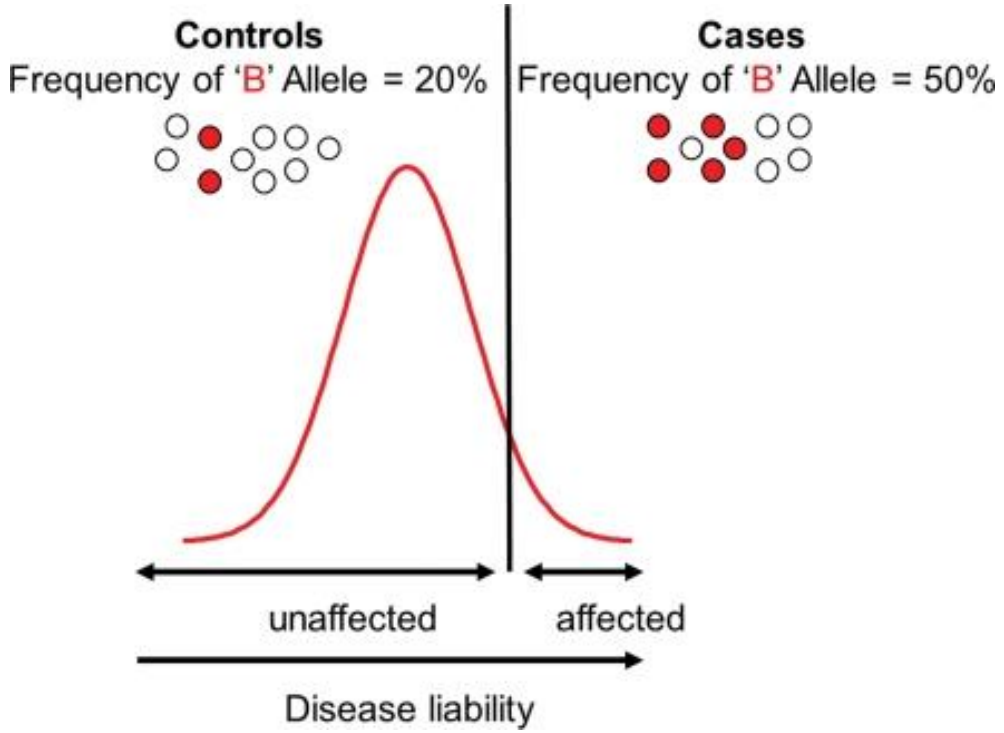
GENOTYPE	PHENOTYPE	
The genotype is an organism's genetic information.	The phenotype is the set of observable physical traits.	
BB homozygous dominant	purple	
Bb heterozygous	purple	
bb homozygous recessive	white	

What is a **SNP**?

What is a **risk/effect allele**?

Single-nucleotide polymorphism





2

Genetic association studies

Modelling

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

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

$$\begin{bmatrix} pheno_0 \\ \vdots \\ pheno_n \end{bmatrix}$$

$= \{0,1\}$ (case-control)

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

$$\begin{bmatrix} A/T \\ \vdots \\ T/T \end{bmatrix}$$

$= \{0,1,2\}$ (genotype, directly typed)

$\in [0,2]$ (dosage, imputed)

$$\begin{bmatrix} 1 \\ \vdots \\ 2 \end{bmatrix}$$

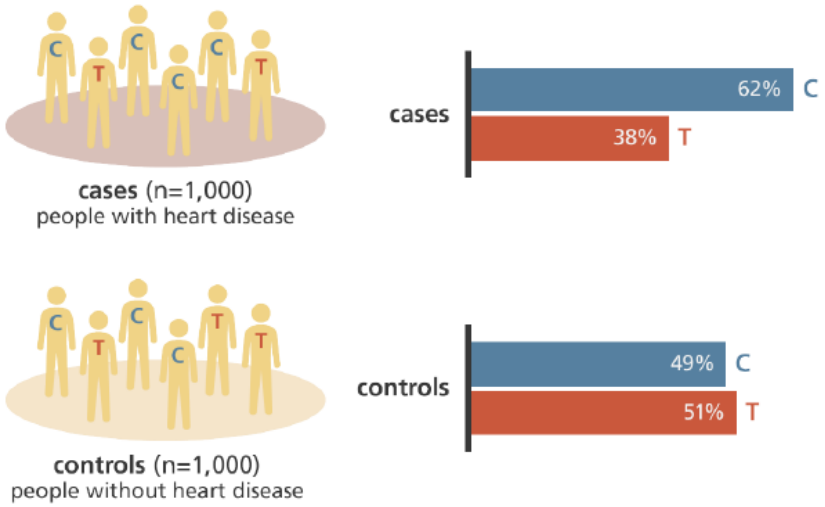
$$\begin{bmatrix} 0.965 \\ \vdots \\ 1.816 \end{bmatrix}$$

- 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)

Case/control studies

Odds ratio (OR): *how much more likely are you to be a case if you carry the risk allele ?*

- Per genotype *g* and disease *Y*, we compute the odds $O = \frac{p}{1-p} = \frac{p_{Y=1|g}}{1-p_{Y=1|g}} = \frac{p_{Y=1|g}}{p_{Y=0|g}}$



	Cases	Controls
T	380	510
C	620	490

$$O_T = \frac{380/n_T}{510/n_T} \qquad O_C = \frac{620/n_C}{490/n_C}$$

$$OR = \frac{n_{affected\ carriers} \times n_{healthy\ non-carriers}}{n_{healthy\ carriers} \times n_{affected\ non-carriers}}$$

$$OR_{C/T} = \frac{620 \times 510}{490 \times 380} = 1.7$$

Case/control studies

OR = ratio of the odds of the two alleles

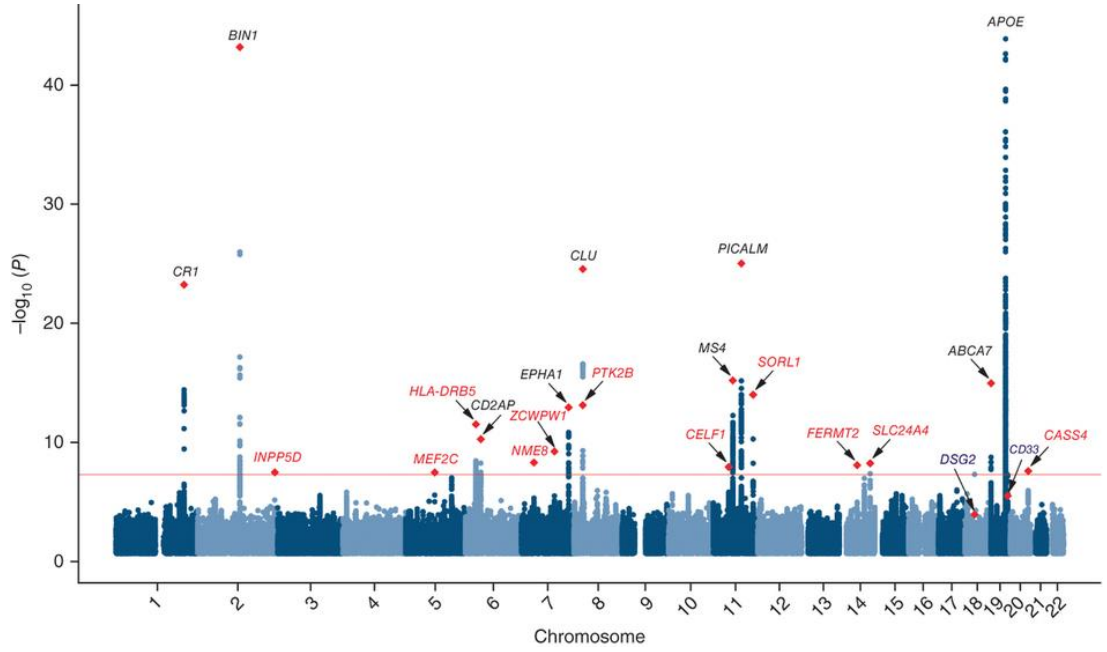
- OR>1: the allele is ‘deleterious’
- OR<1: the allele is ‘protective’

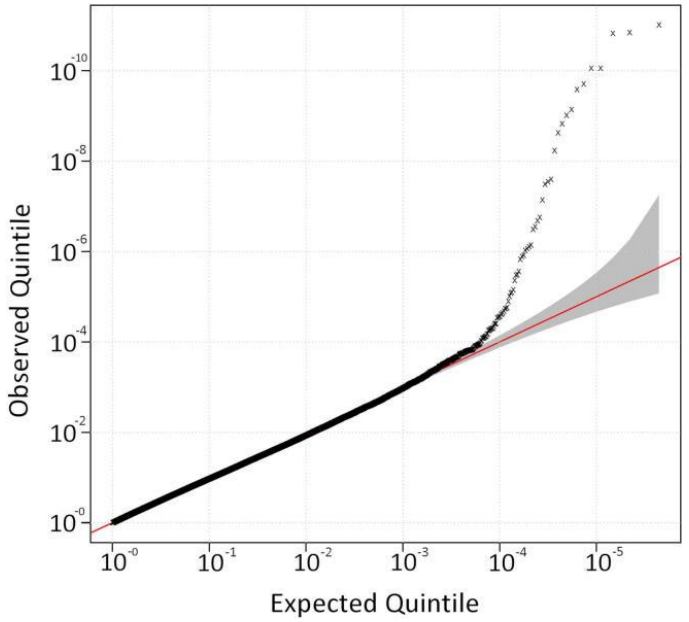
Statistical test: is the OR significantly different from 1?

- Earlier: Fisher’s exact test or Chi-squared test
- Nowadays + for imputed data: linear regression or GLM

GWAS results

1. Quality control (QC) of the data
2. Run model
3. Correct p-value for multiple testing (significance threshold for genomics = $5x10^{-8}$)
4. Visualize results (Manhattan plot)
5. Run sensitivity analysis



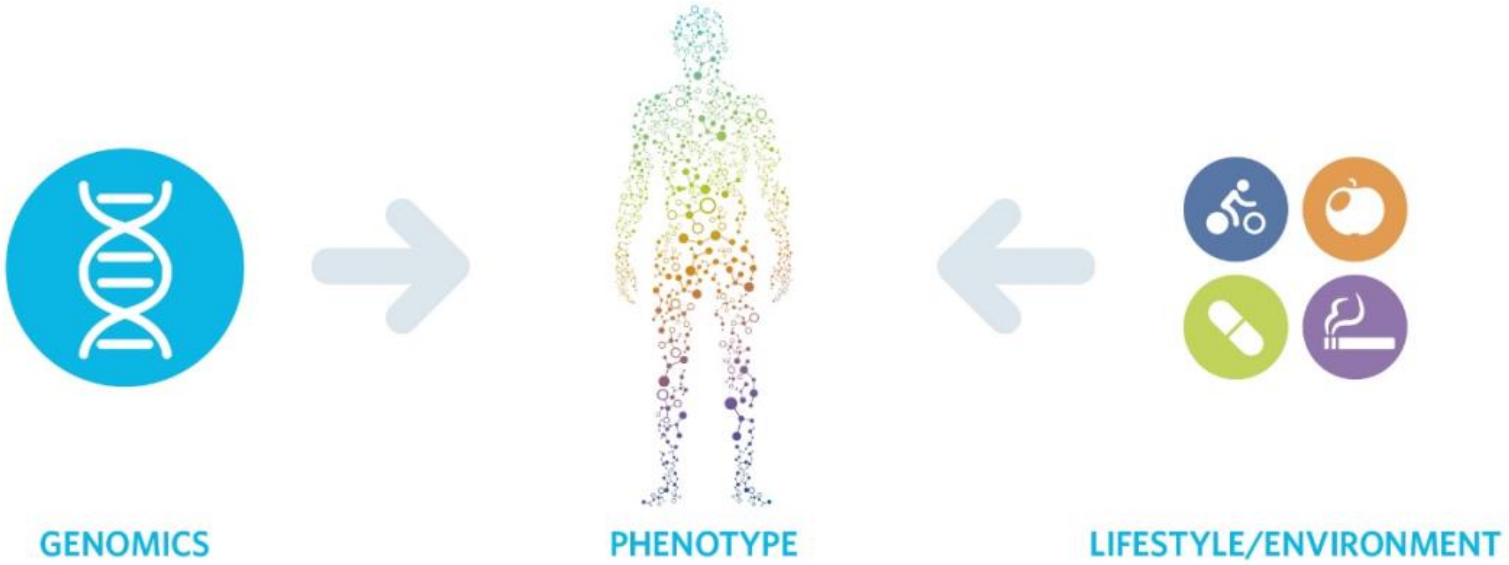


3

Complex traits

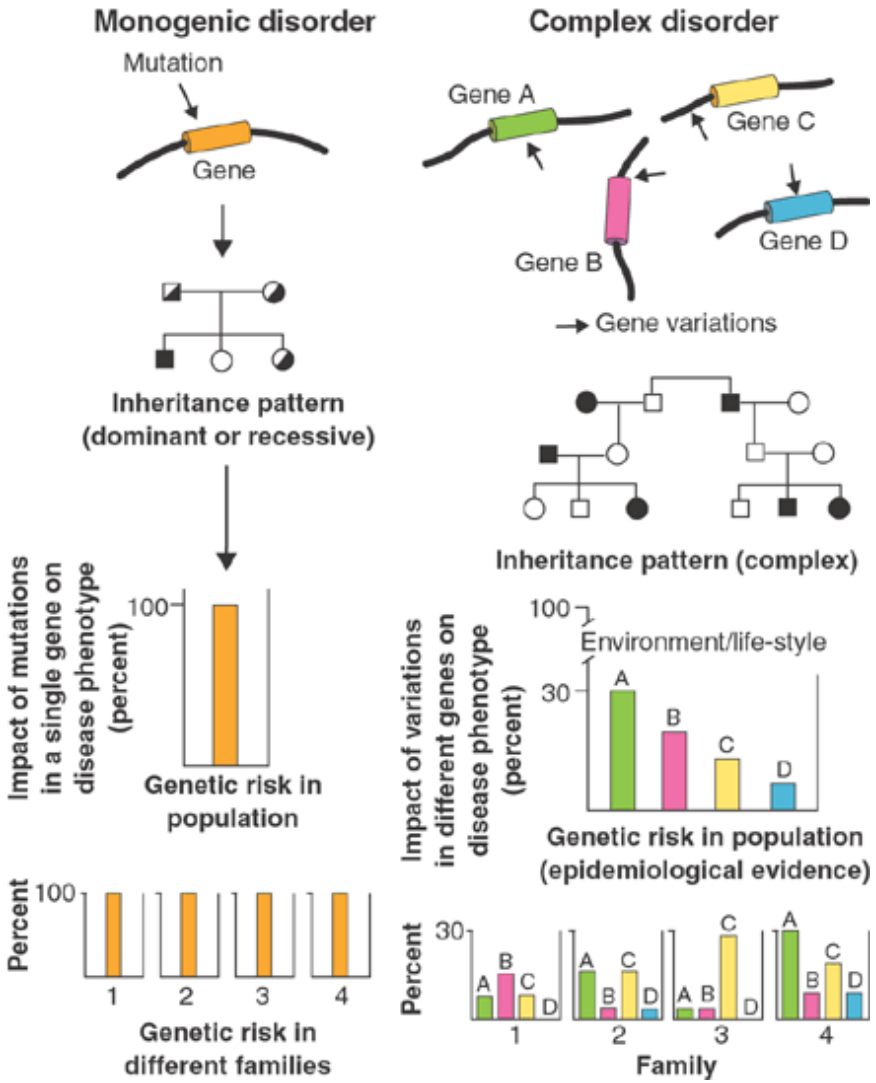
Complex traits

- 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

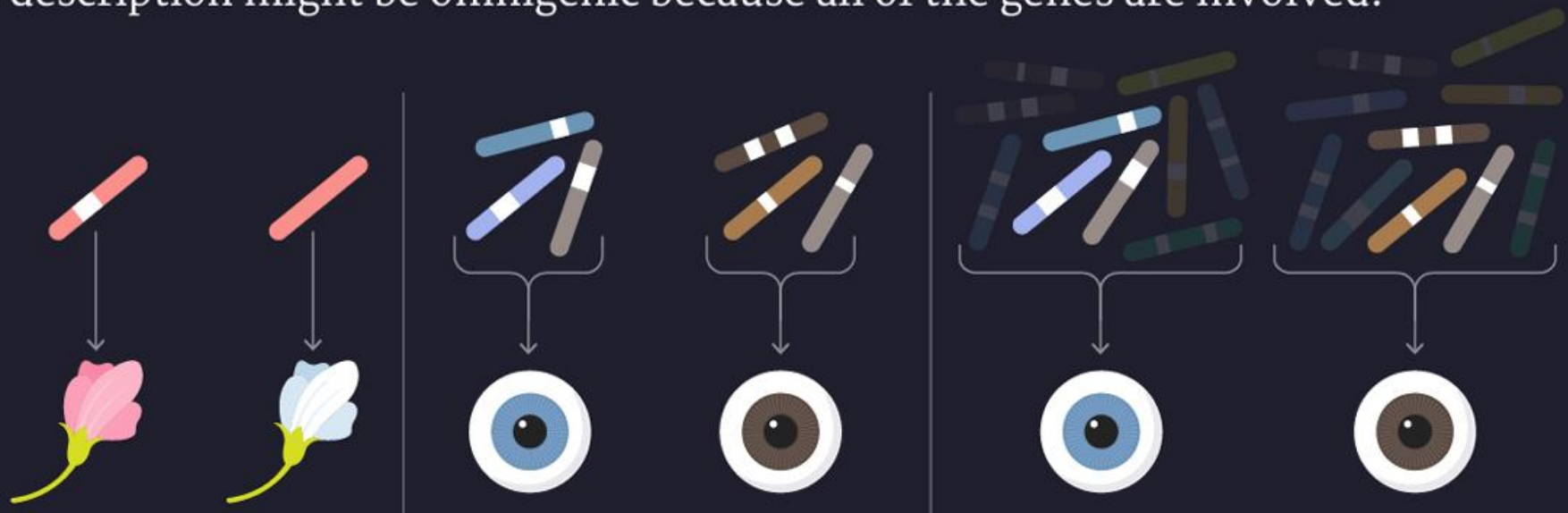
Monogenic disorder vs complex traits



Peltonen & McKusick, Science, 2001

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.

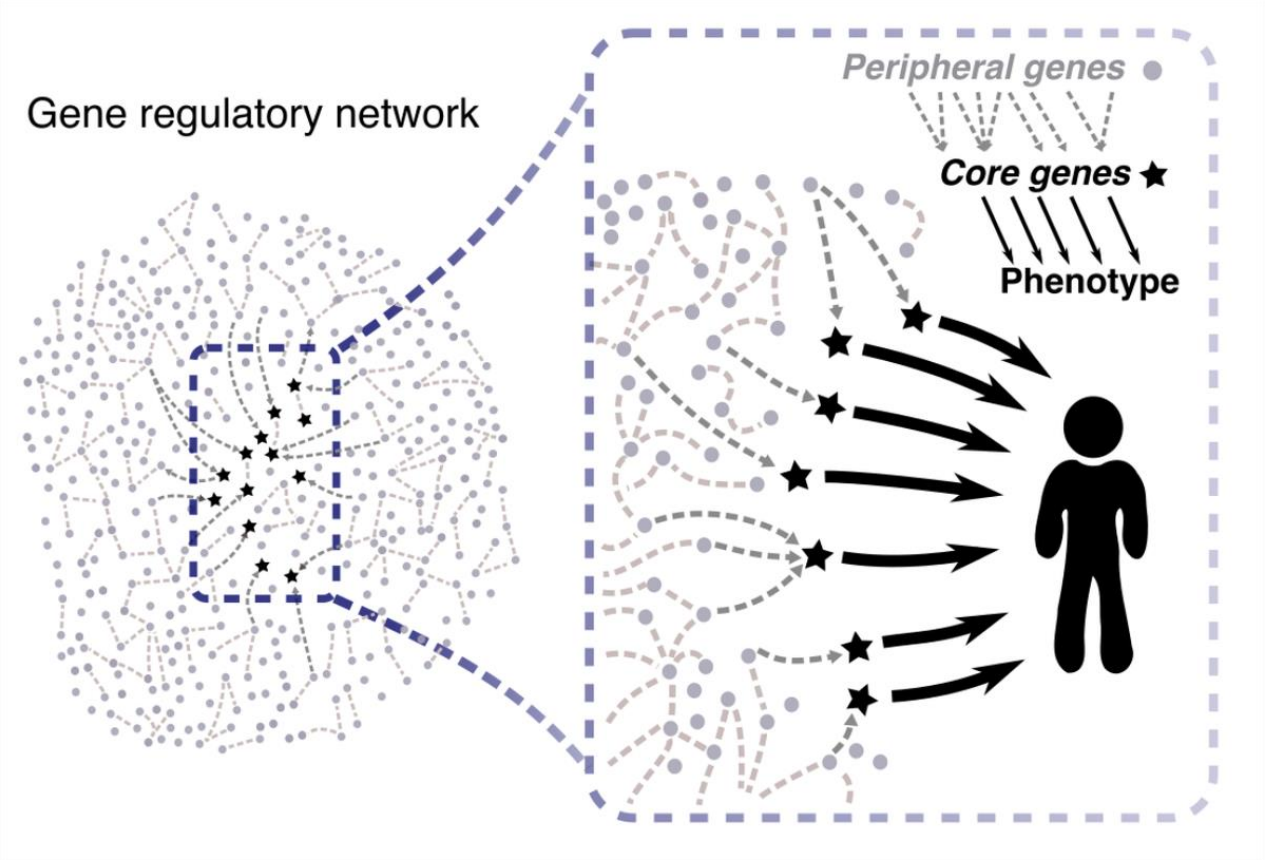
Polygenic

A handful of genes jointly give rise to a trait.

Omnigenic

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

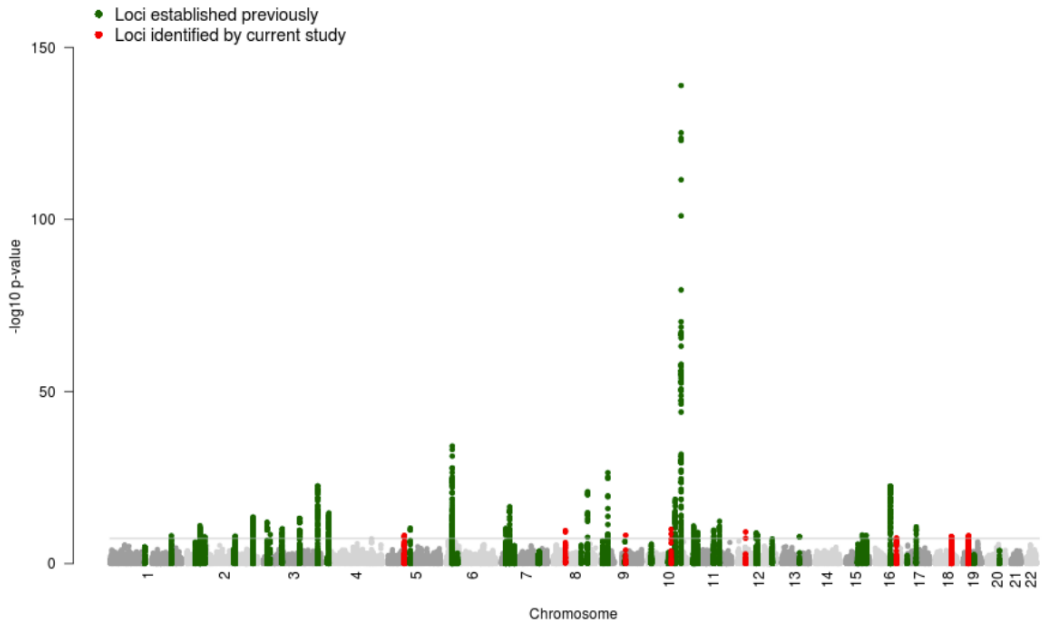
Omnigenic vs Polygenic model



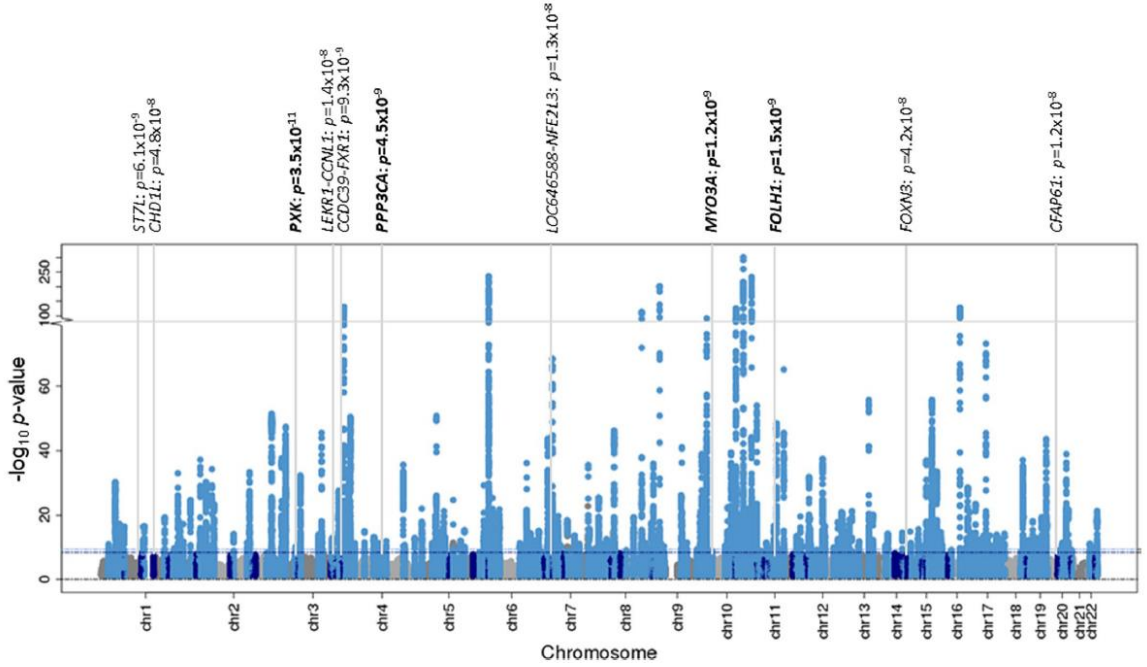
GWAS

Current tool of choice to study complex traits

Type 2 diabetes GWAS *Morris et al. Nat. Genet. 2012*
Number of cases = 34,840
Number of controls = 114,981



Type 2 diabetes GWAS *Mahajan et al. Nat. Genet. 2022*
Number of cases = 180,834
Number of controls = 1,159,055



3

Polygenic scores

3.1

Introduction

Polygenic scores

- Natural follow up for complex traits:
  - Influenced by many genetic variants
  - GWAS → effect for each variant
  - Additive model → risk increases with each copy of the effect allele

→ combine SNPs effects into a score

- Used to: predict quantitative traits (= polygenic score) or disease risk (= polygenic risk score)

Polygenic scores (PGS)

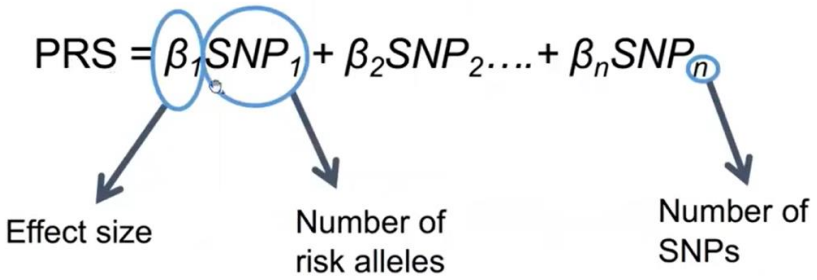
Polygenic score for individual *i*

Total number of SNPs included in PGS

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

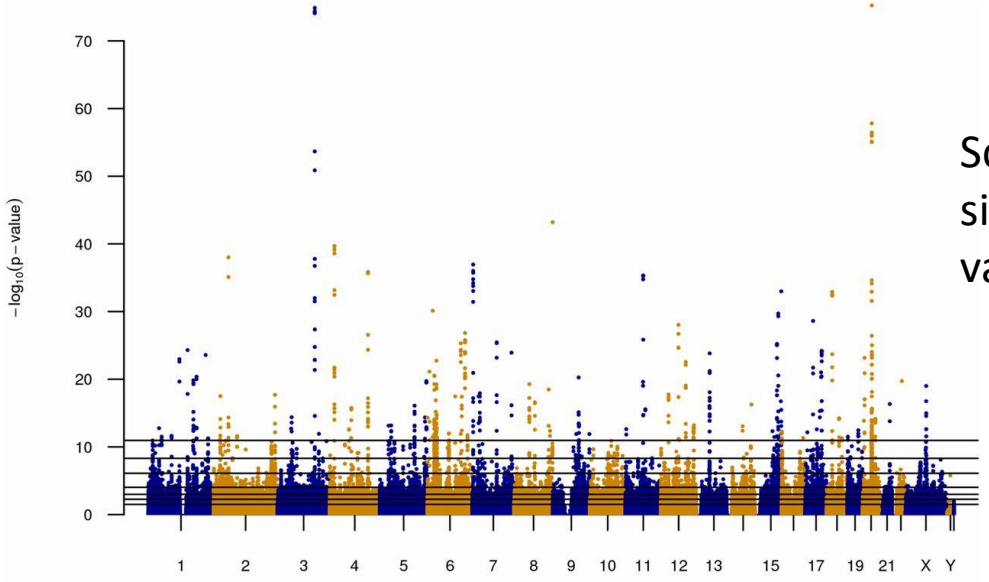
Effect of variant *j* on trait
- Estimated in GWAS

Genotype at SNP *j* for individual *i*
- Coded as 0, 1, 2 depending on the number of risk allele
- Additive model



→ Sum of the number of risk alleles **weighted** by its effects

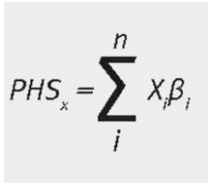
Polygenic scores

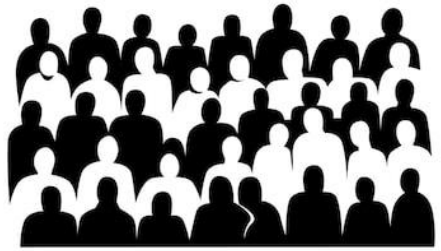


Large GWAS

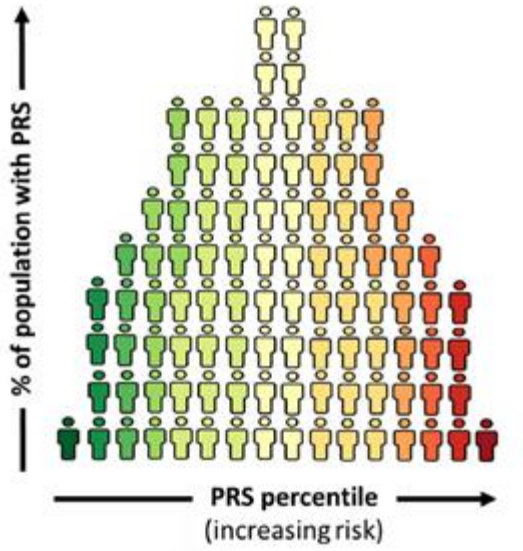
Scores for significant variants

Genetic data





External cohort



Summary-level data = BASE

Individual-level data = TARGET

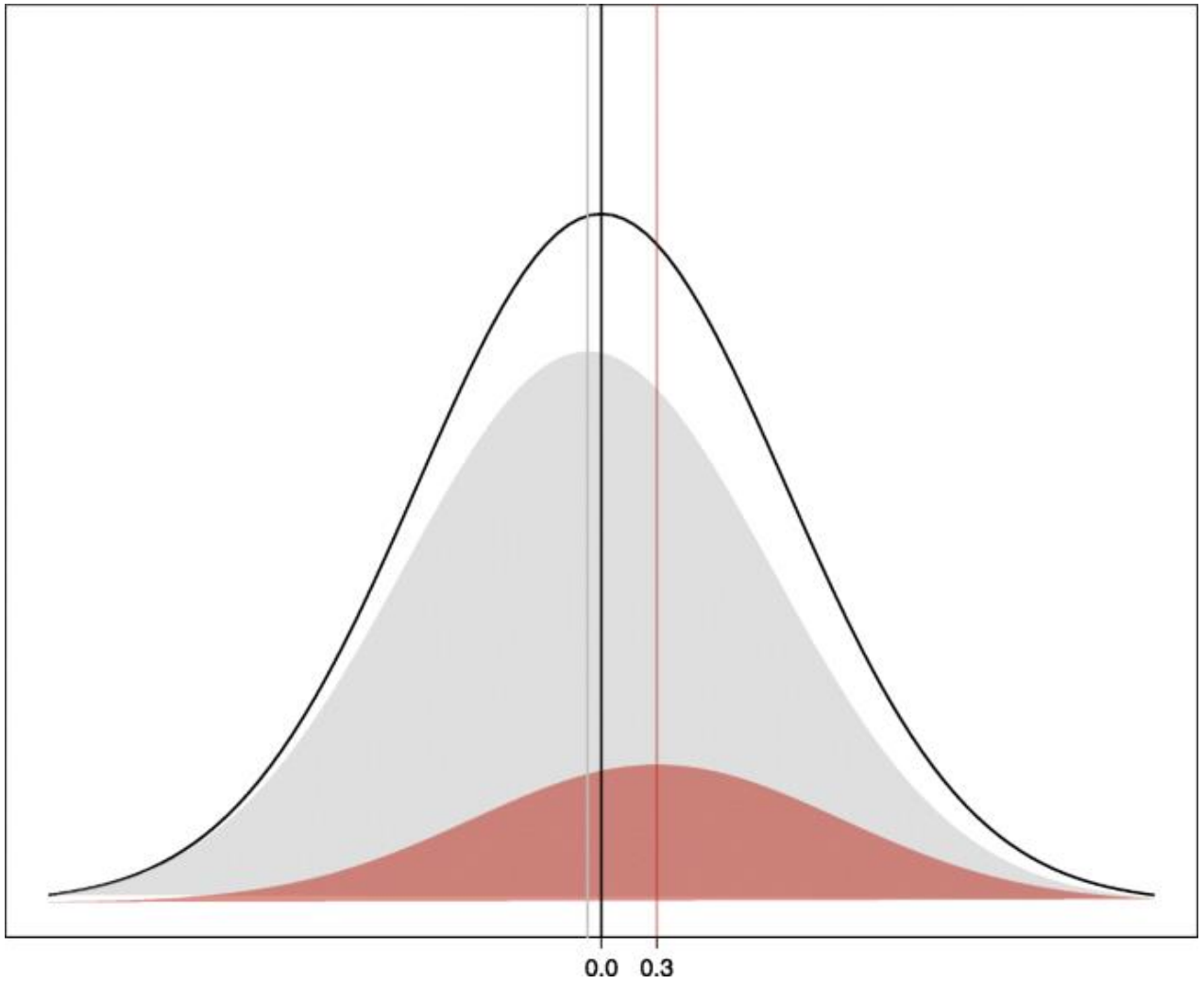
HELMHOLTZ MUNICH

Polygenic risk score

Case/control study

- Grey = controls
- Red = cases
- Overall mean = 0 (standardized)

- Amount of shift = population variance of PGS under log-linear model



3

Polygenic scores

3.2

Construction

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

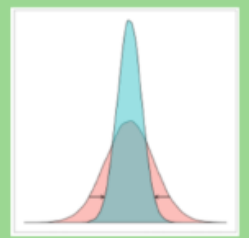
- Both data sets QCed as standard in GWAS
- Some QC requires special care in PRS (e.g., sample overlap, relatedness and population structure)
- Retain set of SNPs that overlap between base and target data

PRS calculation

LD adjustment

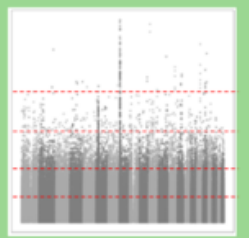
- e.g., clumping

Beta shrinkage



- e.g., LASSO/ridge

P value thresholding

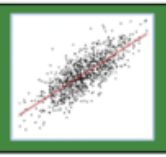


- PRS at multiple *P*

Test

ID	BMI	PRS
101	24.1	0.43
102	28.3	1.61
103	31.2	0.83
104	19.4	3.54

Generate PRS
+
Perform association testing



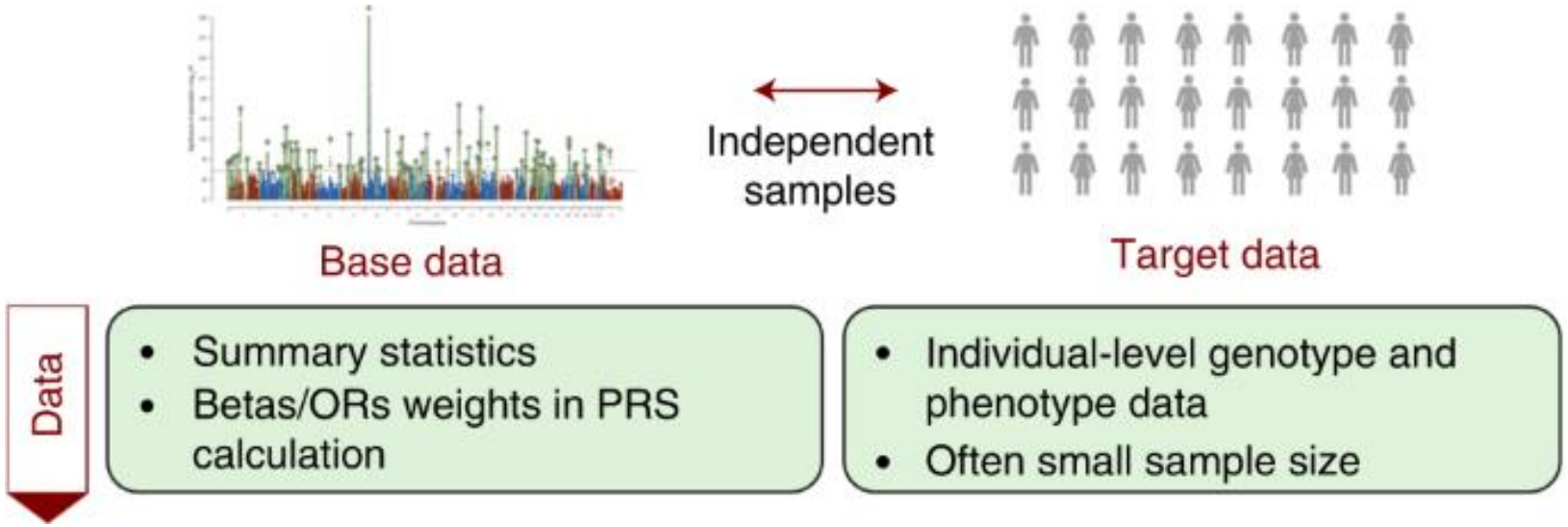
Validate

Out-of-sample PRS testing

- *K*-fold cross-validation
- Test in data separate from base/target

Choi et al. 2020

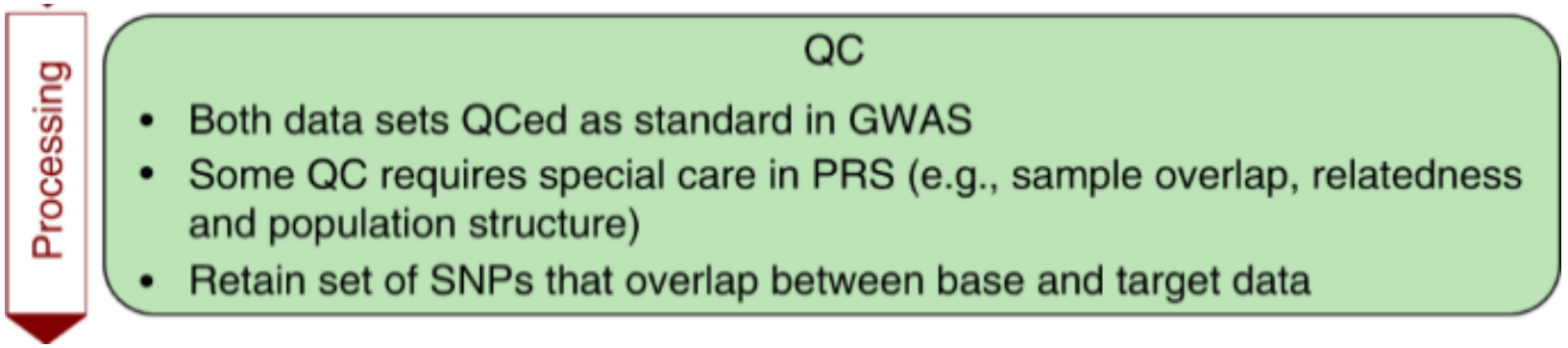
Input data



Sample used to estimate parameters for the PGS
→ Largest GWAS summary statistics
→ We need:
- Effect sizes of the variants: betas/OR
- standard errors
- p-values

Sample where we will apply the PGS
→ Individual level data (genotype data)
→ Often a small sample size

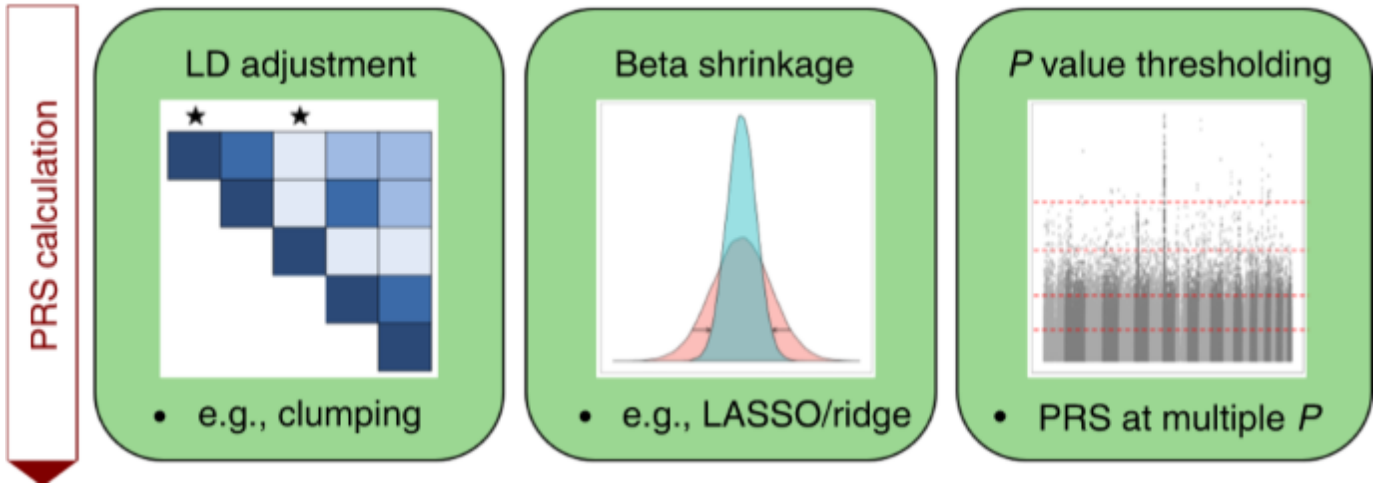
Data processing



- No sample overlap between base and target data
  → Could lead to inflation of effects: ‘overfitting’

- Need homogeneity between base and target samples
  → Hypothesis = samples underlie the same genetic architecture and environmental conditions
  → E.g. same population structure/ancestry

PGS calculation



How to select variants influencing complex traits?

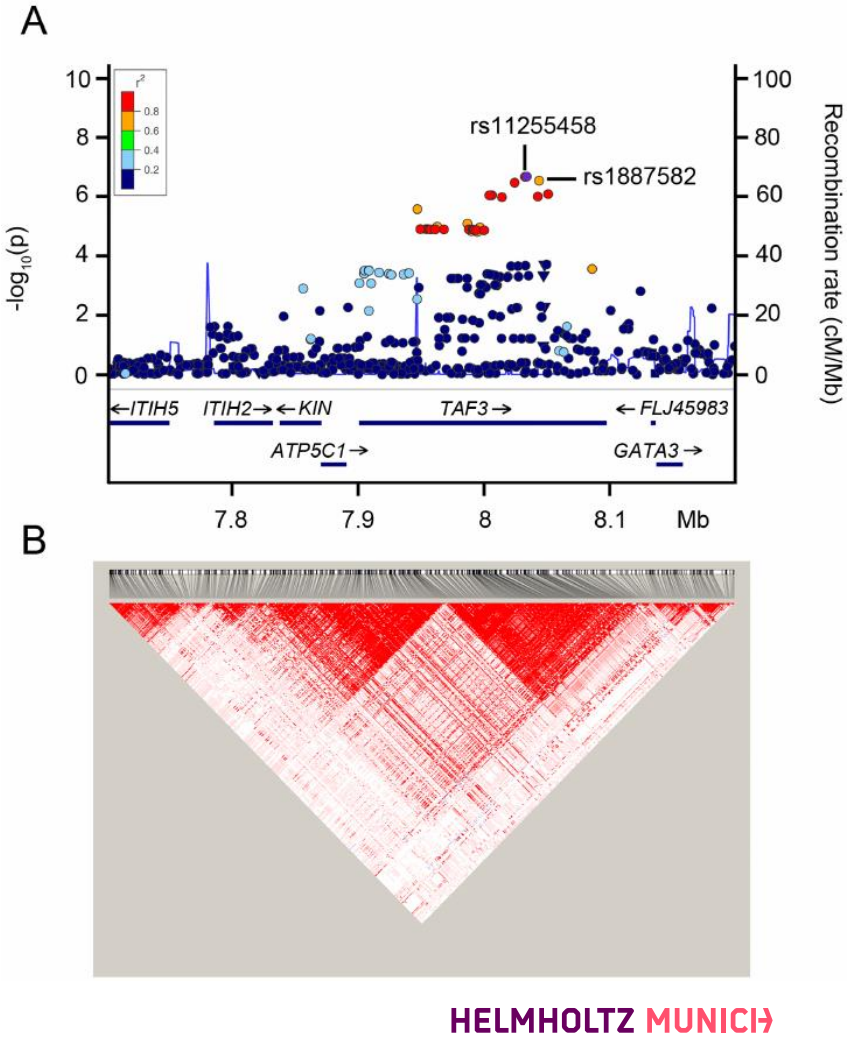
Selection of variants for PGS calculation

- Historically: independent top variants only
  → Challenging in omnigenic and polygenic models
  → Reduced predictive power

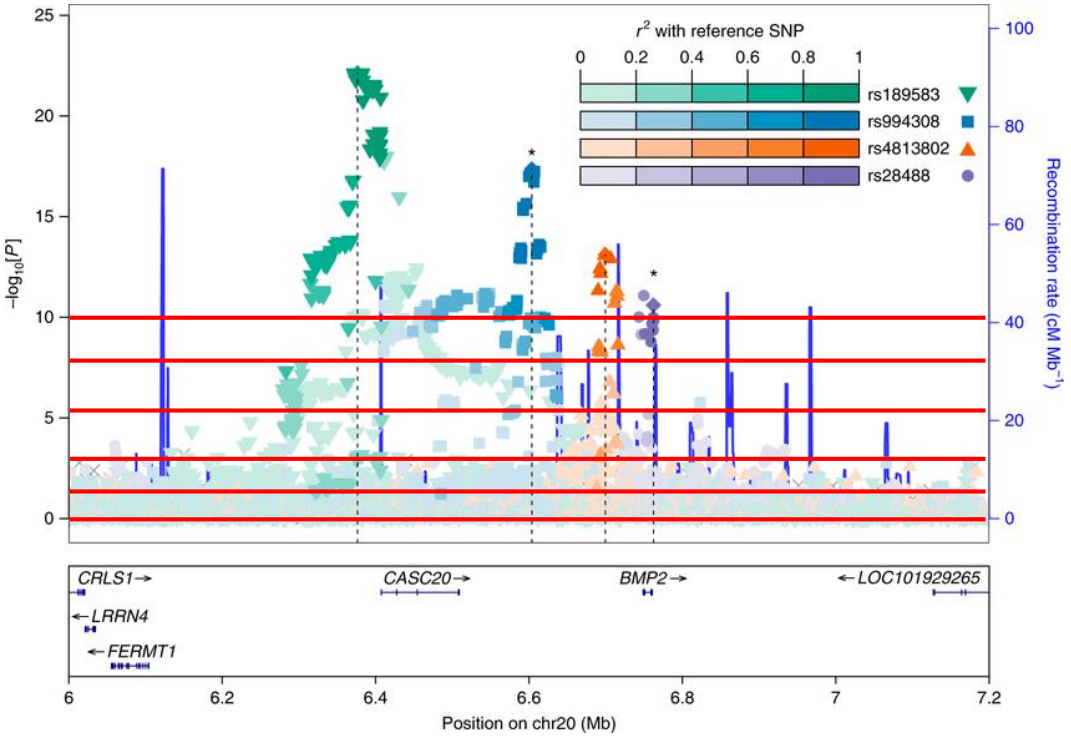
- Solution: use all variants (omnigenic model)
  → Problem: linkage disequilibrium (LD)

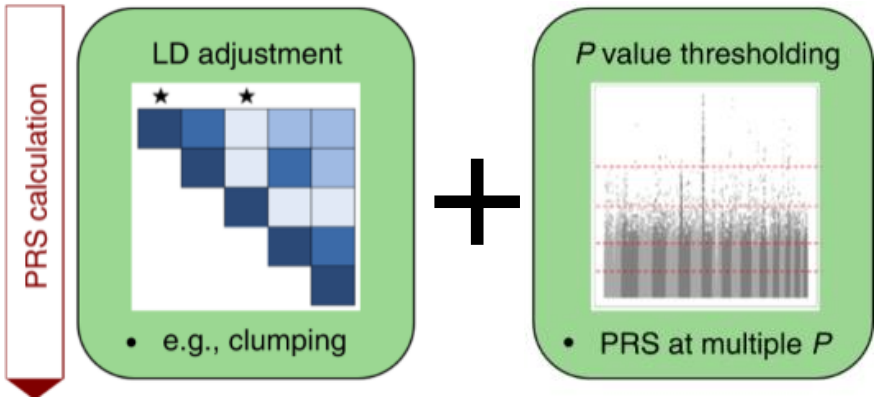
What is LD?

- Now: select **independent variants** (clumping, pruning)
  → One representative for each LD block
  → No overweighting of LD blocks



Clumping + Thresholding (C+T)





P-value aware LD pruning

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

Any other significant variant left in the block ?

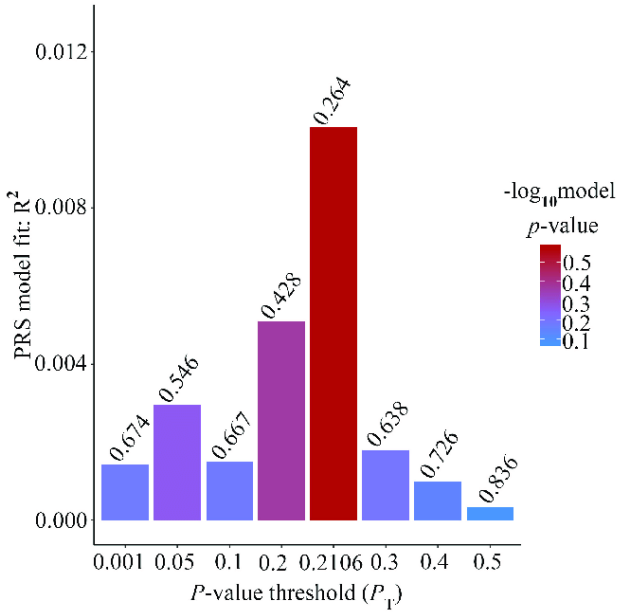
YES

NO

Set of pruned and significantly associated variants

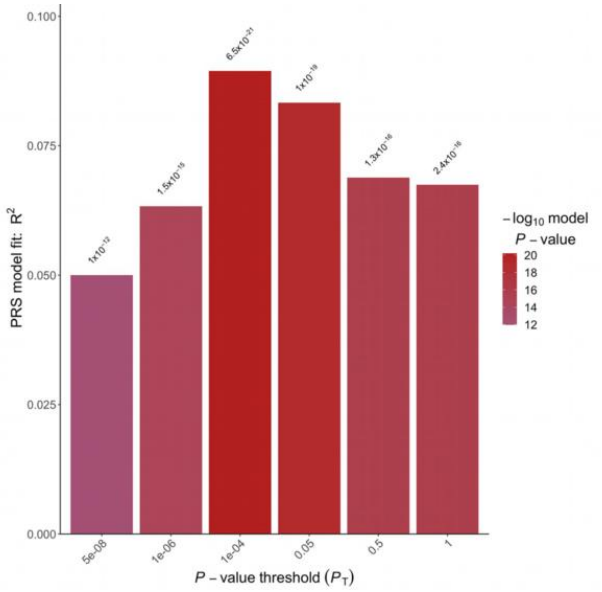
Clumping + Thresholding (C+T)

- Which significance threshold to use to include variants?
  → Optimal threshold depends on the trait
  → More polygenicity → less stringent threshold



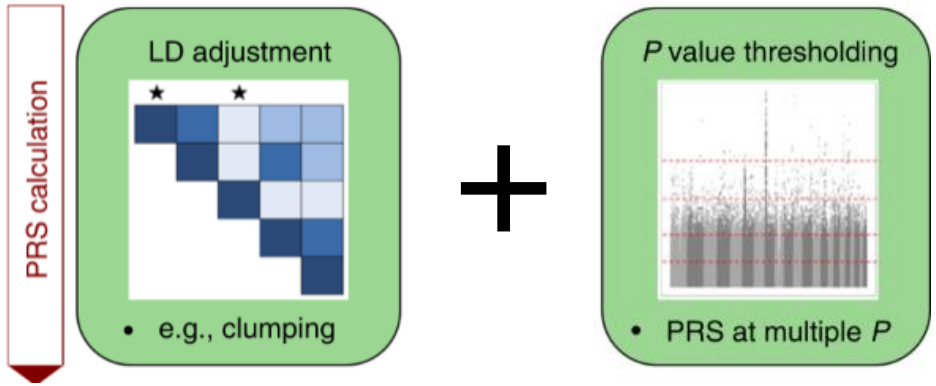
Wang et al. Frontiers in Genetics, July 2019

- Unknown beforehand
  → Try multiple values with validation
  → Integrated into PGS calculation software, e.g. *PRSice*



Maj et al. Frontiers in Cardiovascular Medicine, Feb 2022

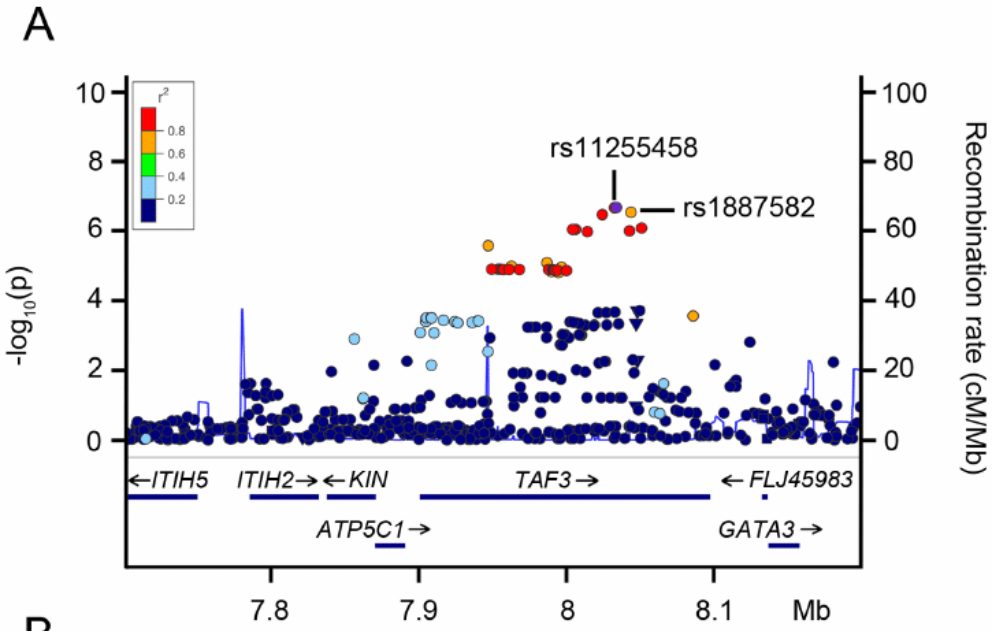
Limitations of C+T



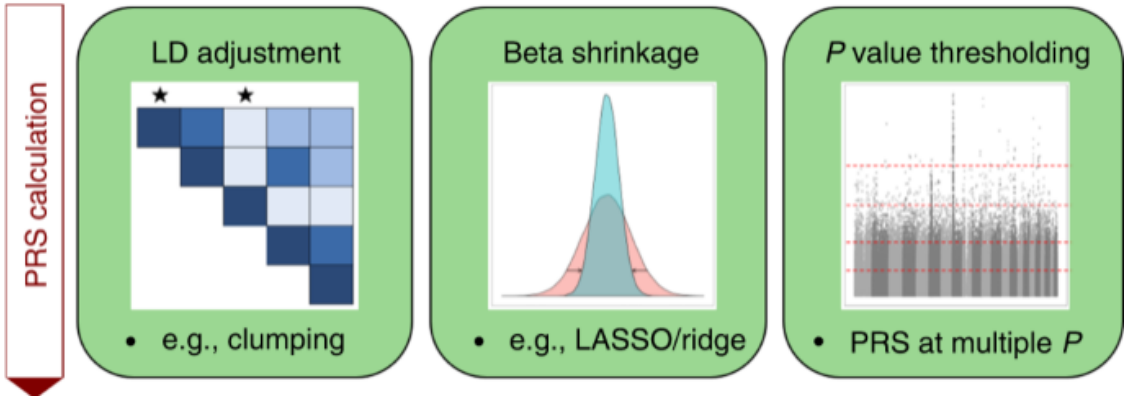
- Potential removal of secondary signals
- Based on the p-values but not the effect sizes
  → The p-value is related to the power of the study
  → Can miss low-effect variants in small sample sizes

- Ideal model = 'whole-genome' model
  → Account for LD

- Sample size is still a limiting factors for improved methods



Bayesian sparse regression methods (beta shrinkage)



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

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

Bayesian sparse regression methods (beta shrinkage)

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

- 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 down variants

→ Shrinkage method that produces scaled effect sizes genome-wide

- Downsides: too many hyperparameters → harder to interpret

Short list of software to calculate PGS

Clumping + thresholding

- PRSice

Bayesian sparse regression method

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

3

Polygenic scores

3.3

Application

Applying PGS

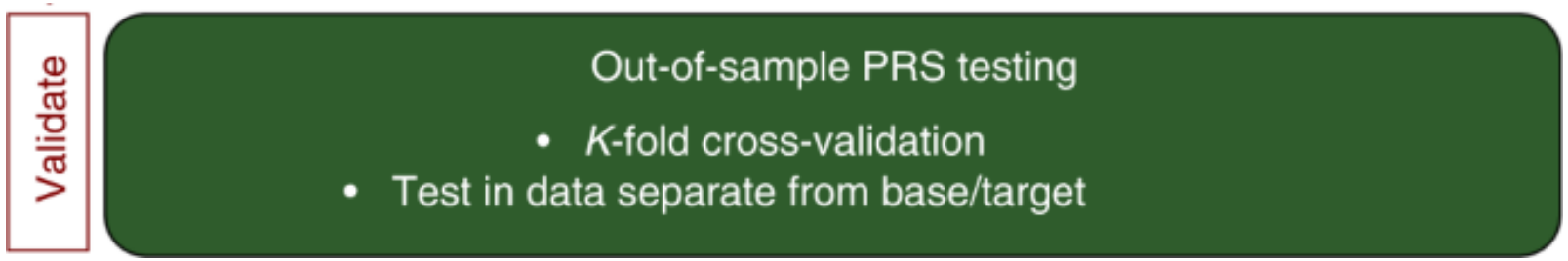
$$PGS_i = \sum_{j=1}^{N_{snps}} G_{ij} * \beta_j$$

- Alleles must be matched between base and target data → beta inversion

$$\beta_{rs1234,A} = 1.56$$
$$alleles_{rs1234} = \{A, T\}$$
$$\Longrightarrow \beta_{rs1234,T} = -1.56$$

- Currently: PGS applied mainly for validation (test predictive power)
- 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)

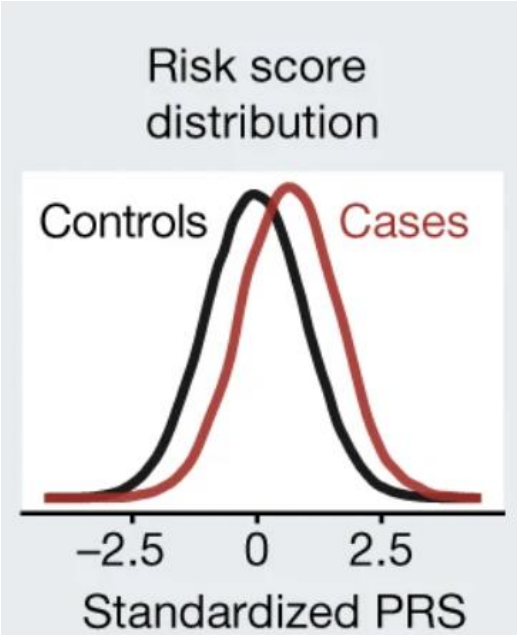
$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

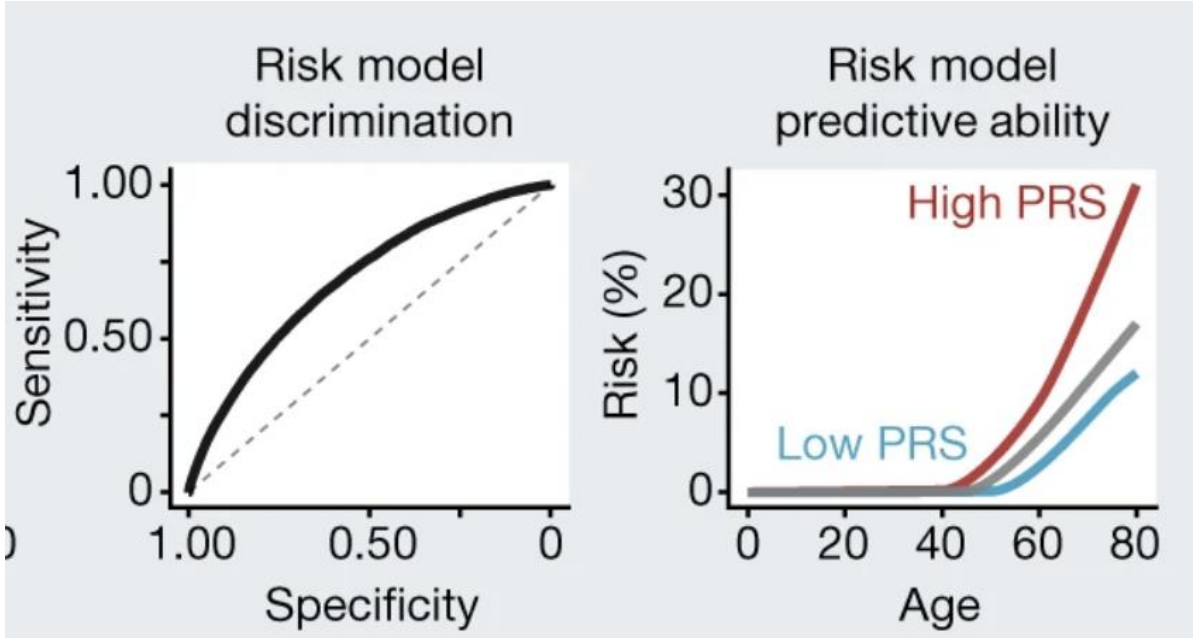
Variability in dependent variable not predicted by the model

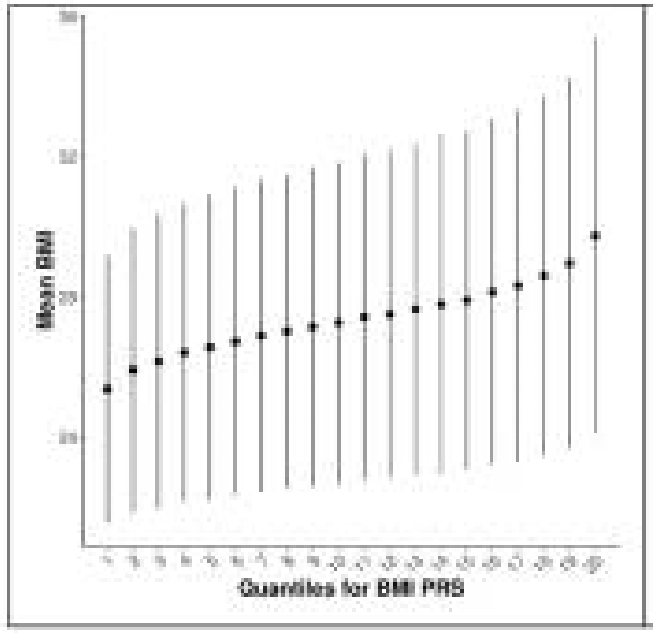
Variability in dependent variable

  → Pseudo-R2: R2 for binary traits
  → Odds ratio between different groups
  → Area under the curve…

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

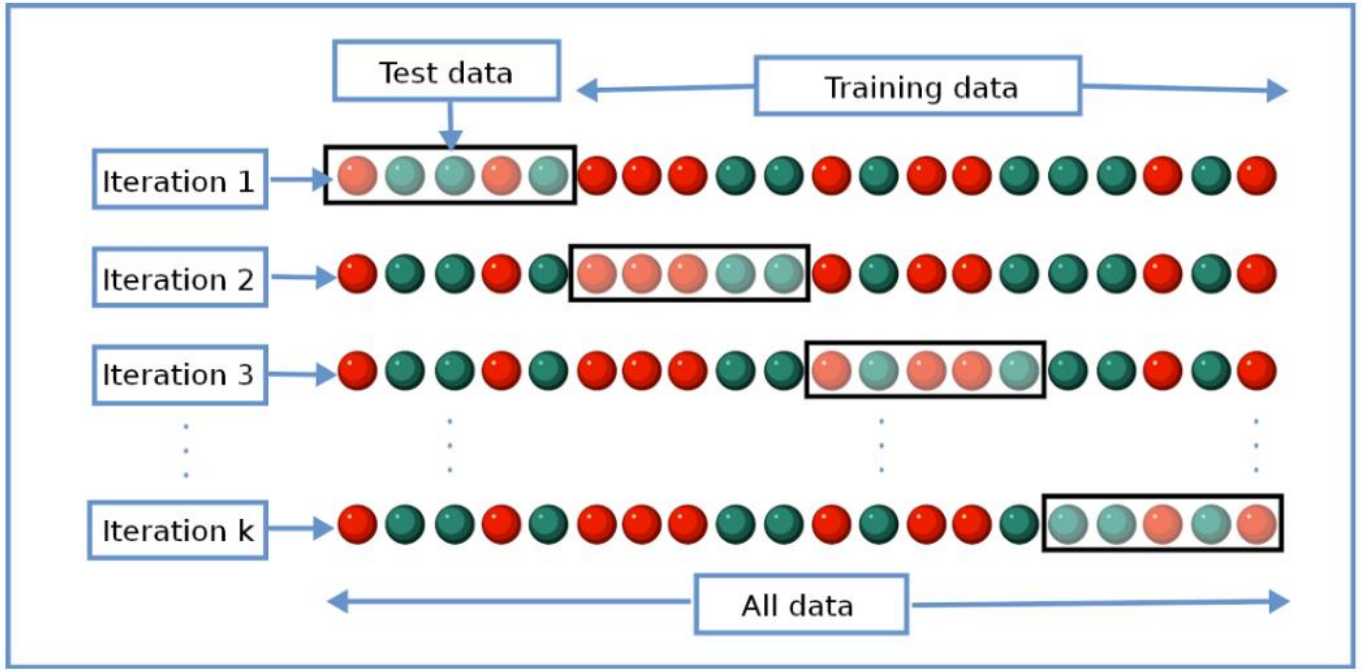
Validation of PGS

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



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

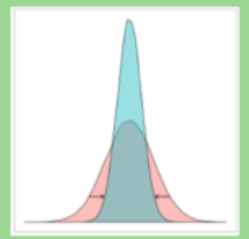
- Both data sets QCed as standard in GWAS
- Some QC requires special care in PRS (e.g., sample overlap, relatedness and population structure)
- Retain set of SNPs that overlap between base and target data

PRS calculation

LD adjustment

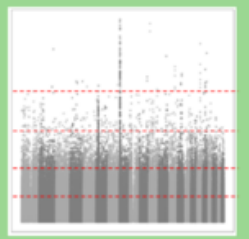
- e.g., clumping

Beta shrinkage



- e.g., LASSO/ridge

P value thresholding

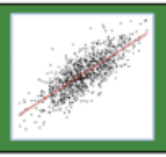


- PRS at multiple *P*

Test

ID	BMI	PRS
101	24.1	0.43
102	28.3	1.61
103	31.2	0.83
104	19.4	3.54

Generate PRS
+
Perform association testing



Validate

Out-of-sample PRS testing

- *K*-fold cross-validation
- Test in data separate from base/target

Choi et al. 2020

3

Polygenic scores

3.4

Limitations

Limitations of PGS

No environmental factors considered

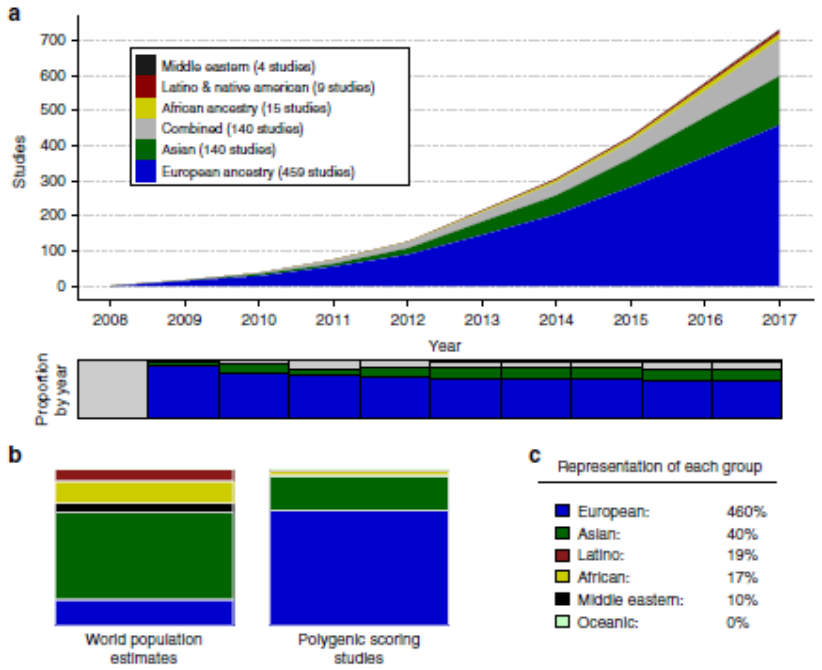
Dependent on homogeneity in discovery and testing samples

Predictive power depends on sample size

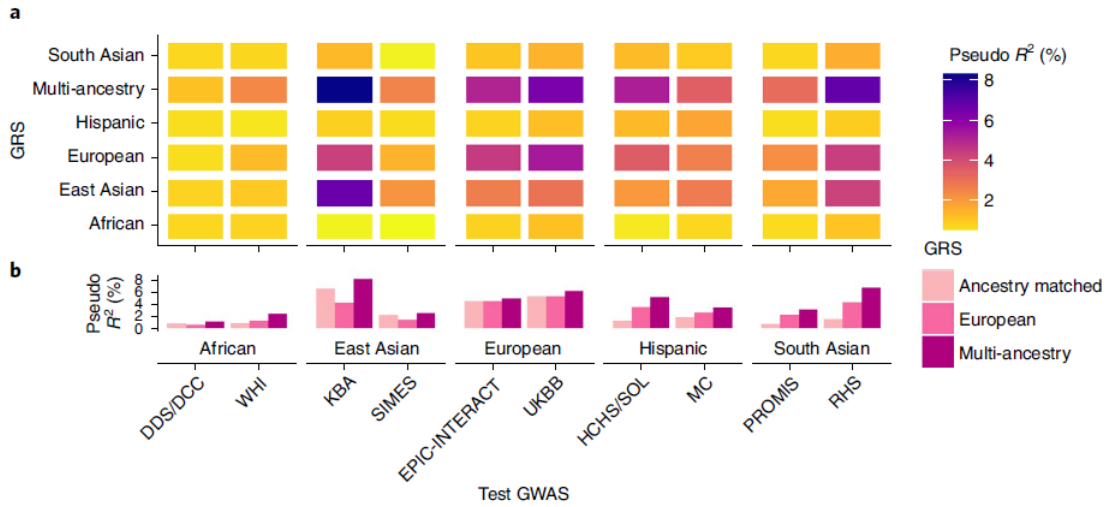
- Low predictive power → limited clinical use
- Low transferability when deviation from original GWAS cohort (e.g. ancestry)

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
  - → meta-regression, …
  - → Decrease health disparities



Duncan et al. Nat. Comm. 2019



Mahajan et al. Nat. Genet. 2022

HELMHOLTZ MUNICH

3

Polygenic scores

3.5

Workshop

Timeline

- Introduction (Exercise 1): 10 minutes
- Manual score in R: 30 minutes (Exercises 2-5)
- Score in Plink: 20 minutes (Exercises 6-7)
- PGS and Polygenicity: 20 minutes (Exercises 8-9)

Thank you.