HELMHOLTZ
MUNICH

Genetic architecture of complex traits and polygenicity

Ana Aruda and Ozvan Bocher
6th of December 2022

Agenda

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

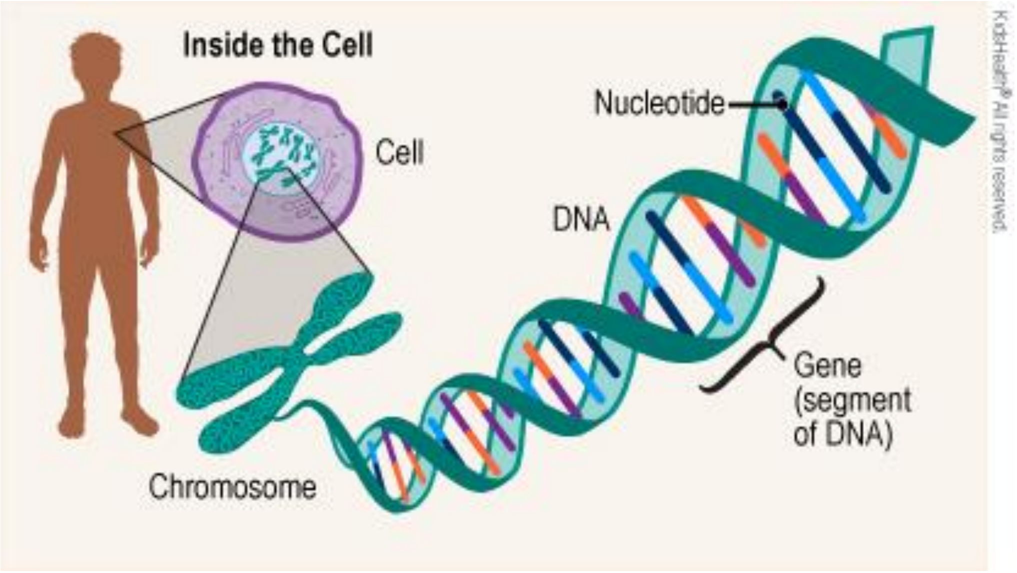
1

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.

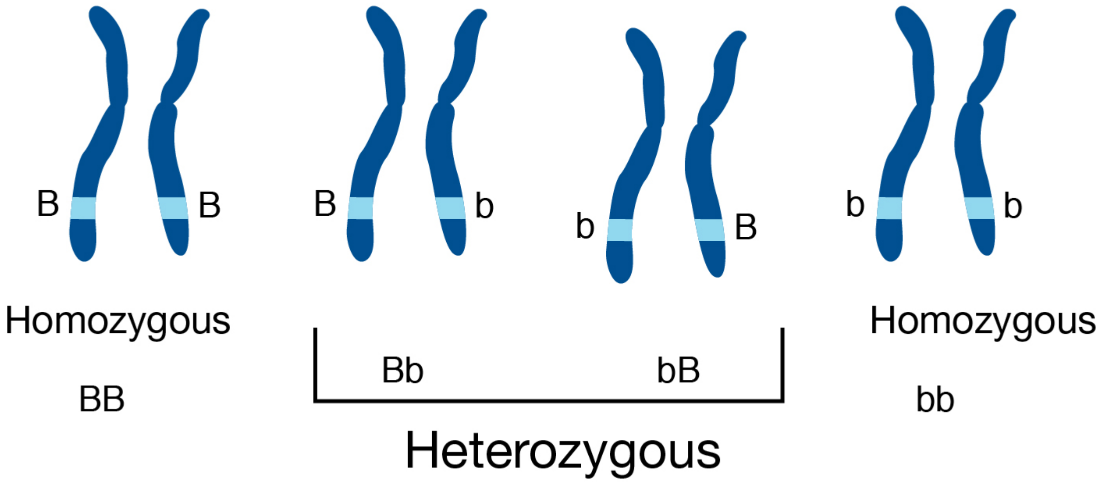


KidsHealth® All rights reserved.

What is an **allele**?

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

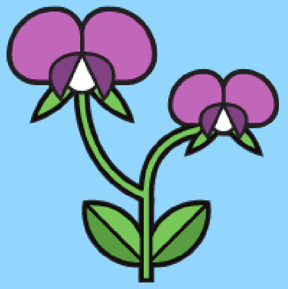
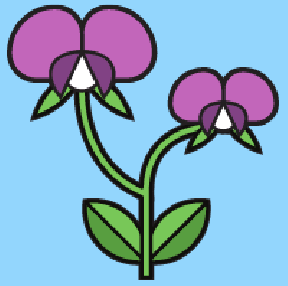
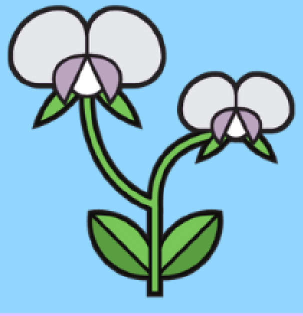
Humans are **diploid organisms**, which means that they have **two alleles at each genetic position**, or locus, with one allele inherited from each parent.



Allele b count			
BB	bB	Bb	bb
0	1	1	2

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

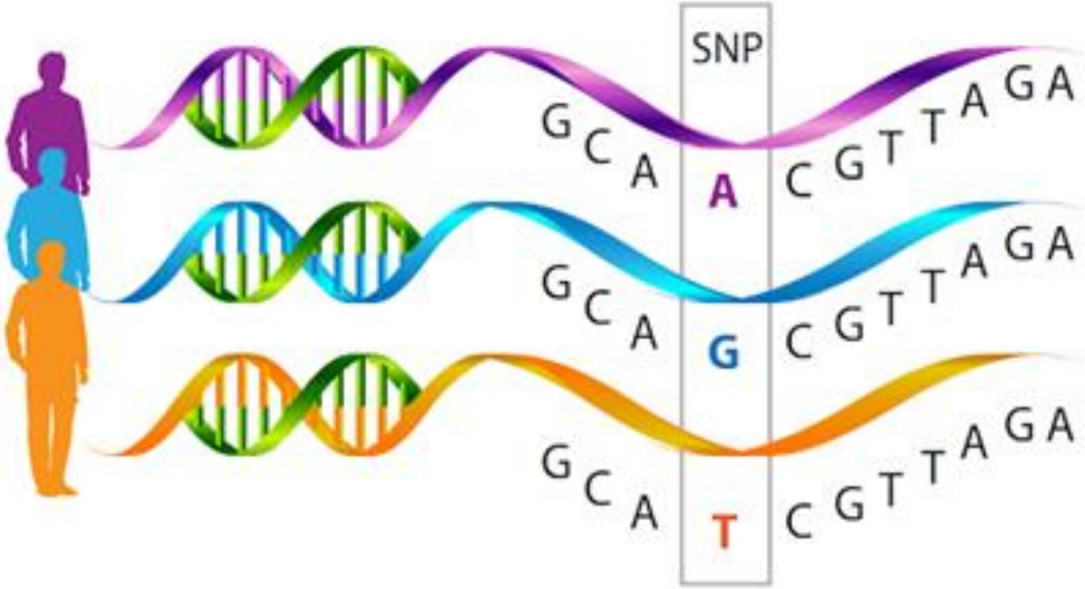
Genotype vs Phenotype

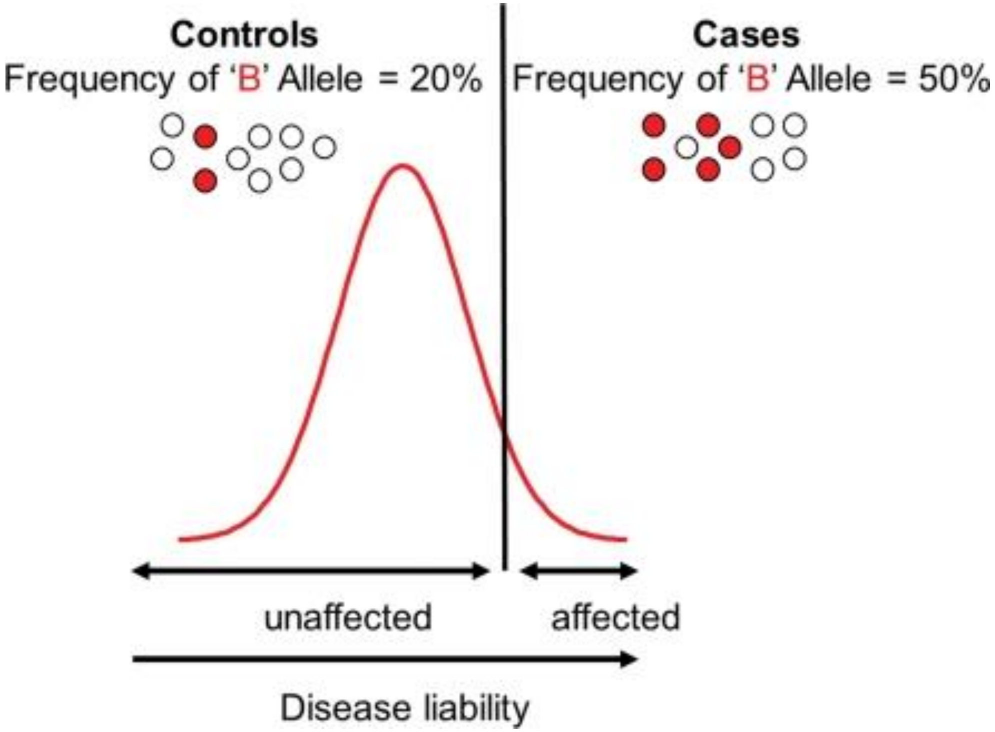
GENOTYPE	PHENOTYPE	
The genotype is an organism's genetic information.	The phenotype is the set of observable physical traits.	
BB homozygous dominant	purple	
Bb heterozygous	purple	
bb homozygous recessive	white	

sciencenotes.org

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** ?

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

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

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

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

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

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

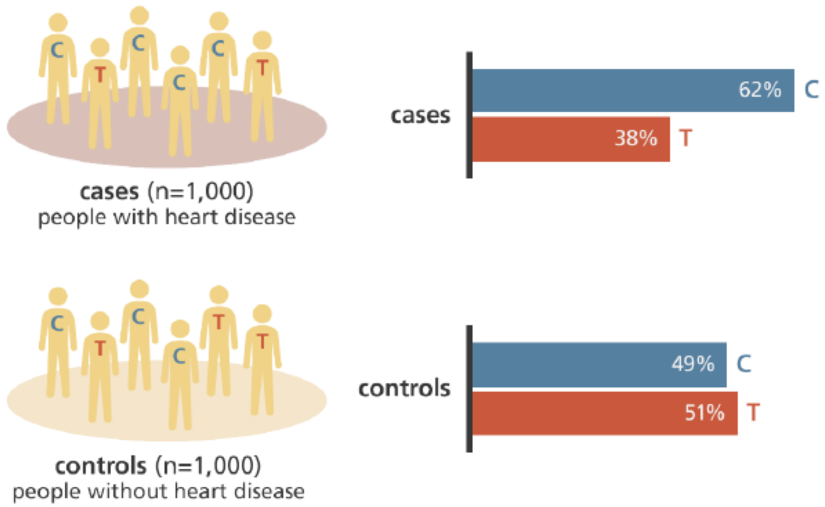
$\in [0,2]$ (dosage, imputed) $\begin{bmatrix} 0.965 \\ \vdots \\ 1.816 \end{bmatrix}$

- 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)

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 (n=1,000)
people with heart disease

controls (n=1,000)
people without heart disease

	Cases	Controls
T	380	510
C	620	490

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

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

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

Case/control studies

OR = ratio of the odds of the two alleles

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

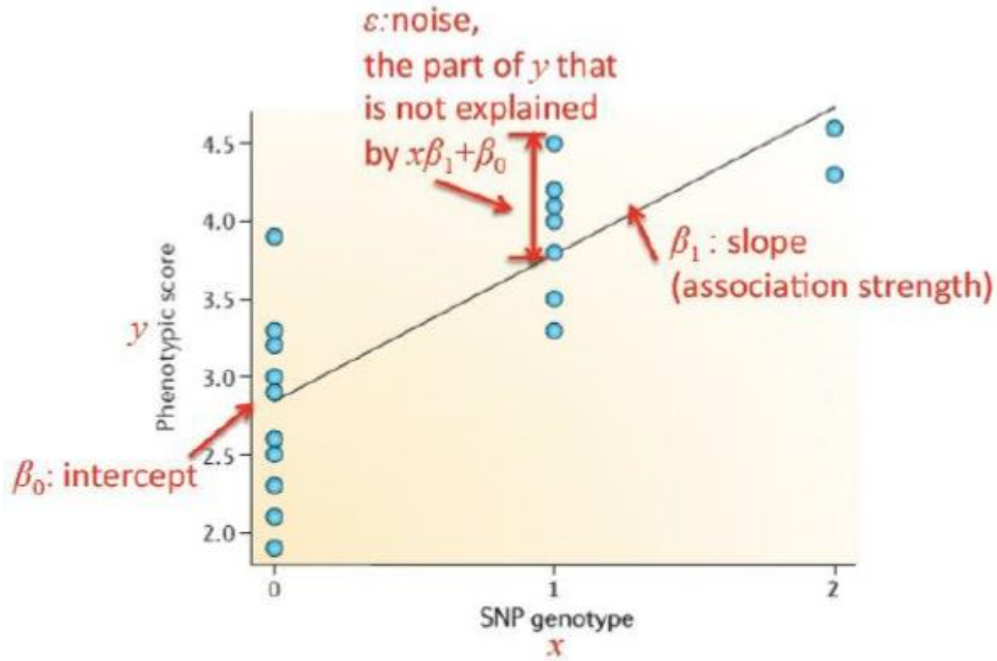
Statistical test: is the OR significantly different from 1?

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

Continuous traits

- A linear regression model is defined as:

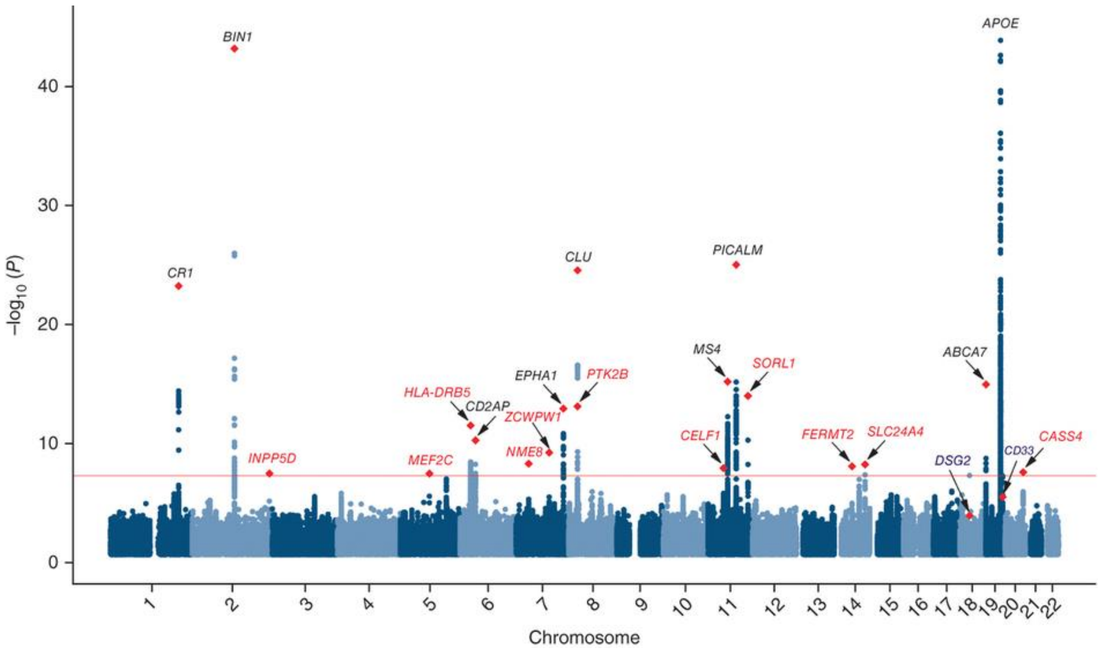
$$y = x\beta_1 + \beta_0 + \varepsilon$$

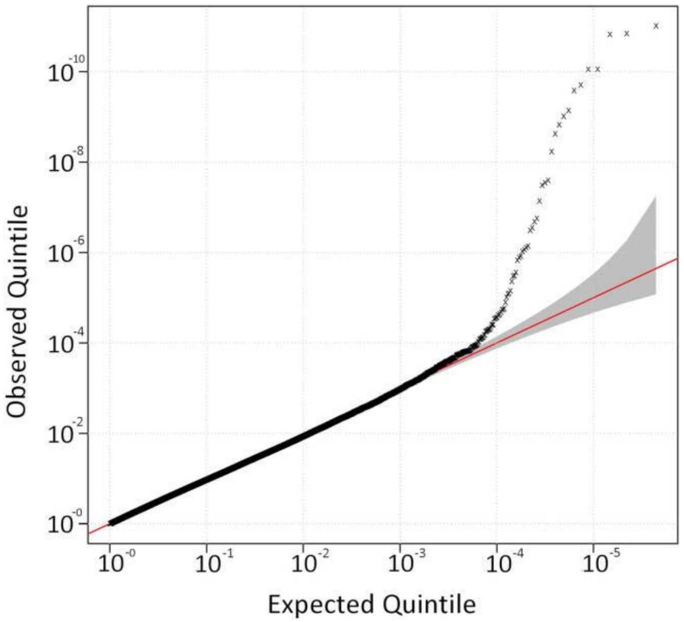


- Data:
  - y is a continuous trait
  - x is the SNP genotype at a given locus
- Parameters:
  - β_1 is the regression coefficient, represents the strength of association between y and x
    - ➢ $\beta > 1$: for every one supplementary allele, the phenotype will increase by the beta coefficient value
    - ➢ $\beta < 1$: for every one supplementary allele, the phenotype will decrease by the beta coefficient value
  - β_0: intercept term (is often ignored)
- Assumptions:
  - The individuals in the study are not related
  - The phenotype y has a normal distribution

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



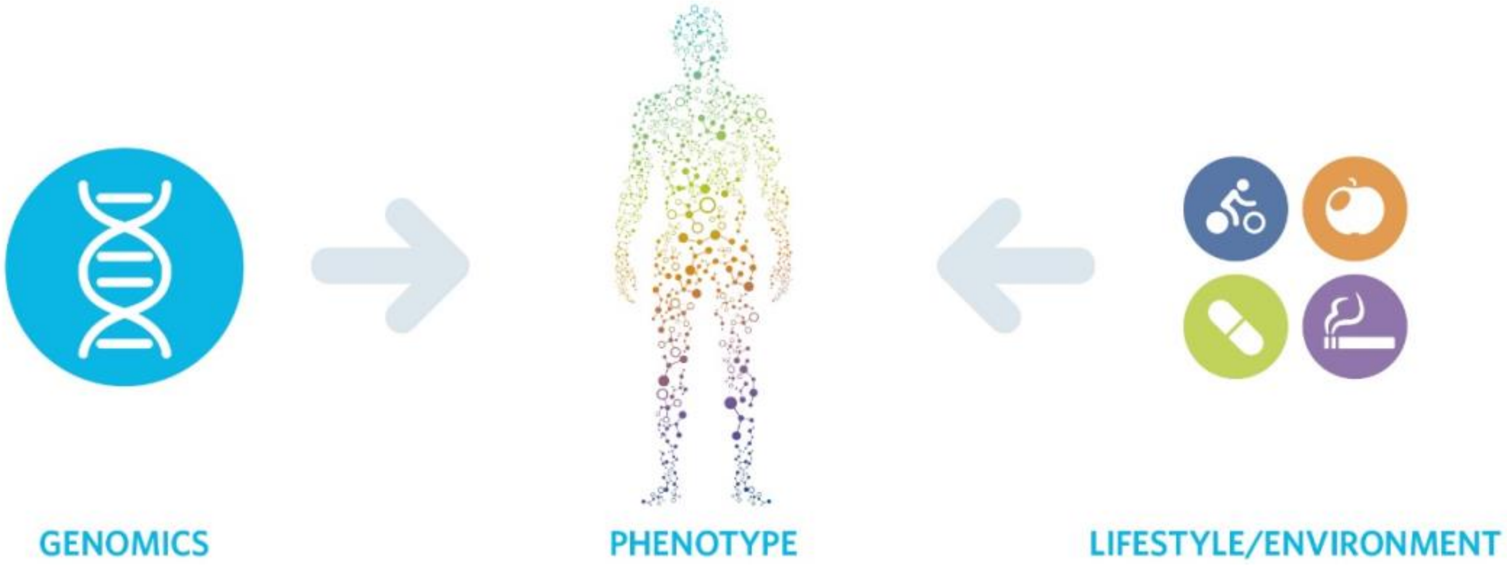


3

Complex traits

Complex traits

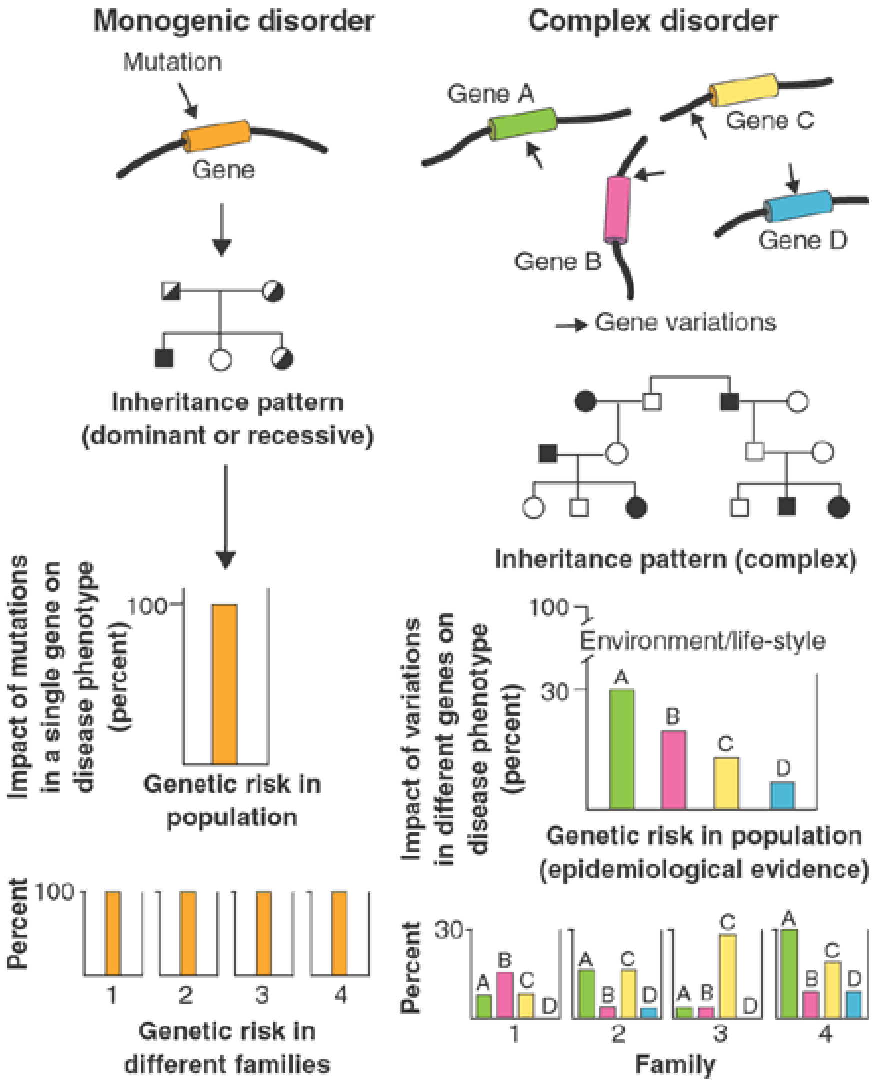
- Complex traits = interaction between (often many) **genetic** and **environmental** factors



- Examples: body shape, type 2 diabetes, Alzheimer’s disease…
- Complex diseases tend to be common

  → Tool of choice = GWAS

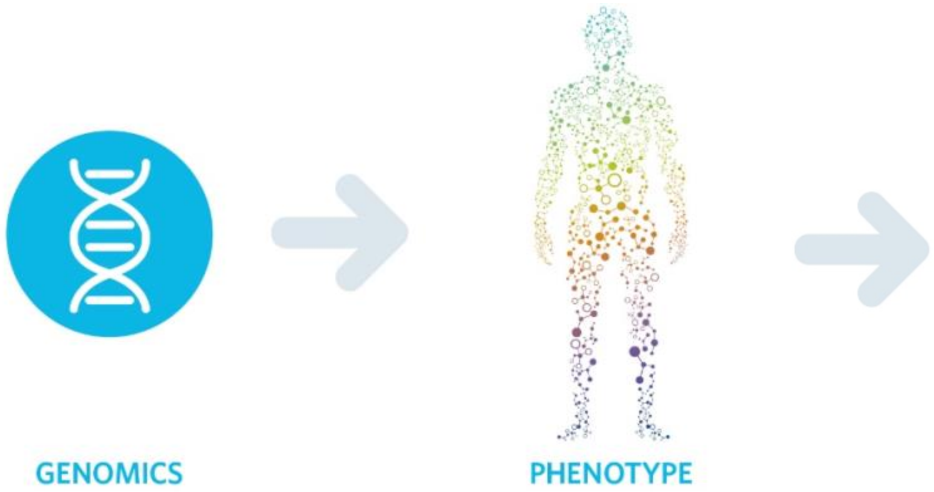
Monogenic disorder vs complex traits



Monogenic disorder
Mutation
Gene
Inheritance pattern (dominant or recessive)
Impact of mutations in a single gene on disease phenotype (percent)
100
Genetic risk in population
Percent
100
1
2
3
4
Genetic risk in different families
Complex disorder
Gene A
Gene C
Gene B
Gene D
Gene variations
Inheritance pattern (complex)
Impact of variations in different genes on disease phenotype (percent)
100
Environment/life-style
30
A
B
C
D
Genetic risk in population (epidemiological evidence)
Percent
30
A
B
C
D
1
2
3
4
Family

Peltonen & McKusick, Science, 2001

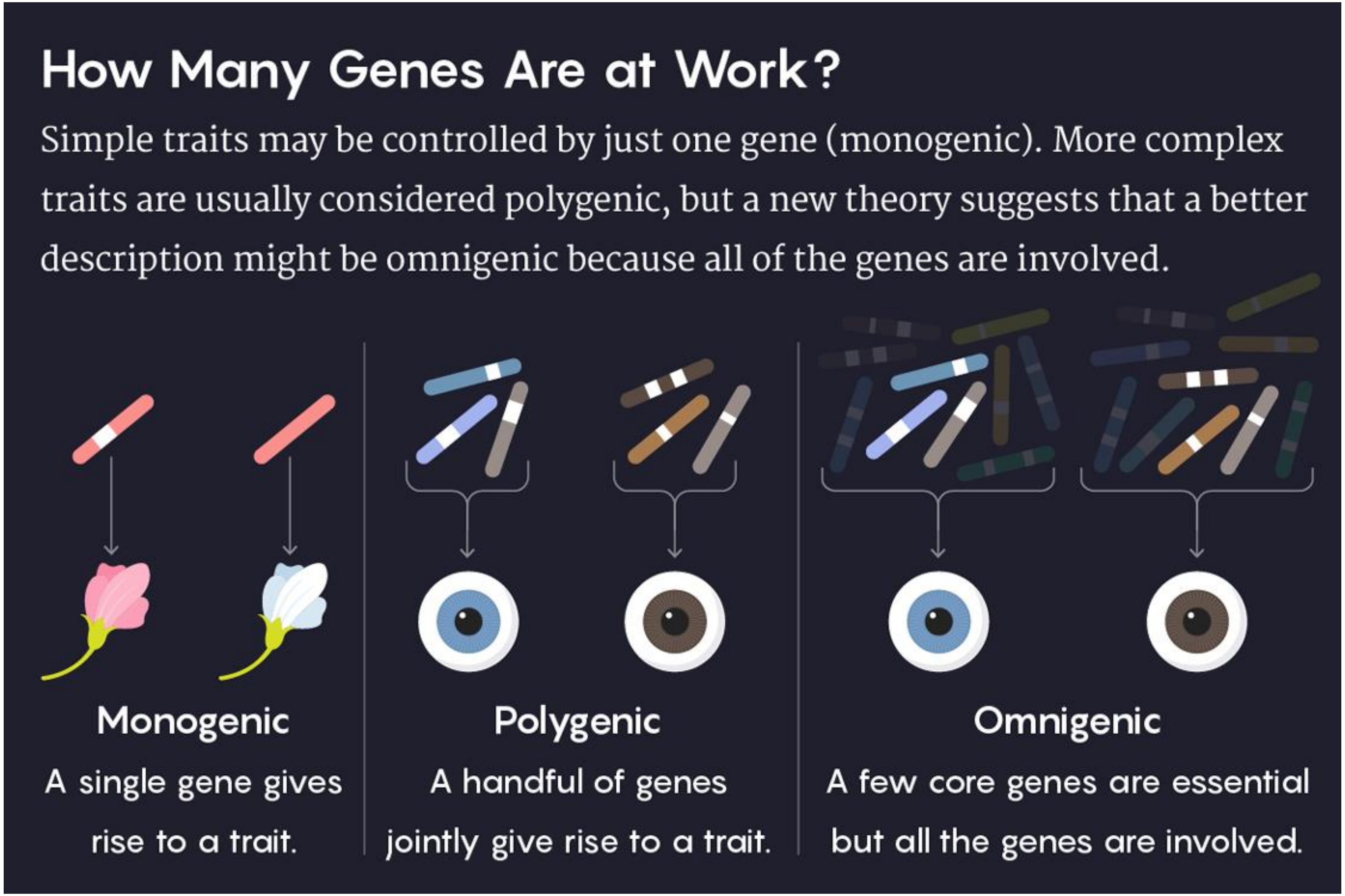
Heritability



- Phenotype = Genetic Effect + environmental effect
- Heritability: $h^2 = \frac{Var[Genetic\ effect]}{Var[Phenotype]}$
  - Proportion of variance in the phenotype that comes from genetics
  - Variance explained by all genetics variations
- SNP heritability: h_g^2
  - How much variance/heritability is explained by a set of SNPs
  - $h_g^2 < h^2$

→Estimate heritabilities with mixed models

→If h_g^2 or h^2 are large, then genetics plays large role on phenotype



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.

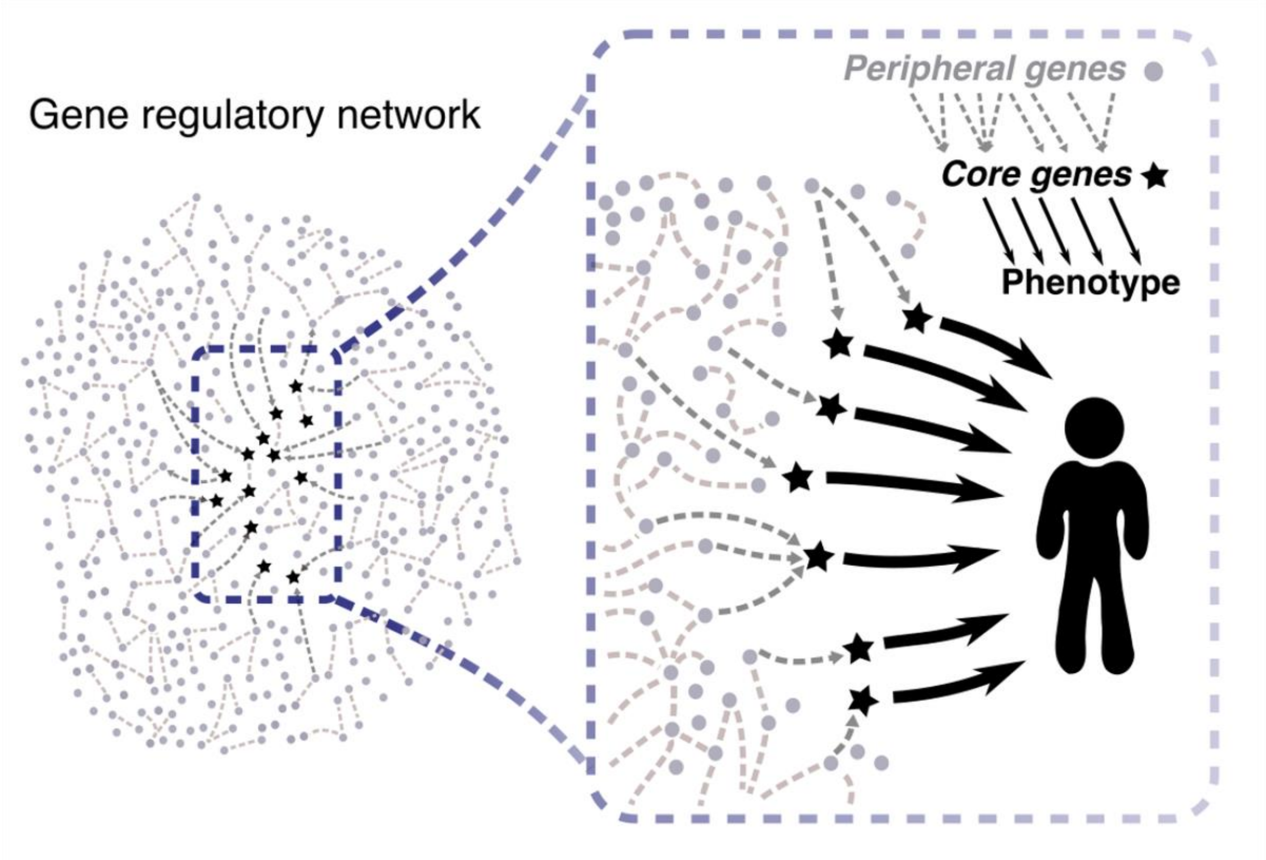
Polygenic

A handful of genes jointly give rise to a trait.

Omnigenic

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

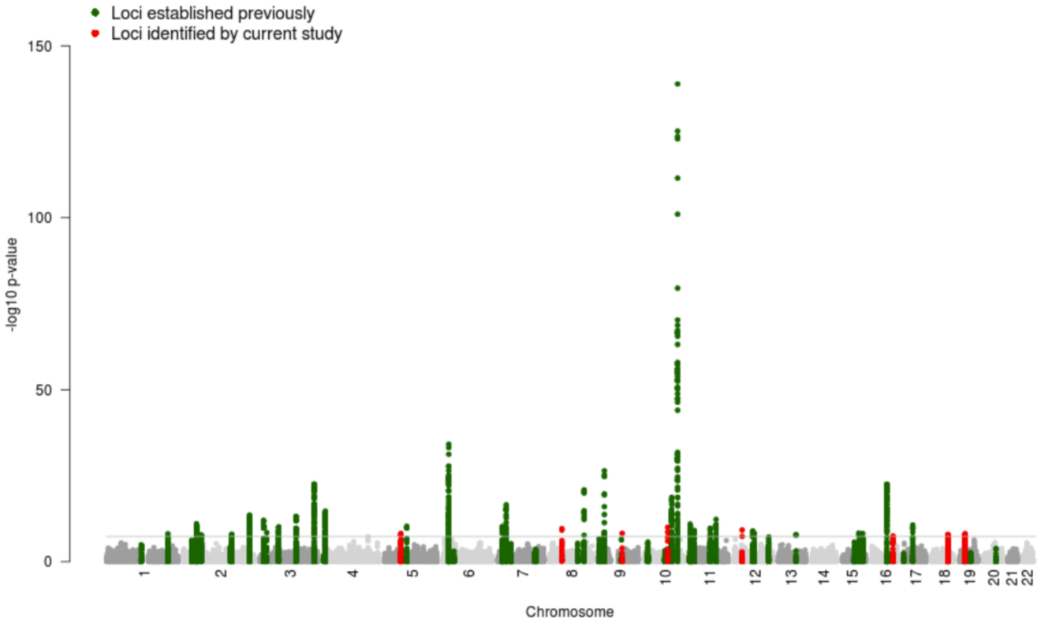
Omnigenic vs Polygenic model



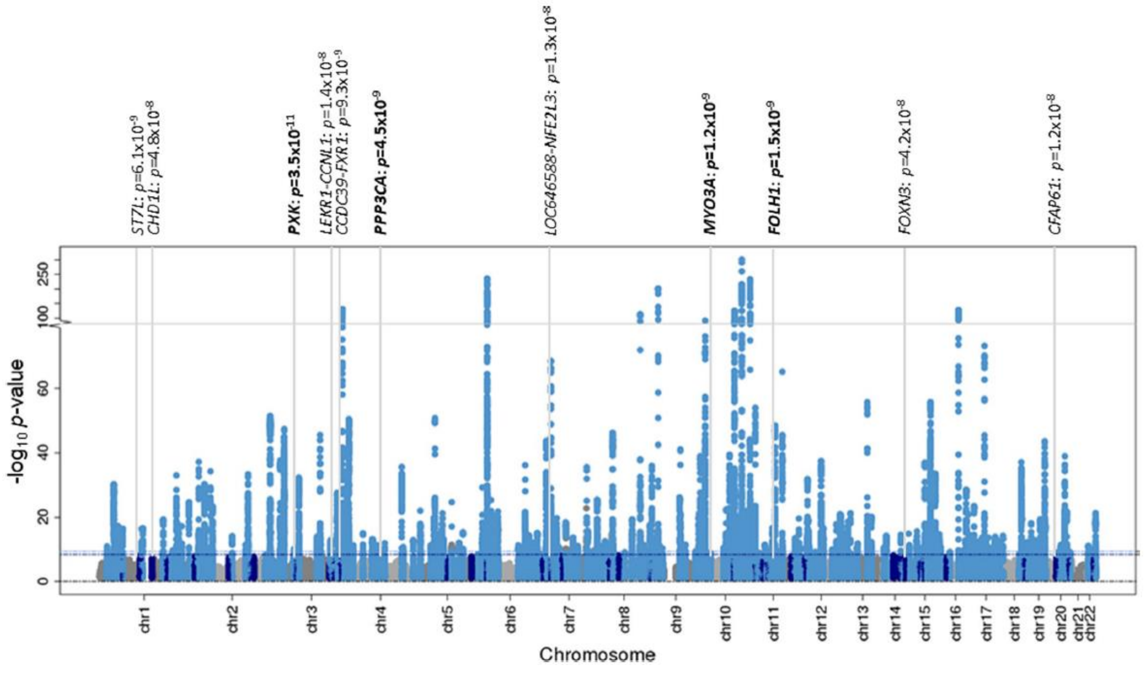
GWAS

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



HELMHOLTZ MUNICH

3

Polygenic scores

3.1

Introduction

Polygenic scores

- Natural follow up: combine SNPs effects into a score
  - Many genetic variants influence a complex trait
  - GWAS gives an effect for each variant → use those estimates!
  - Additive model → each copy of the effect allele increases risk

- Used to : predict quantitative traits or disease risk (= polygenic risk score)
- Larger sample size for GWAS → increased predictive power of PGS
- All SNPs have in principle non-zero weights (very small contribution)

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 or disease risk (= polygenic risk score)
- Larger sample size for GWAS → increased predictive power of PGS
- All SNPs have in principle non-zero weights (very small contribution)

Polygenic scores (PGS)

Polygenic score for individual *i*

Total number of SNPs included in PGS

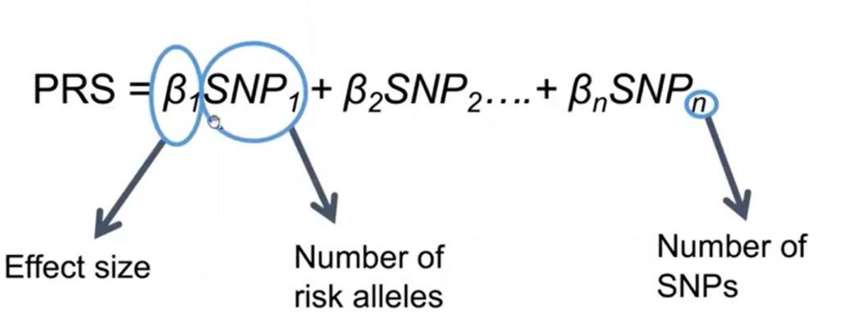
$$PGS_i = \sum_{j=1}^{N_{snps}} G_{ij} * \beta_j$$

Effect of variant *j* on trait

- Estimated in GWAS

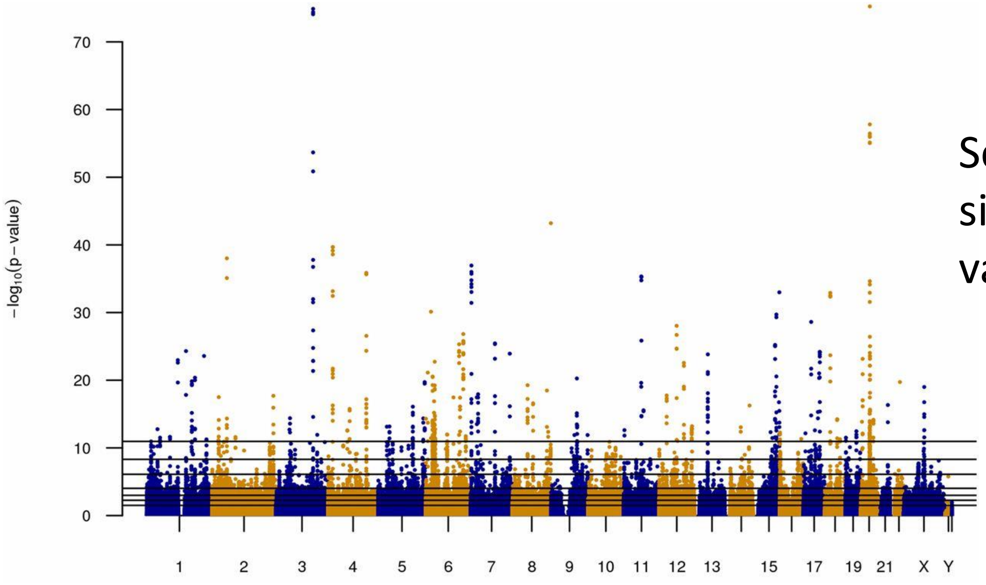
Genotype at SNP *j* for individual *i*

- 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

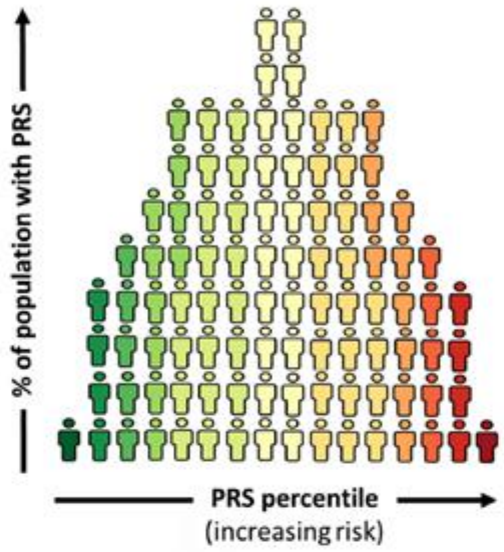
Polygenic scores

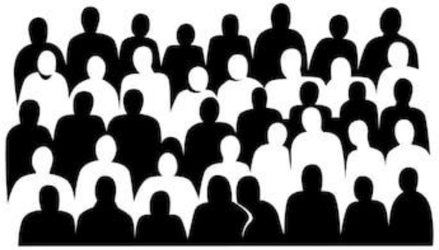


Large GWAS

Scores for significant variants × Genetic data =

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





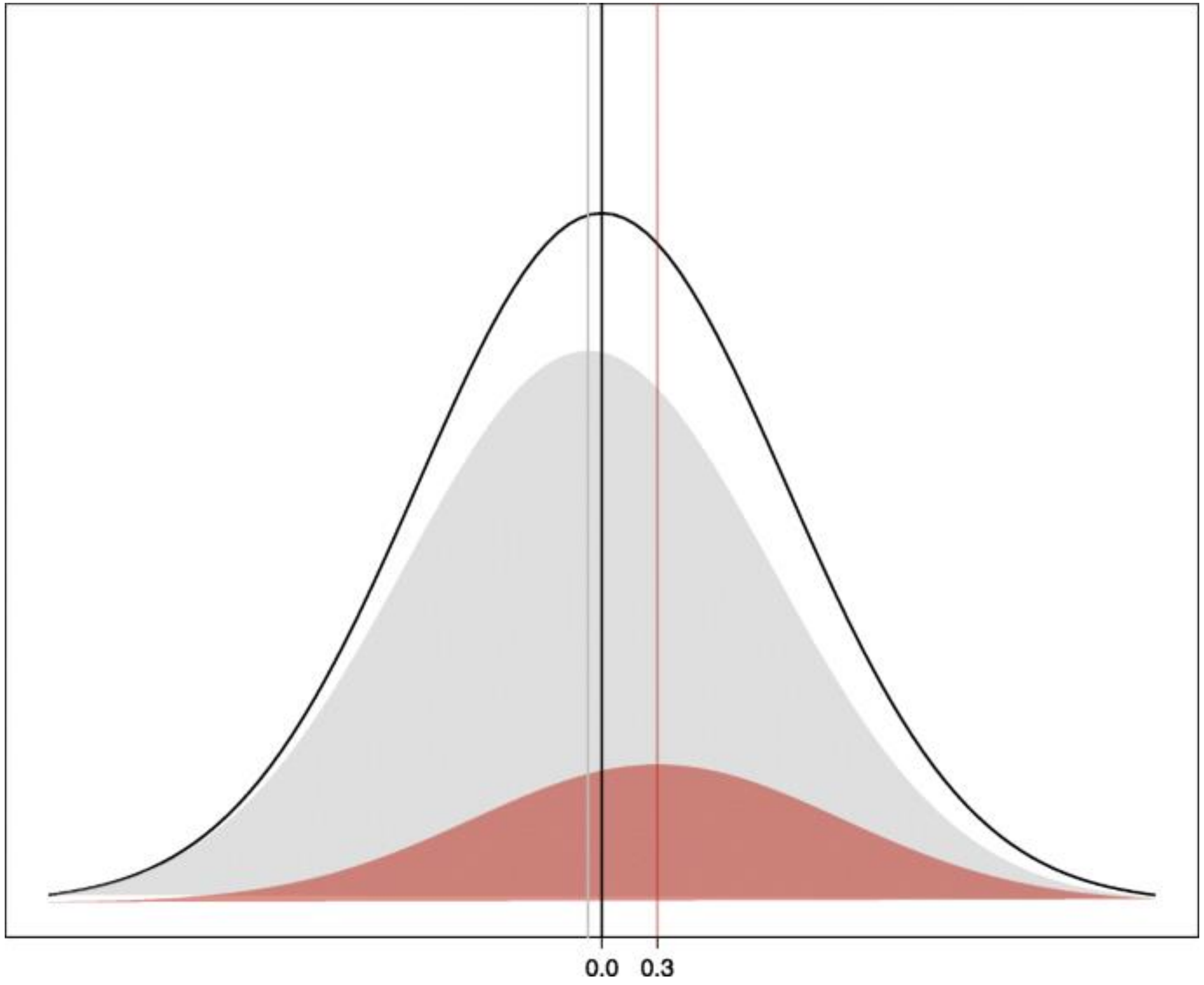
External cohort

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

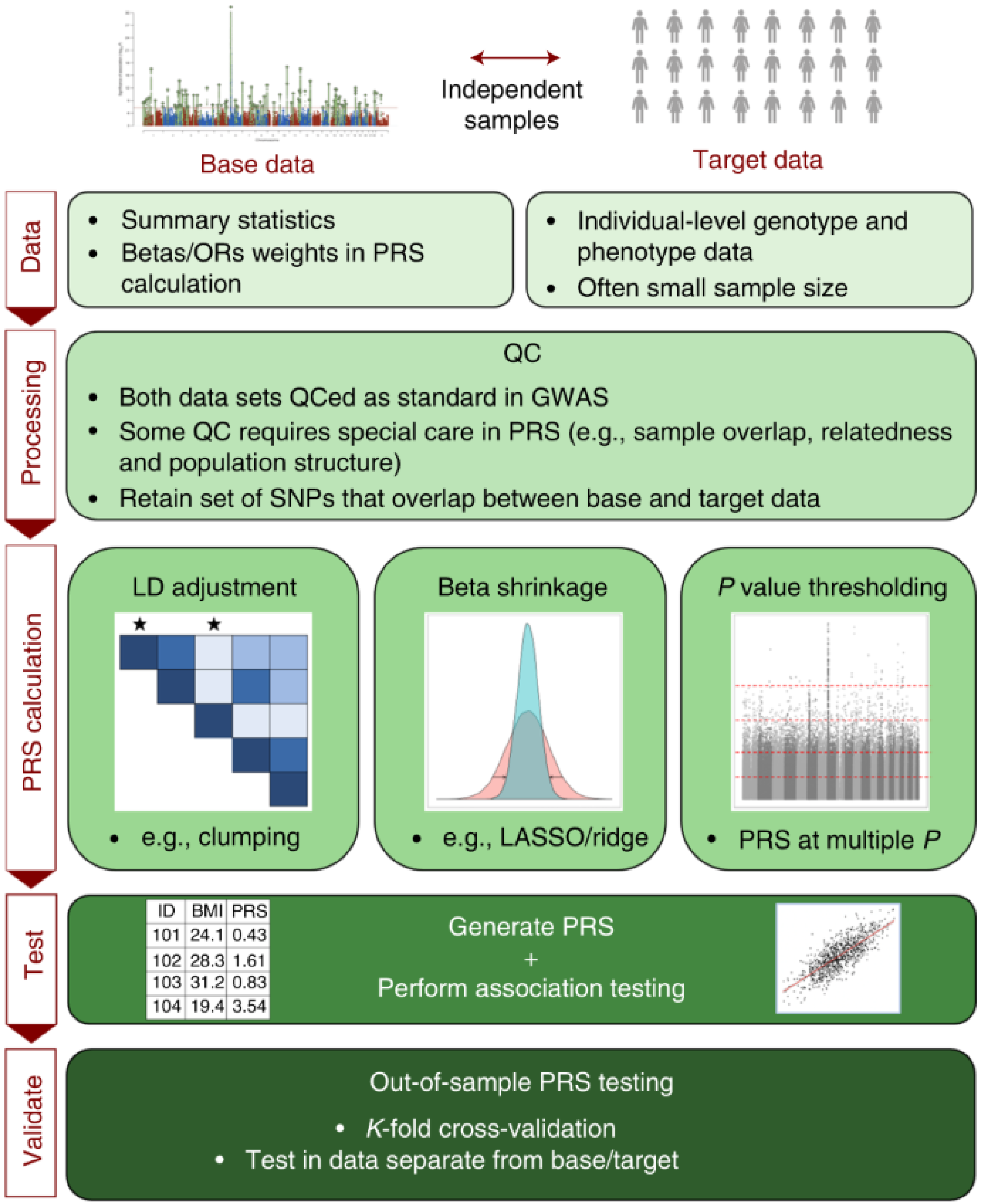


3

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

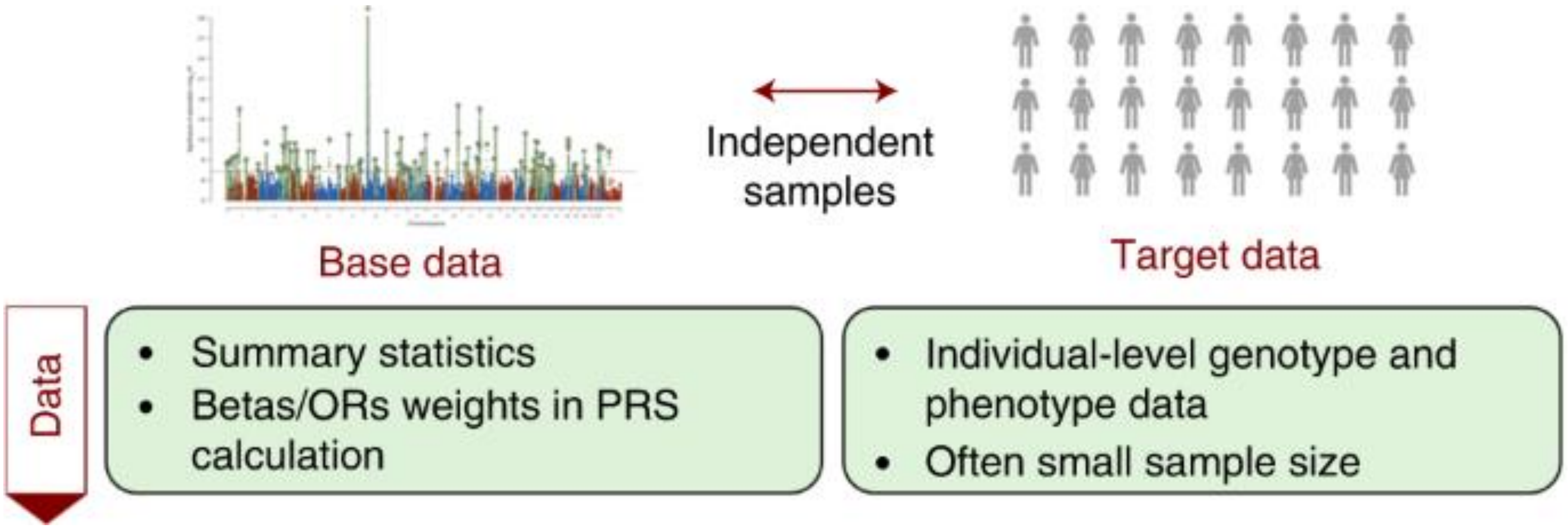
Validate

Out-of-sample PRS testing

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

Choi et al. 2020

Input data



- Summary statistics
- Betas/ORs weights in PRS calculation

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

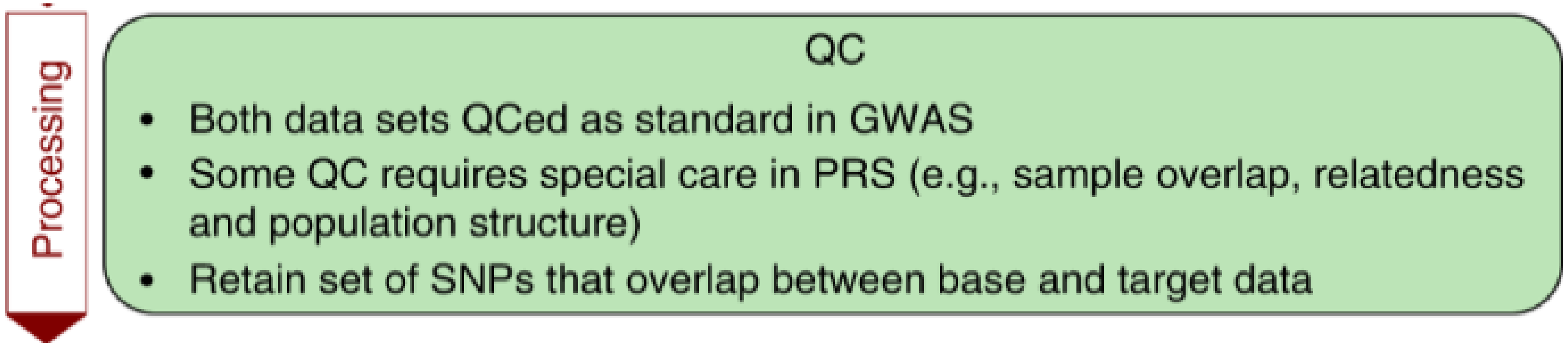
Sample used to estimate parameters for the PGS
→ Largest GWAS summary statistics
→ We need:

- 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

Goal: apply on real patients

Data processing



QC

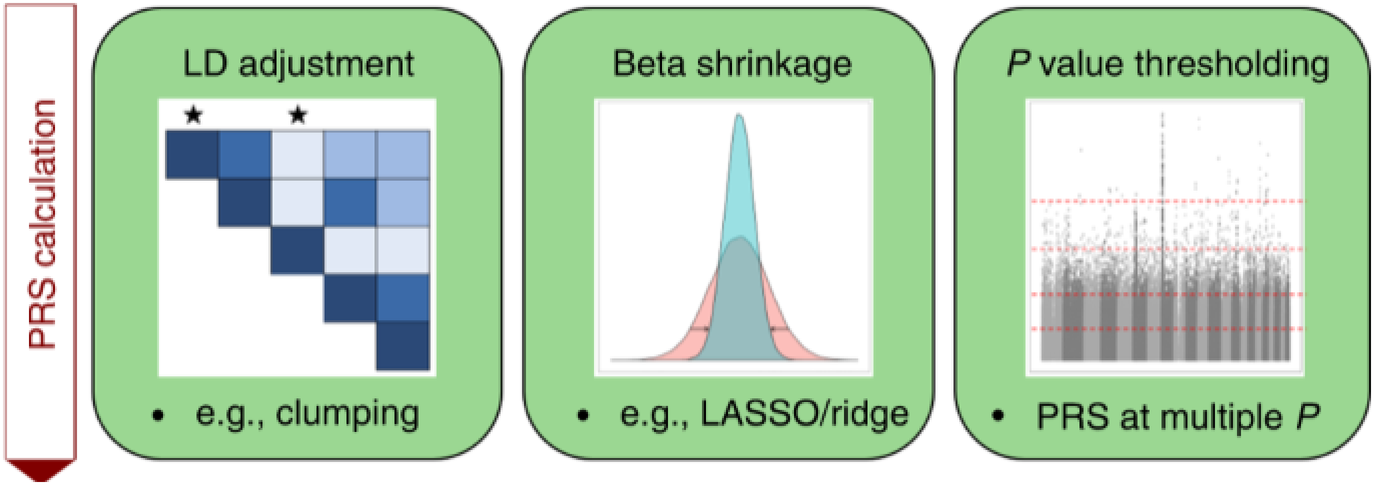
- Both data sets QCed as standard in GWAS
- Some QC requires special care in PRS (e.g., sample overlap, relatedness and population structure)
- Retain set of SNPs that overlap between base and target data

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

- Need homogeneity between base and target samples
  → Hypothesis = sample underlying genetic architecture
  → Also suppose homogeneity in environment

- Population structure
  → Match the ancestry between base and target samples
  → Heterogeneity between population = overall poor transferability from one ancestry to another
  → Move to trans-ancestry PGS

PGS calculation



How to select variants influencing complex traits?

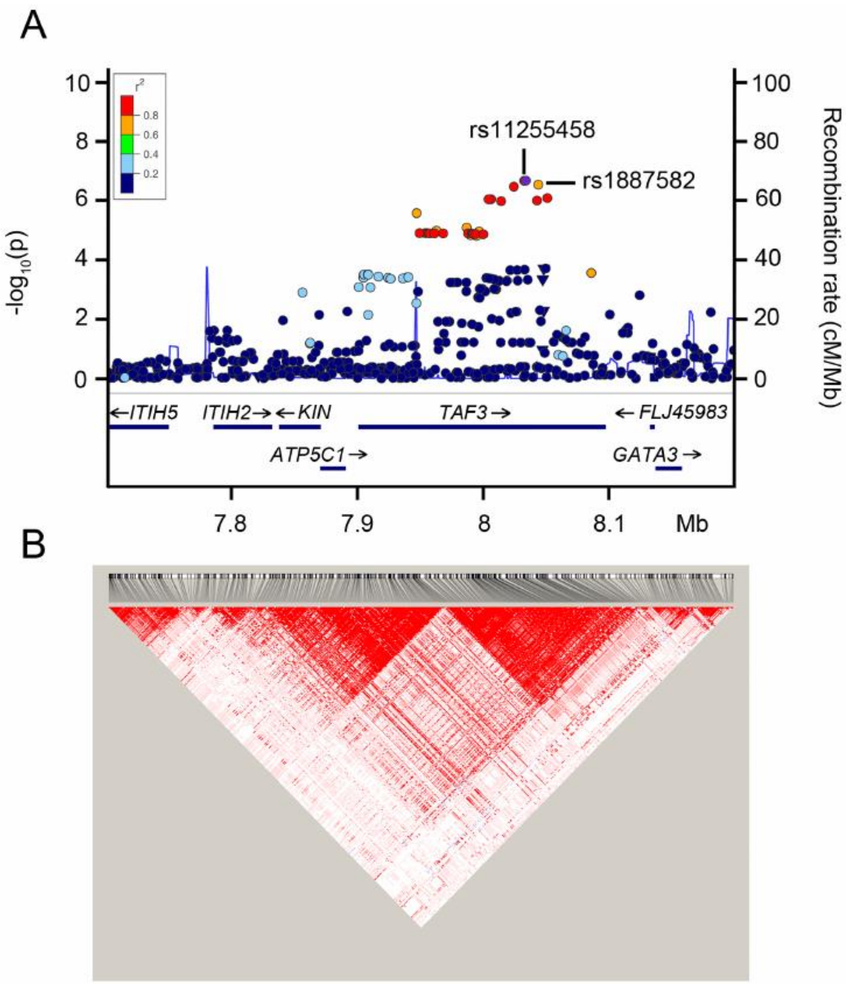
Selection of variants for PGS calculation

- Historically: independent top variants
  → Challenging in omnigenic and polygenic models
  → With more power, more peaks appear

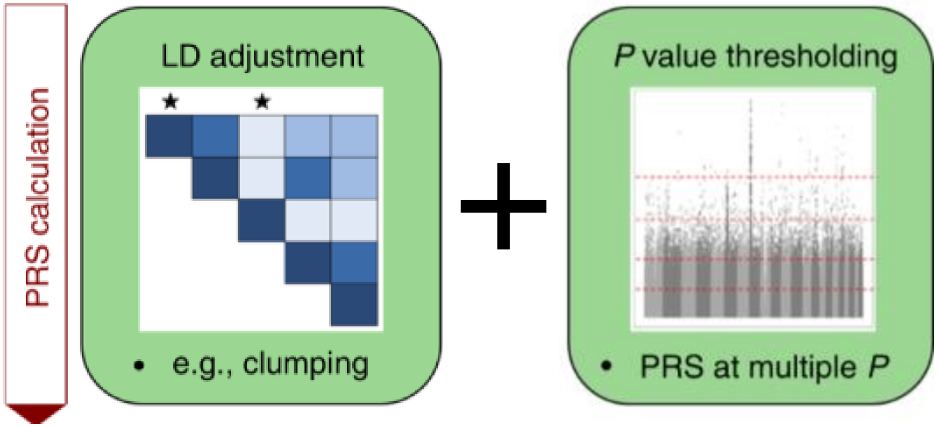
- Solution: use all variants (omnigenic model)
  → Linkage disequilibrium (LD) issue

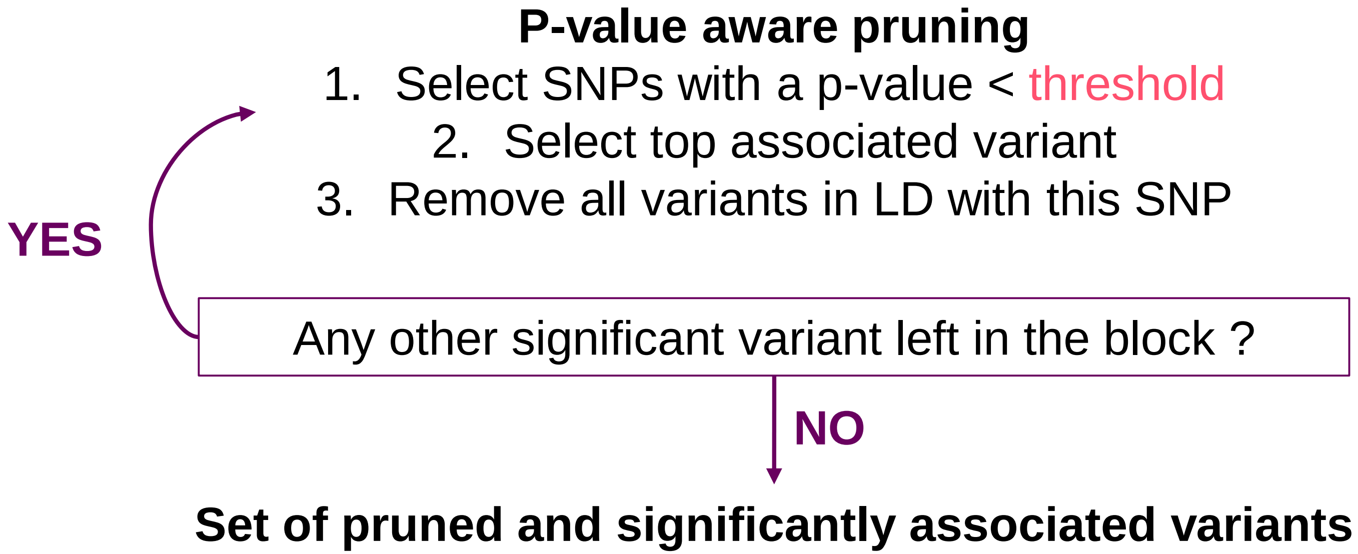
What is LD?

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



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

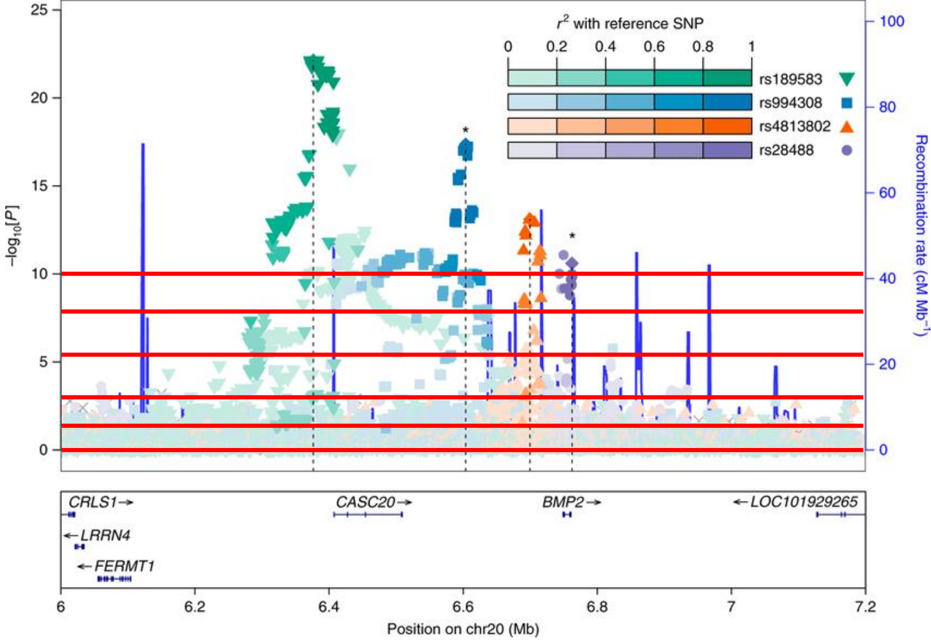
Any other significant variant left in the block ?

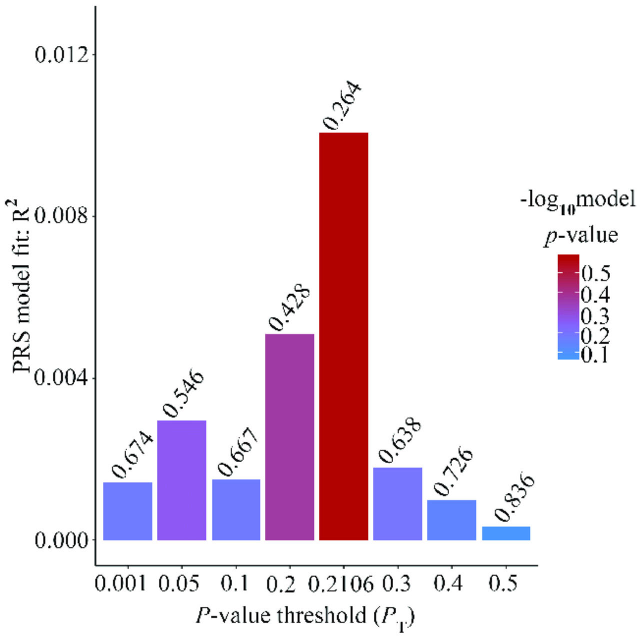
NO

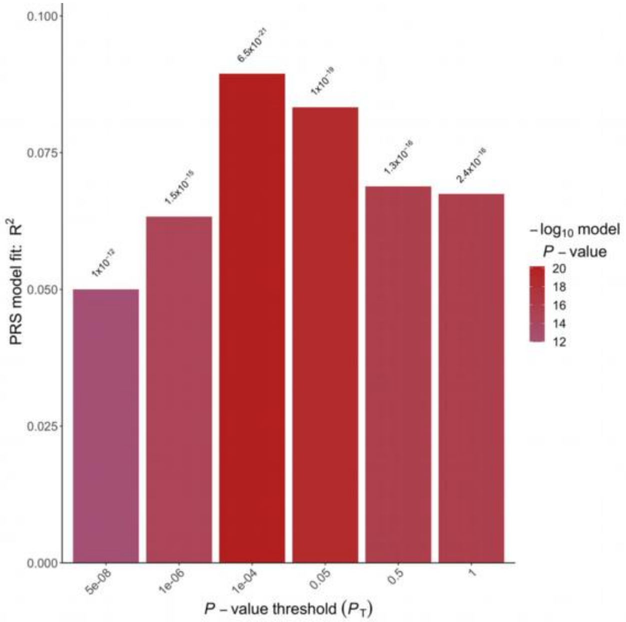
Set of pruned and significantly associated variants

Clumping + Thresholding (C+T)

- Which significance threshold to use ?
  - → Optimal threshold depends on the trait
  - → More polygenicity = more variants → increase threshold
- Unknown beforehand
  - → Try multiple values with validation
  - → Integrated into PGS calculation software, e.g. *PRSice*







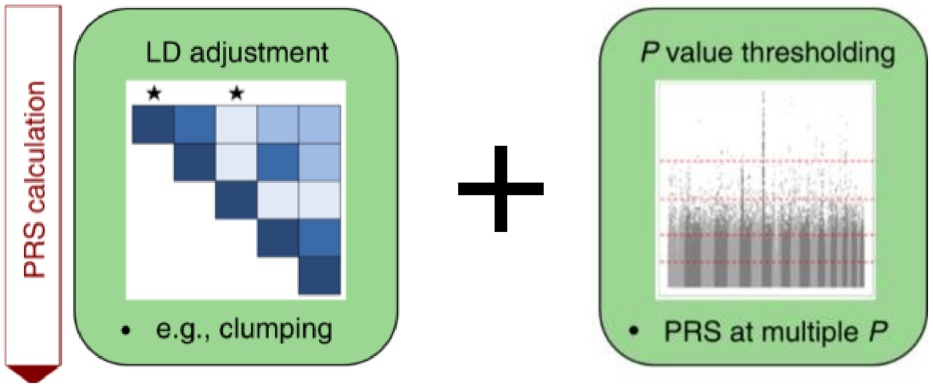
HELMHOLTZ MUNICH

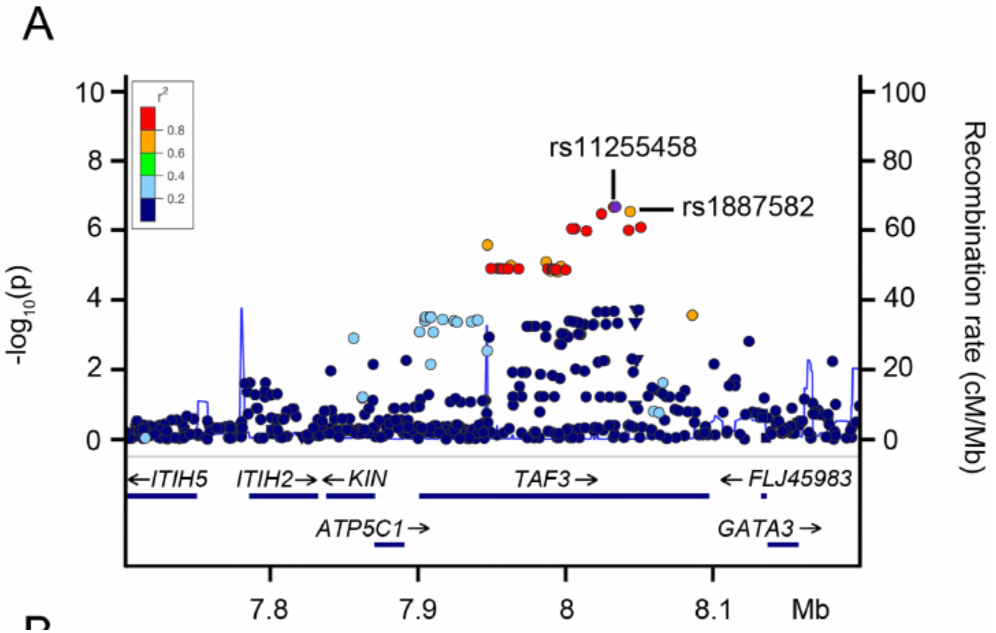
Wang et al. Frontiers in Genetics, July 2019

Maj et al. Frontiers in Cardiovascular Medicine, Feb 2022

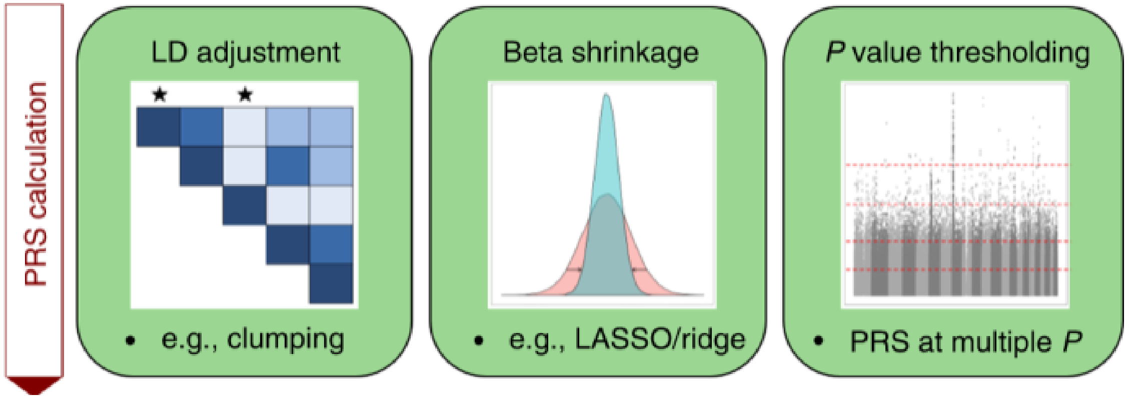
Limitations of C+T

- 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
- Ideal model = ‘whole-genome’ model
  - → Account for LD
  - → Perform “shrinkage” estimation for association coefficients
- 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 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 variants

→ Shrinkage method that produces scaled weights genome-wide

- Downsides: too many hyperparameters → harder to interpret

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

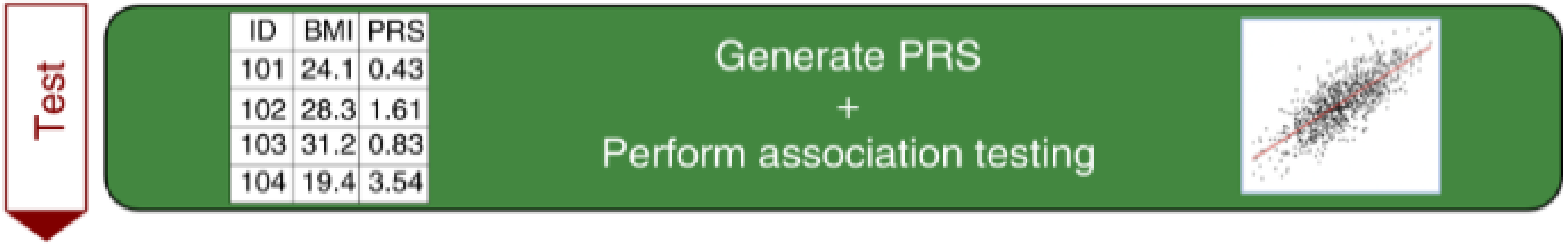
3

Polygenic scores

3.3

Application

Applying PGS



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

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

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

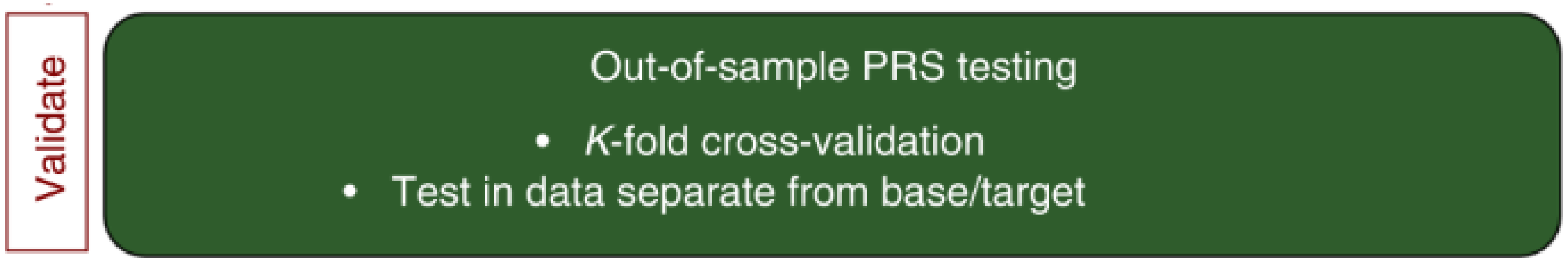
$$\beta_{rs1234,A} = 1.56$$
$$alleles_{rs1234} = \{A, T\}$$
$$\Rightarrow \beta_{rs1234,T} = -1.56$$

- 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



Validate

Out-of-sample PRS testing

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

- Values to assess the prediction of PGS:

→ R2: amount of phenotypic variance explained by PGS (continuous traits)

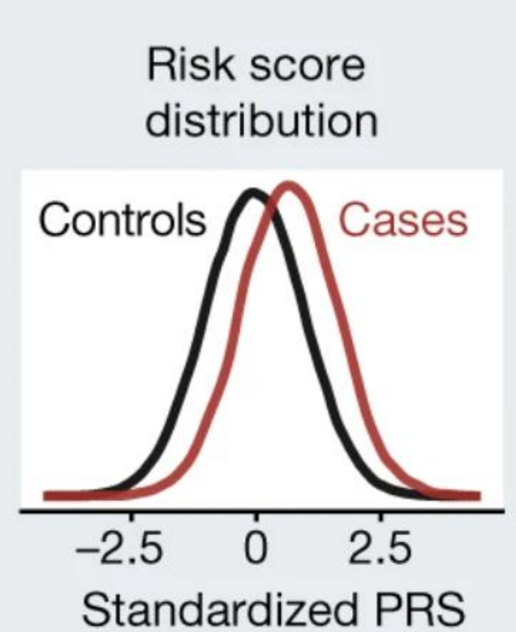
$$R^2 = 1 - \frac{\boxed{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\boxed{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

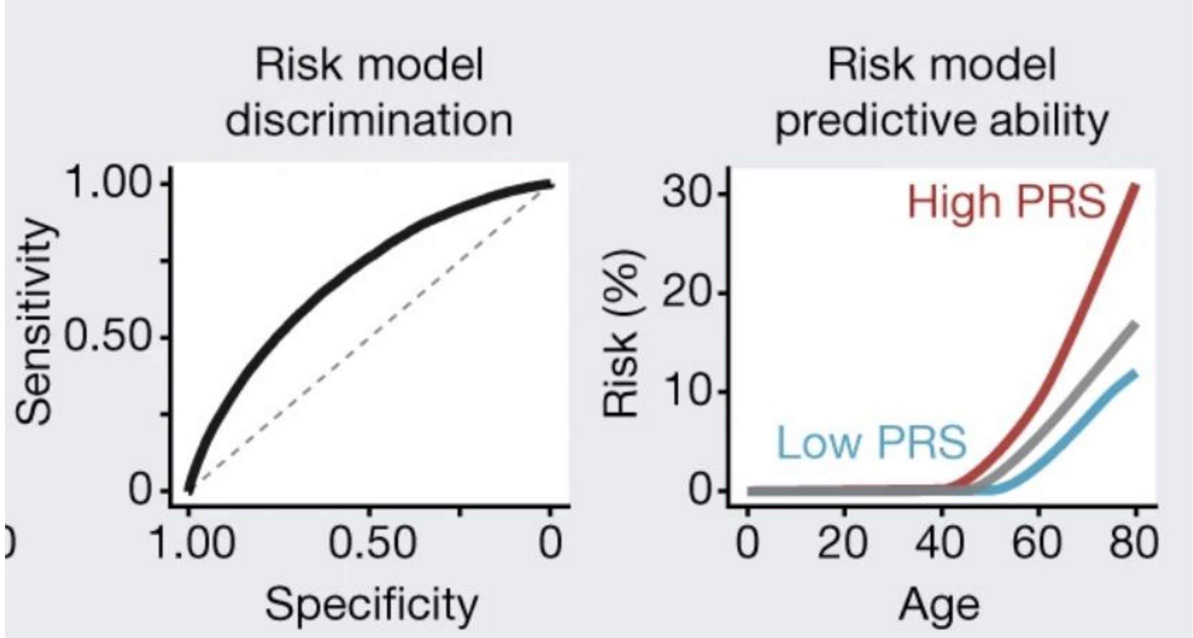
Variability in dependent variable not predicted by the model

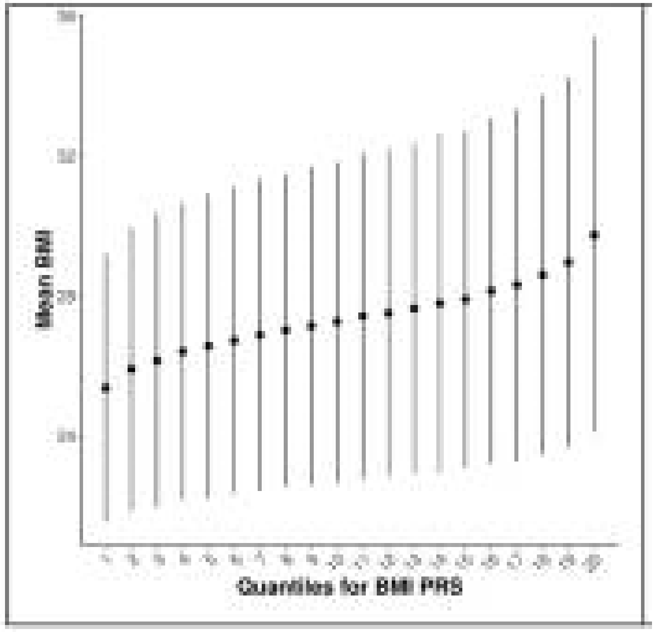
Variability in dependent variable

→ 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

Validate

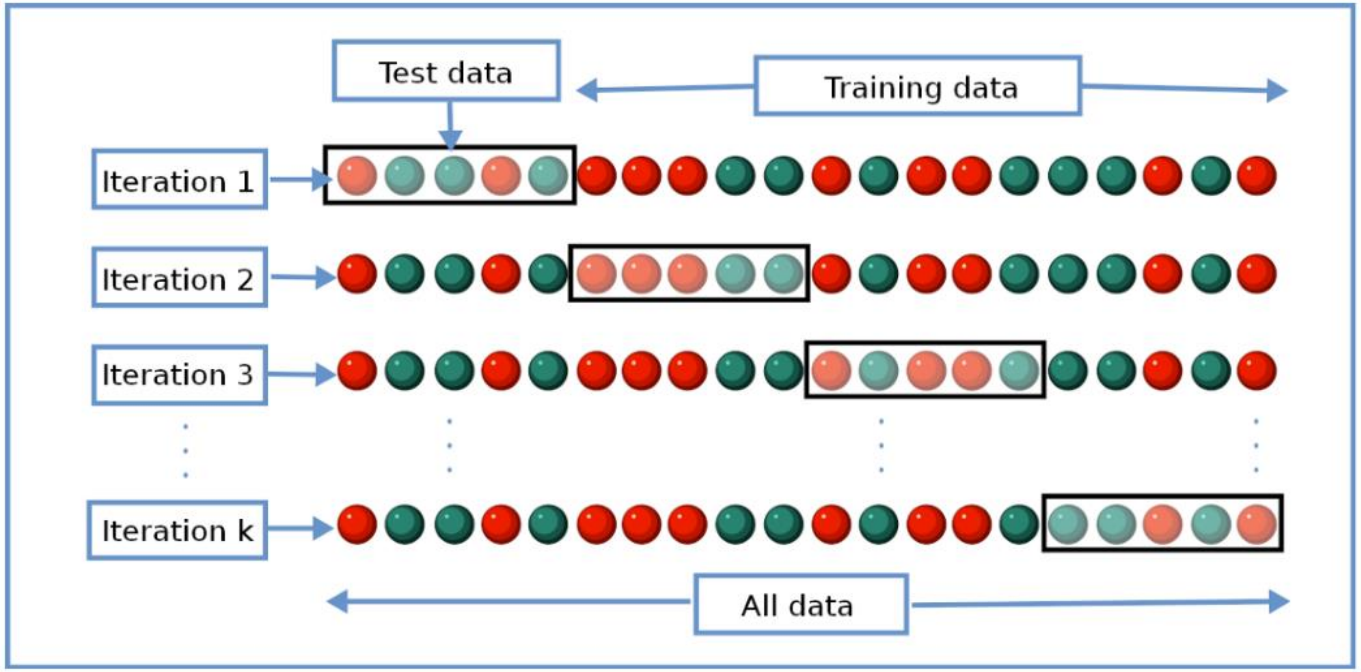
Out-of-sample PRS testing

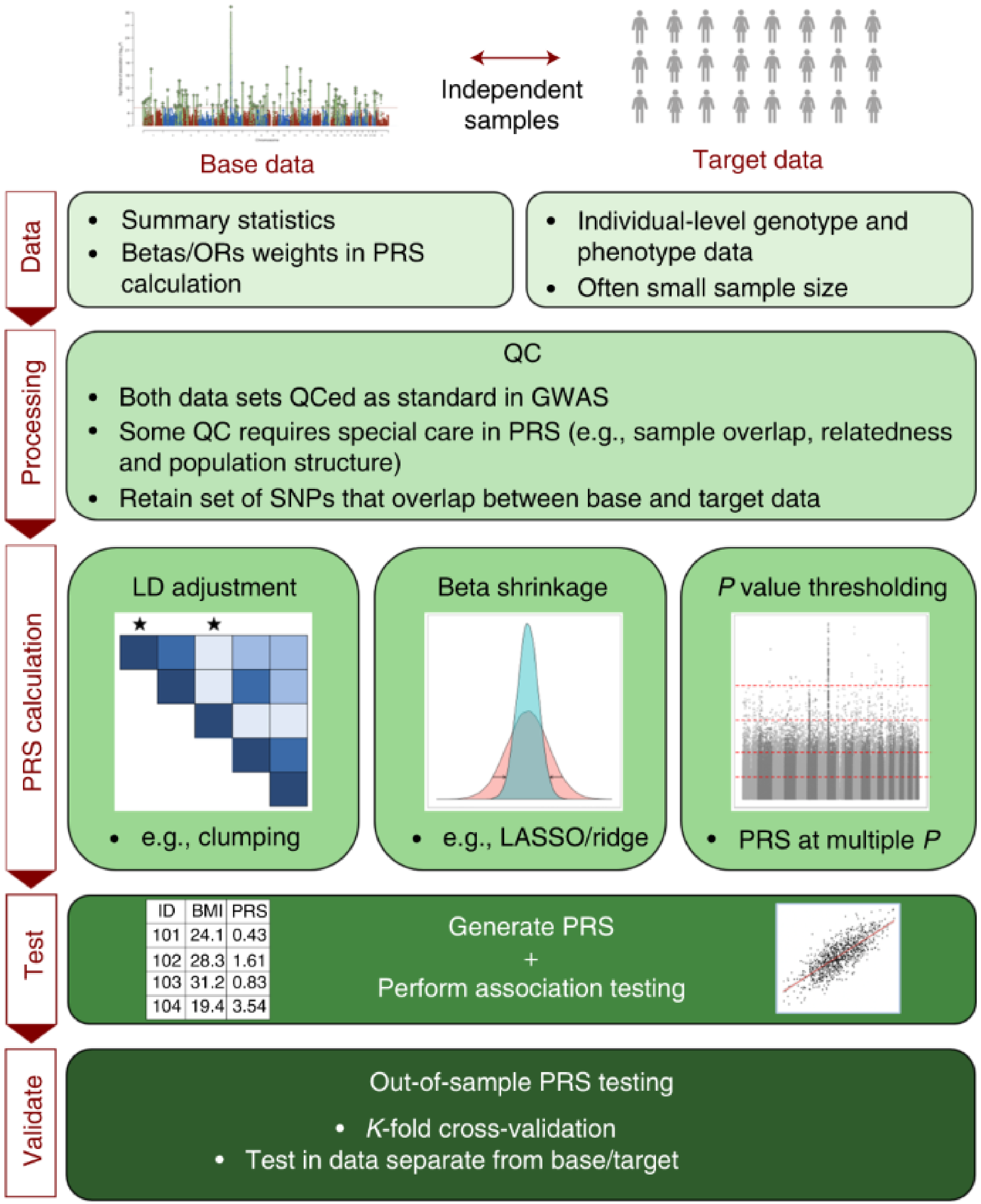
- *K*-fold cross-validation
- Test in data separate from base/target

Validation of PGS

- K-fold cross-validation

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





- Summary statistics
- Betas/ORs weights in PRS calculation

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

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

LD adjustment

- e.g., clumping

Beta shrinkage

- e.g., LASSO/ridge

P value thresholding

- PRS at multiple *P*

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

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

- PGS relies on assumptions:
  → No environmental factors considered
  → Genetic associations = genetic causation
  → Homogeneity in discovery and testing samples

- Depends on:

Heritability

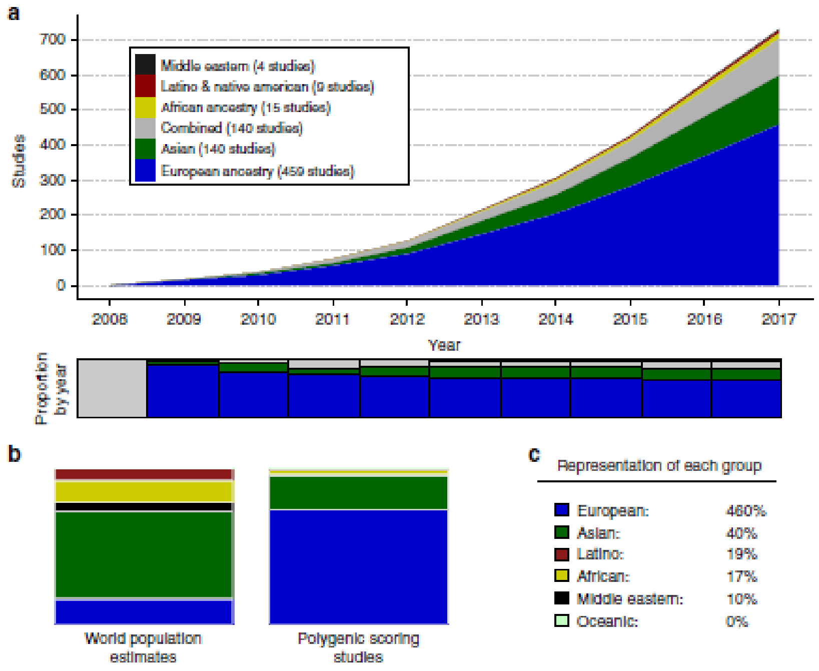
Effect-size distribution

Sample size

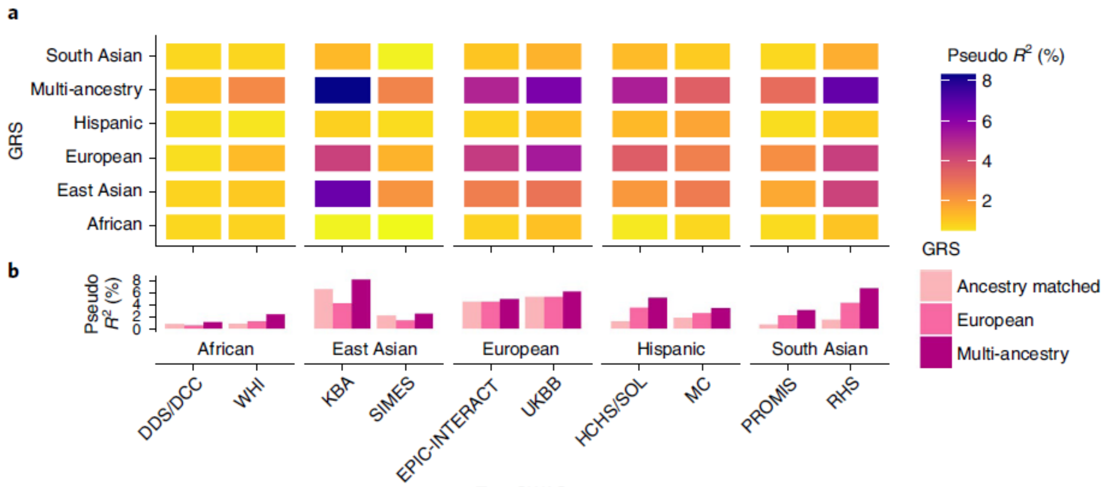
- Low predictive power → limited clinical use
- Focus on common variants only
- Low transferability when deviation from original GWAS cohort (e.g. ancestry)

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

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.