Polygenic risk scores

A/Prof Ben Brumpton

Overview

- Definition - Type of heritability

- Deriving a Polygenic Risk Score (PRS)

- Examples of risk prediction in genetic studies

Polygenic risk score

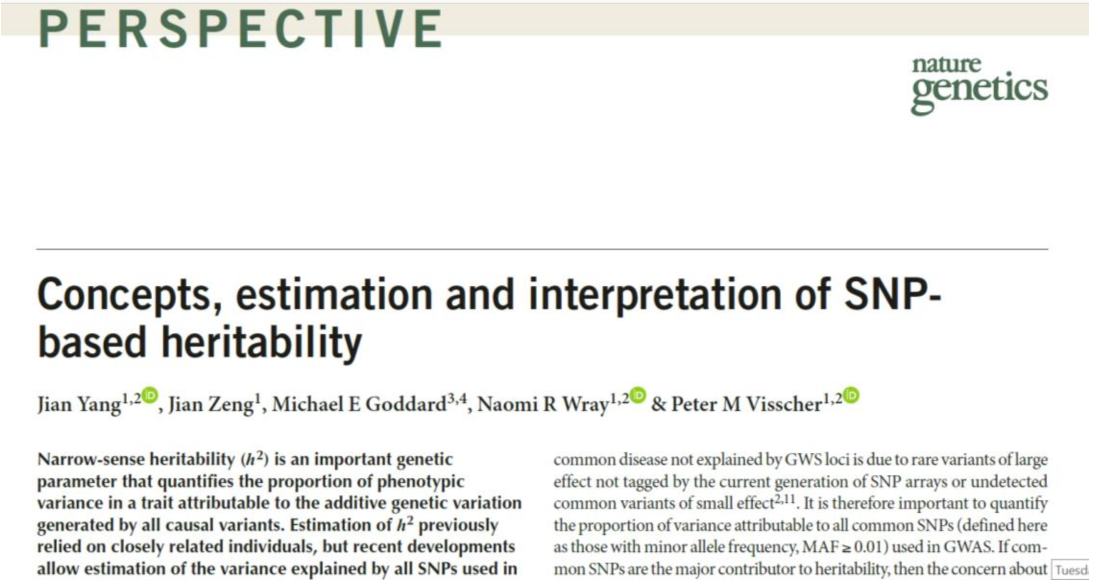
- Estimating GWAS heritability
    - Polygenic risk score (PRS)

The proportion of phenotypic variability that can be accounted for by genetic differences.

Definitions – SNP-based and GWAS heritability

- SNP-based heritability (h^2_{SNP}) is the variance explained by any set of SNPs, typically in ‘unrelated individuals’.
  - All SNPs on a genotyping array
  - Whole exome sequencing
  - Whole genome sequencing

- GWAS heritability (h^2_{GWAS}) is the variance explained by genome-wide significant loci in genome-wide association studies.



PERSPECTIVE

nature genetics

Concepts, estimation and interpretation of SNP-based heritability

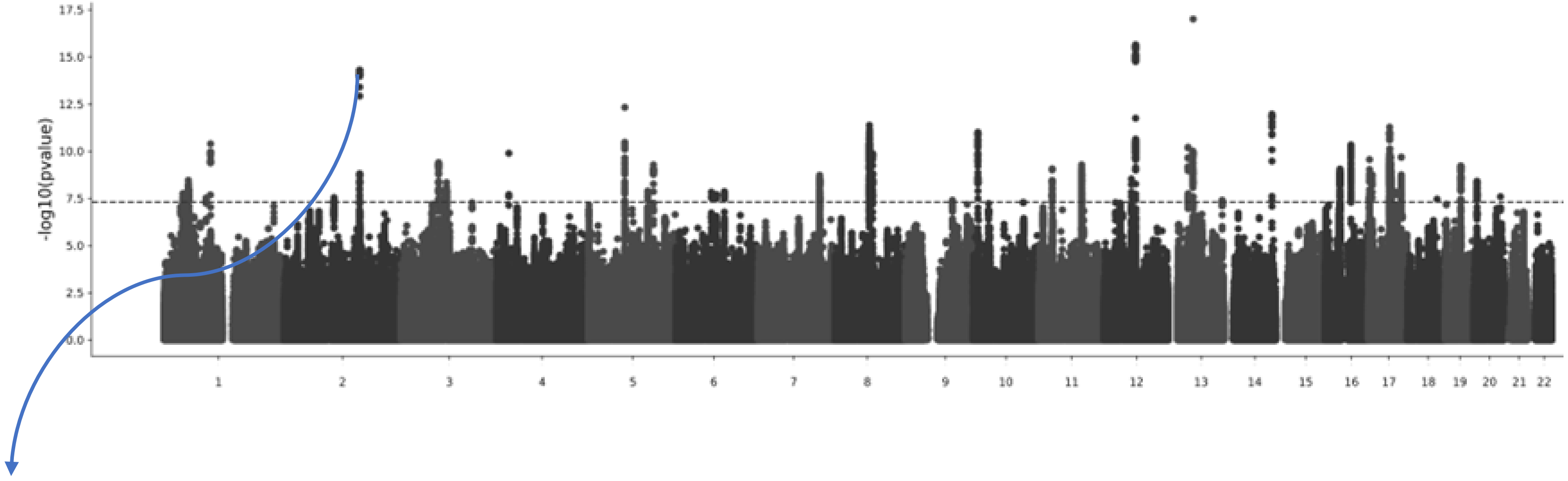
Jian Yang[1,2], Jian Zeng[1], Michael E Goddard[3,4], Naomi R Wray[1,2] & Peter M Visscher[1,2]

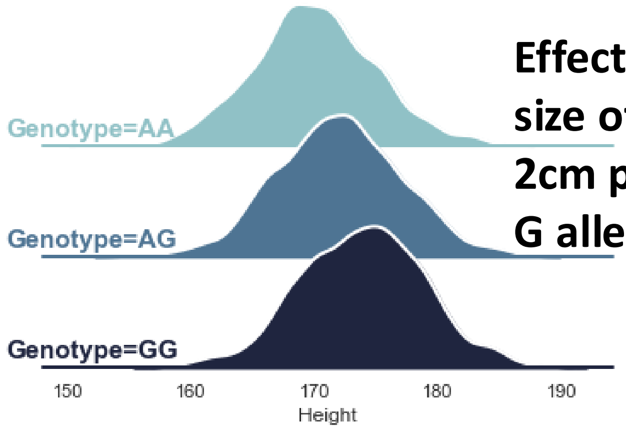
Narrow-sense heritability (h^2) is an important genetic parameter that quantifies the proportion of phenotypic variance in a trait attributable to the additive genetic variation generated by all causal variants. Estimation of h^2 previously relied on closely related individuals, but recent developments allow estimation of the variance explained by all SNPs used in

common disease not explained by GWS loci is due to rare variants of large effect not tagged by the current generation of SNP arrays or undetected common variants of small effect[2,11]. It is therefore important to quantify the proportion of variance attributable to all common SNPs (defined here as those with minor allele frequency, MAF ≥ 0.01) used in GWAS. If common SNPs are the major contributor to heritability, then the concern about

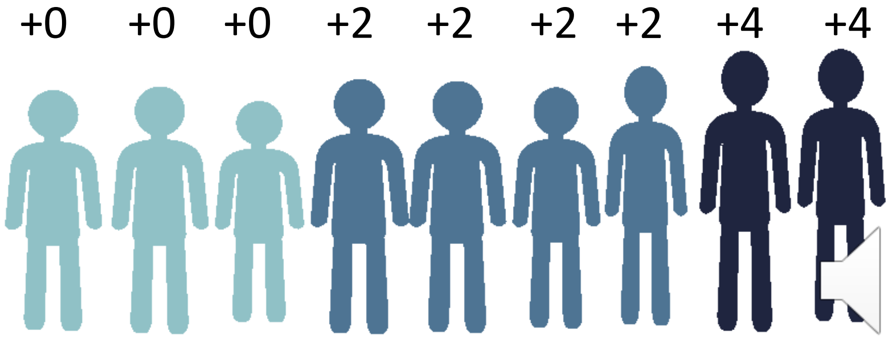
Yang et al. Nat Genet. (2017) 49:1304-1310

Intuition for polygenic risk scores

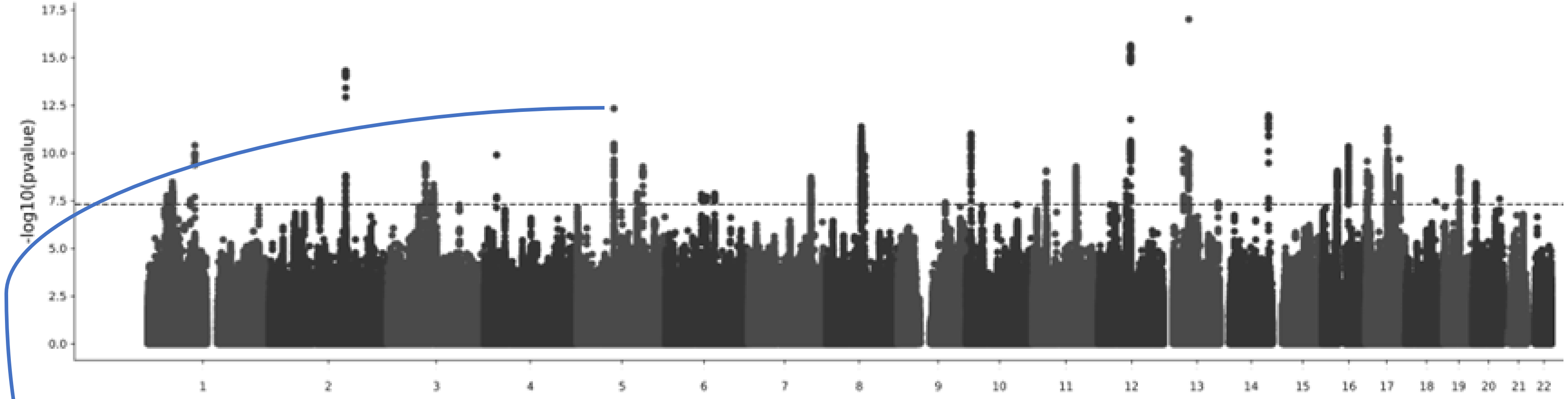


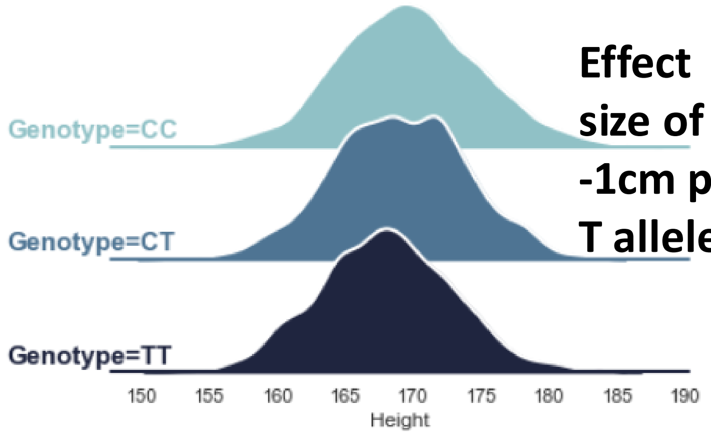


Effect size of 2cm per G allele



Intuition for polygenic risk scores

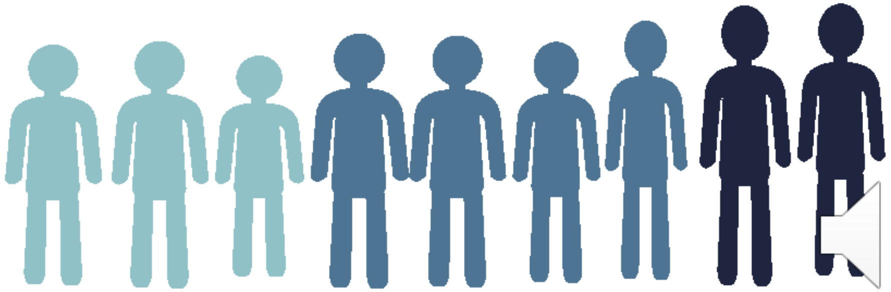




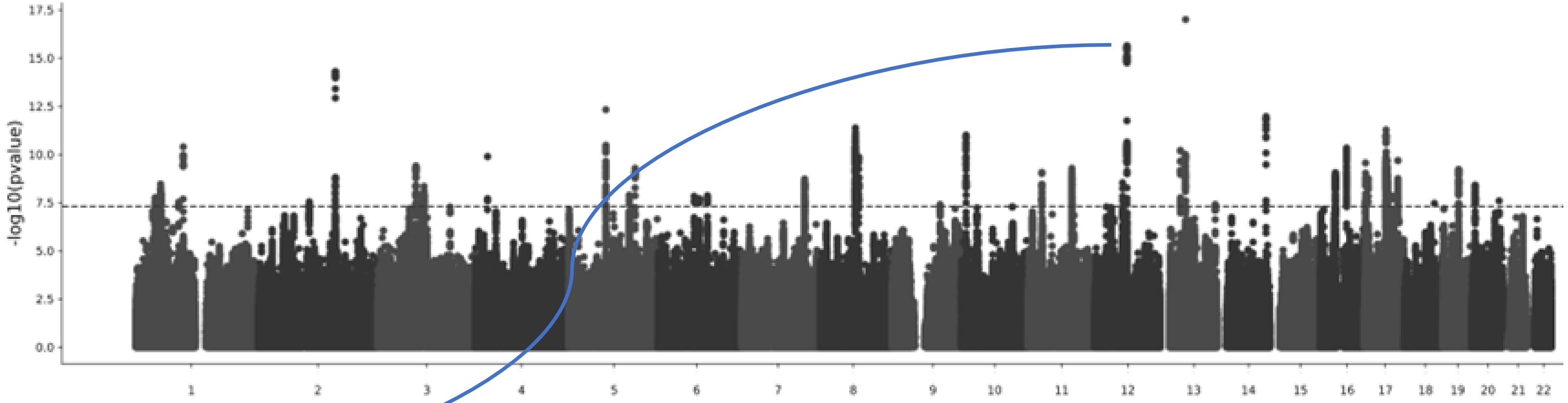
Effect size of -1cm per T allele

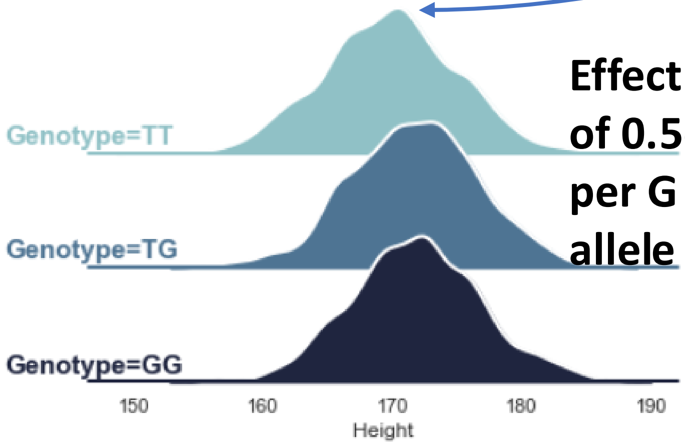
-2 +0 -2 -1 +0 +0 -1 -1 +0

+0 +0 +0 +2 +2 +2 +2 +4 +4



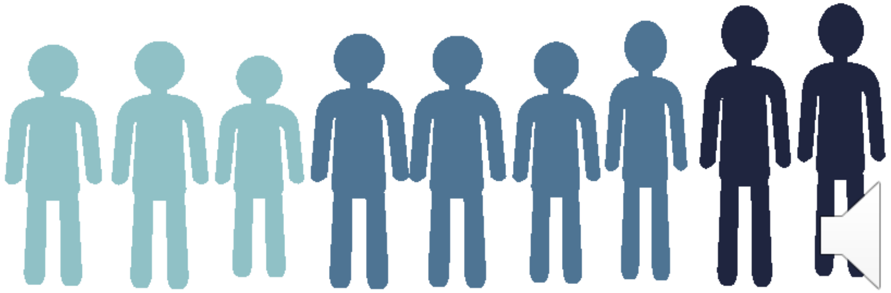
Intuition for polygenic risk scores



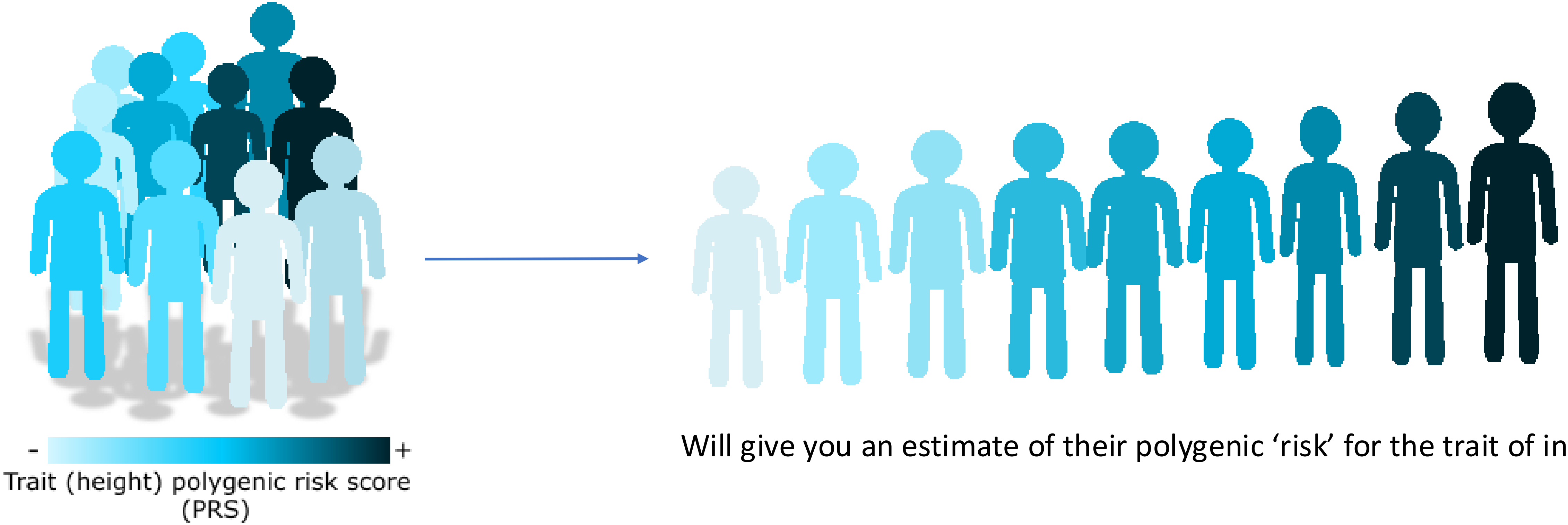


Effect size of 0.5cm per G allele

+0.5	+0.5	1	+0	+0	+0.5	+0	+1	+1
-2	+0	-2	-1	+0	+0	-1	-1	+0
+0	+0	+0	+2	+2	+2	+2	+4	+4



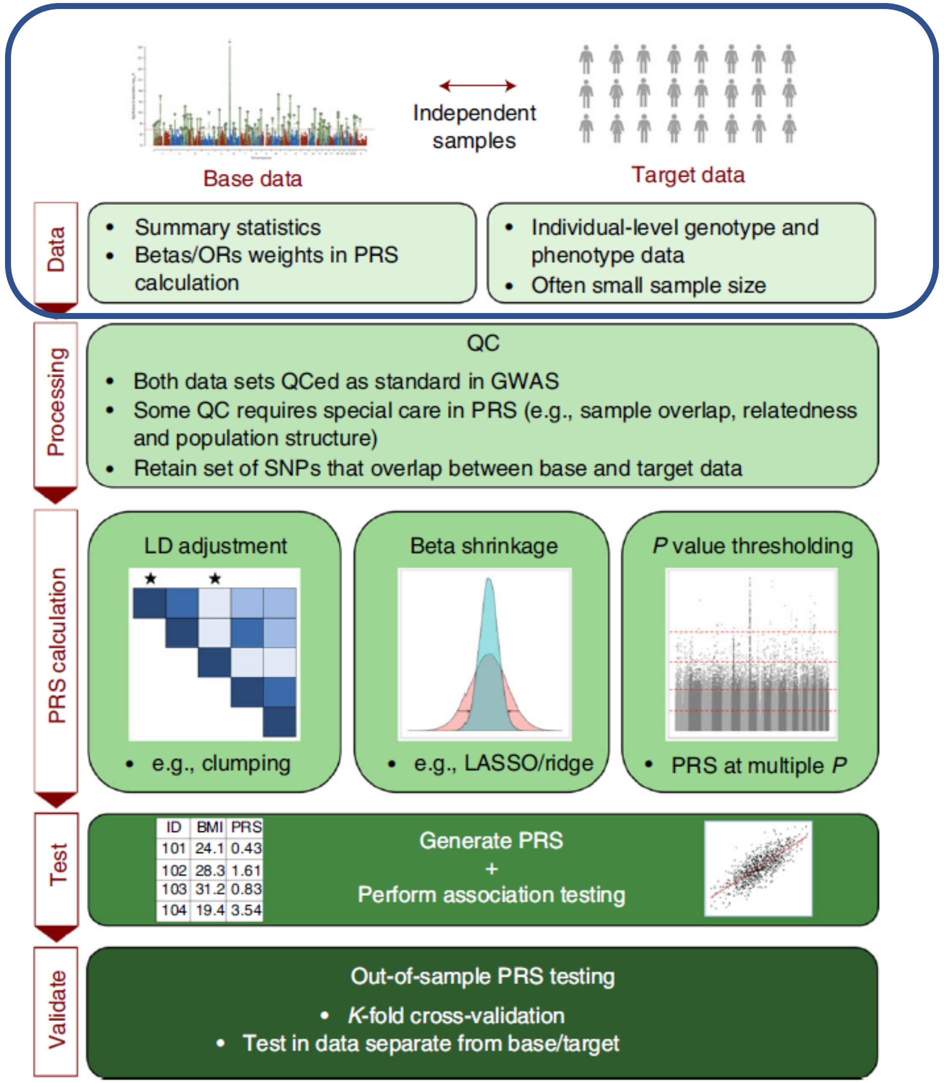
Intuition for polygenic risk scores



Will give you an estimate of their polygenic ‘risk’ for the trait of interest

Polygenic risk score – estimated effect of many genetic variants on an individual's phenotype, calculated as a weighted sum of trait-associated alleles.

How to calculate a polygenic risk score



- 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

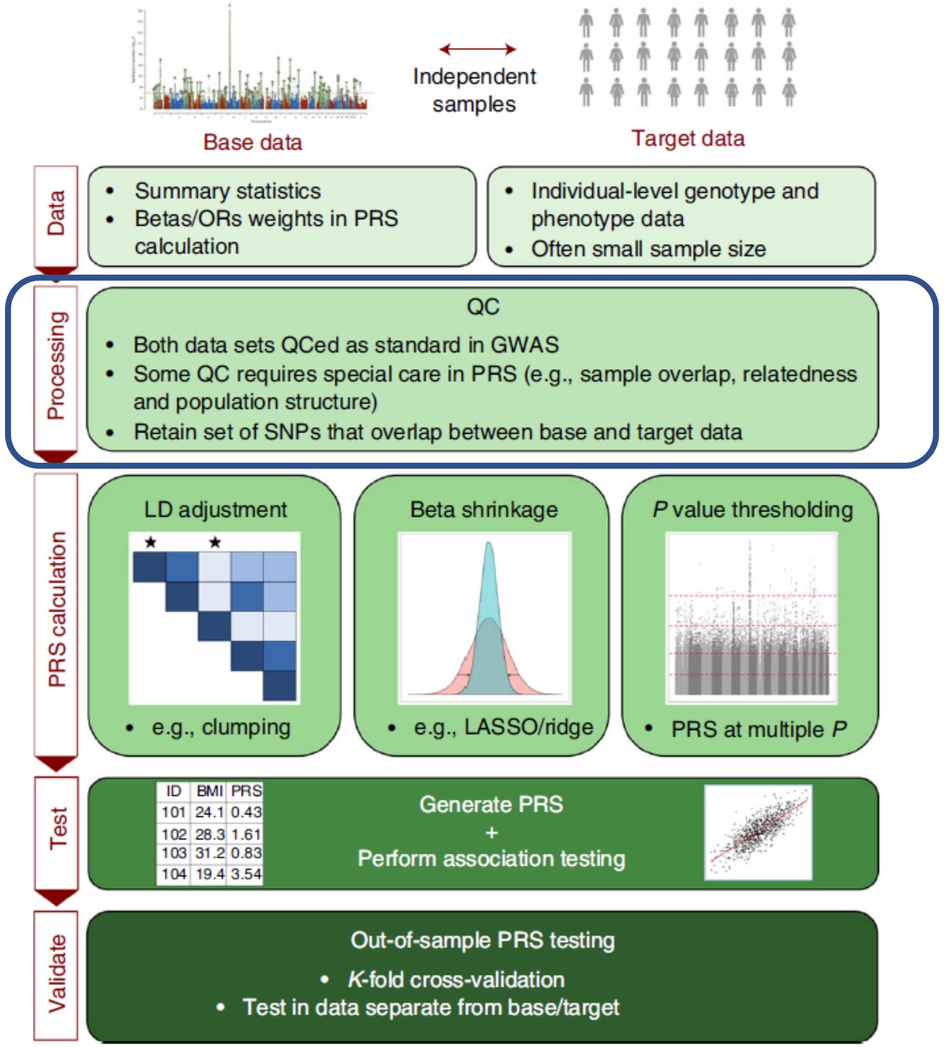
What you need:

1. Base data: consisting of summary statistics (e.g., effect size estimates and P-values) of genotype-phenotype associations at genetic variants genome-wide

2. Target data: individual level genotype and phenotype data, which is independent of the GWAS sample

Choi et al. Nat Protocols. (2020) 15:2759-2772

How to calculate a polygenic risk score



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

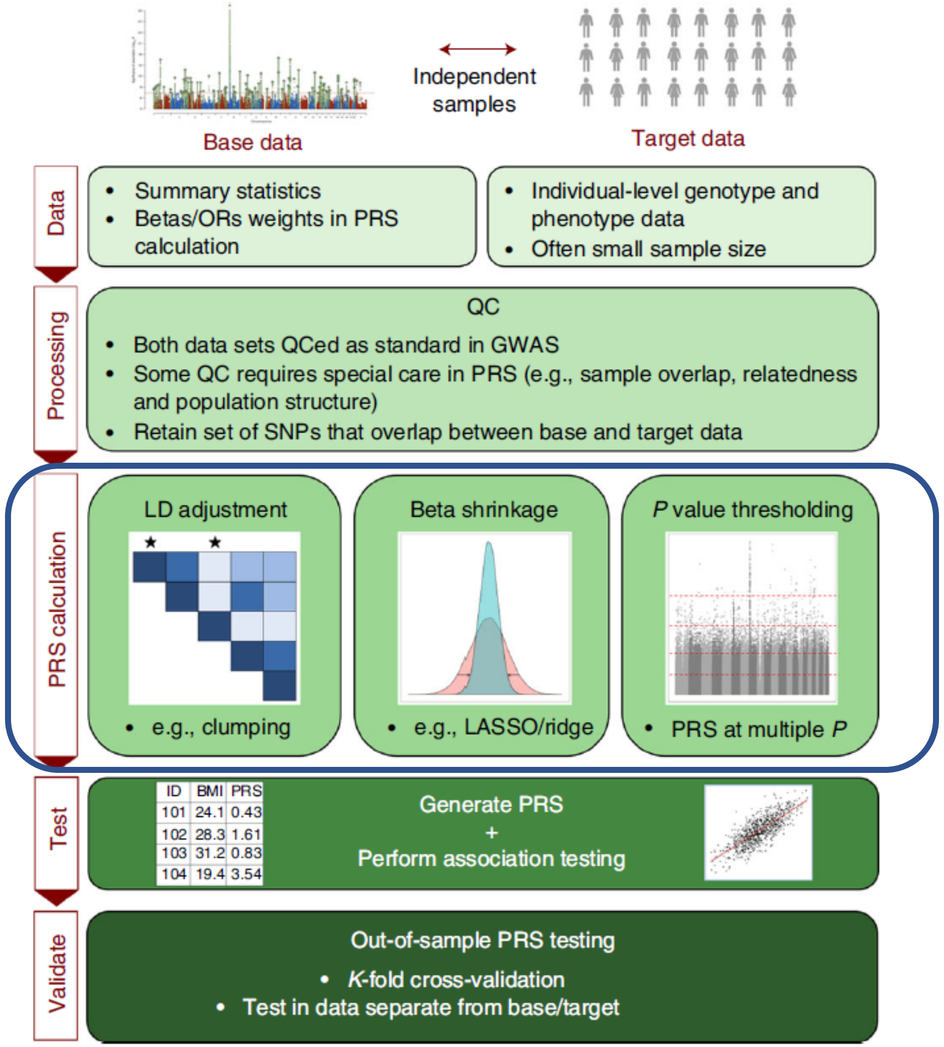
- *K*-fold cross-validation
- Test in data separate from base/target

Quality control:

- GWAS should have $h^2_{SNP} > 0.05$
- Target data should have N>100
- Ensure both base and target data have genomic positions assigned on the same genome build
- Apply standard GWAS QC on both base and target data (i.e. genotyping rate, missingness, HWE, heterozygosity, MAF, imputation quality)
- Remove ambiguous (C/G or A/T SNPs) and duplicate SNPs
- Consider removing mismatches of sex and the sex chromoromes
- Sample overlap -> inflation in the association between PRS and the trait of interest (proportional to the fraction of overlap).

Choi et al. Nat Protocols. (2020) 15:2759-2772

How to calculate a polygenic risk score



- 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

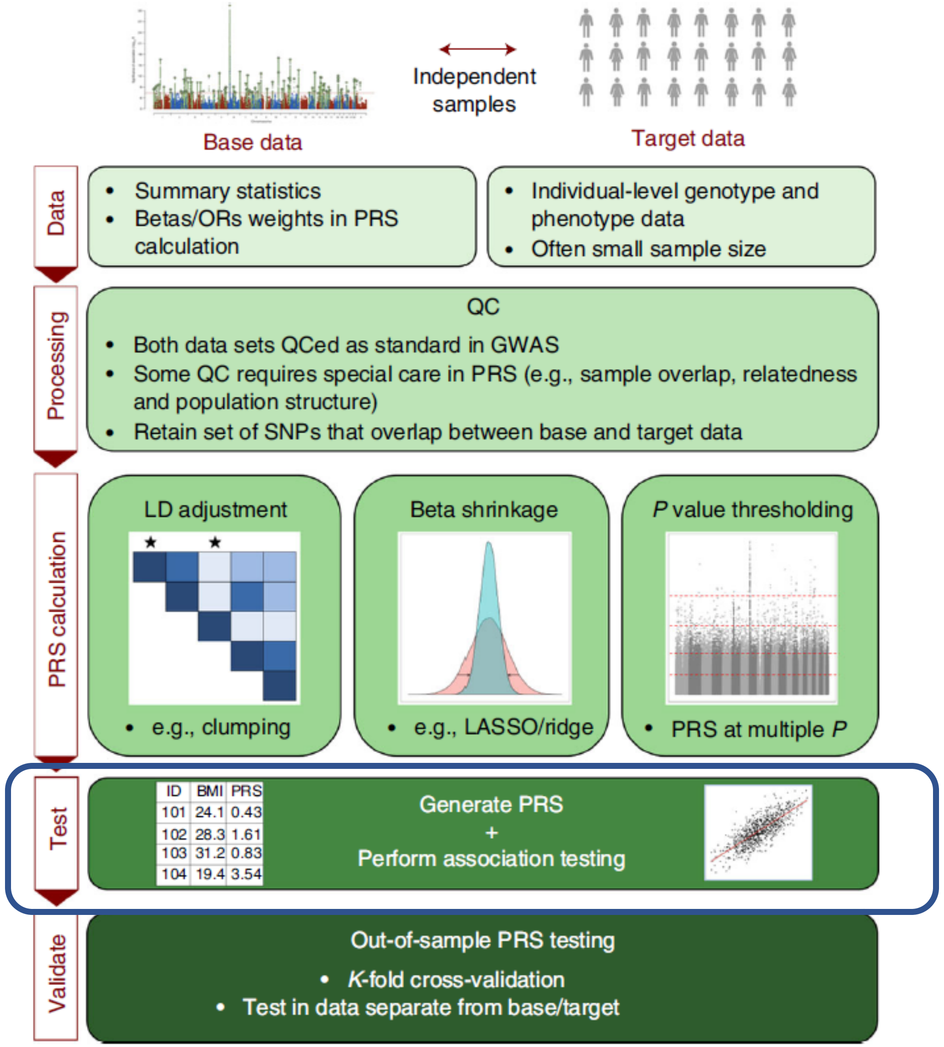
- *K*-fold cross-validation
- Test in data separate from base/target

Methods for calculating PRS:

1. Adjustment for GWAS estimated effect sizes (beta shrinkage)
   - Shrinkage of all SNPs use statistical shrinkage techniques (LASSO, ridge regression, Bayesian approaches). Dependent on underlying mixture of null and true effect sizes
   - Thresholding = using a predefined P-value threshold and setting all other effect sizes to zero.
2. Tailoring the PRSs to the target population
3. Accounting for LD
   - If independent effects were estimated in the GWAS or by subsequent fine-mapping, then PRS calculation can be a simple summation of those effects
   - Clumping = thinning data based on LD and prioritizing SNPs at each locus with the smallest P-value
   - Include all SNPs and account for LD between them

Choi et al. Nat Protocols. (2020) 15:2759-2772

How to calculate a polygenic risk score



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

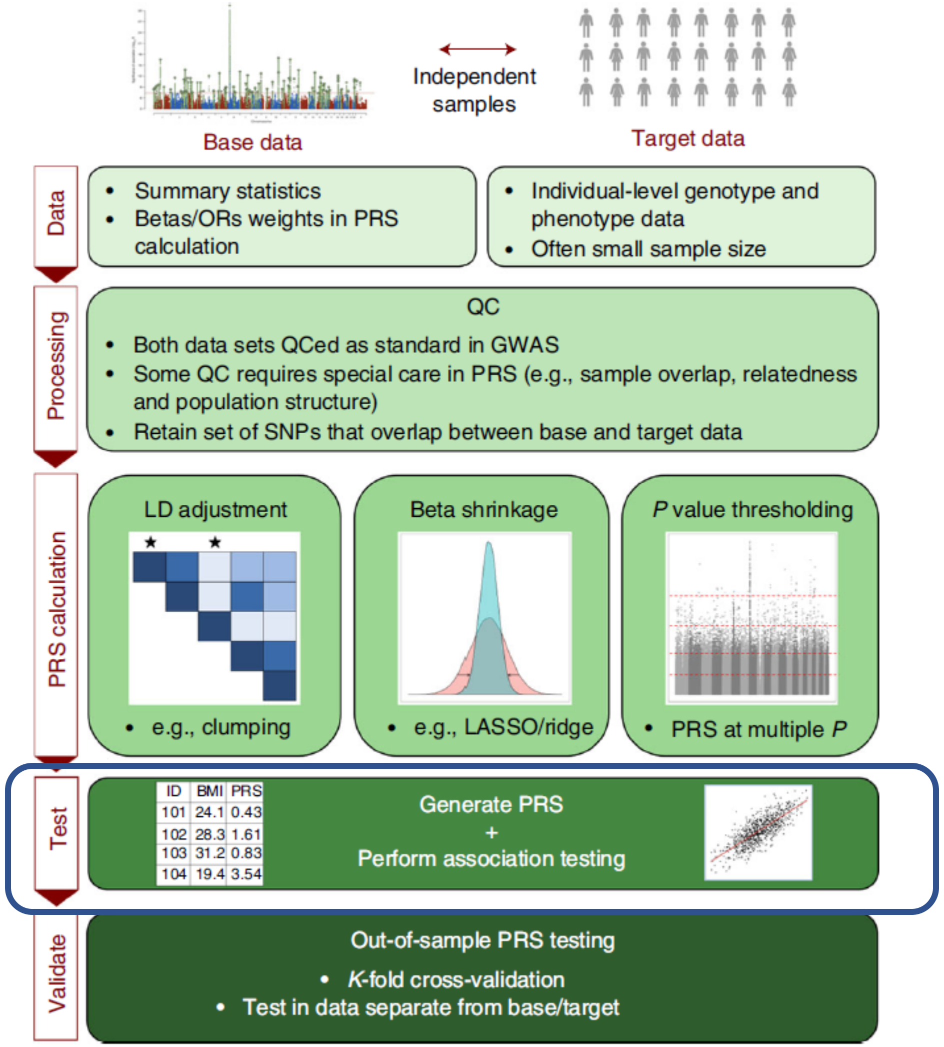
- *K*-fold cross-validation
- Test in data separate from base/target

Calculating PRS:

- PRSice: standard C+T approach
- LDpred: Bayesian shrinkage model
- PRS-CS: Bayesian regression with continuous shrinkage prior
- JAMPred: Two-step Bayesian modelling
- lassosum: Penalized regression
- Software not dedicated to PRS (e.g. Plink, bigsnpr R package)

Choi et al. Nat Protocols. (2020) 15:2759-2772

How to calculate a polygenic risk score



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

Performing association testing:

- Test for association between PRS and trait(s) in target data, adjusting for covariates (genetic PCs, age, sex etc).
  - P-value for testing a null hypothesis of no association
  - Phenotypic variance explained (adjusted R^2)
  - Effect size estimate per unit of PRS or between specific start (high vrs low risk individuals)
  - Area under the receiver operator curve (AUC)

Choi et al. Nat Protocols. (2020) 15:2759-2772

Visualizing the results of PRS analysis

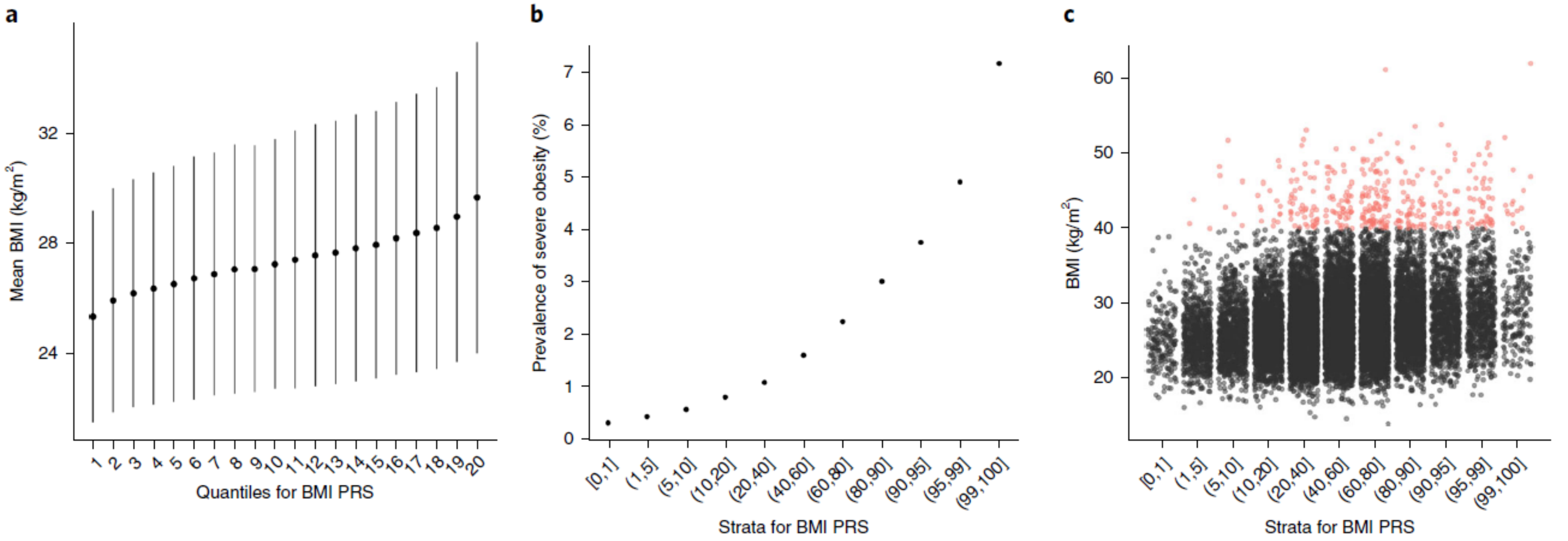


Fig. 5 | Three different ways of representing the same data. The data correspond to body mass index (BMI; in kg/m^2) PRSs calculated in 386,266 individuals in the UK Biobank data, derived using the GIANT BMI GWAS as base data. **a**, Quantile plot with 20 quantiles of increasing BMI PRS versus mean BMI (y-axis). **b**, Strata plot with unequal strata of increasing BMI PRS versus prevalence (%) of severe obesity (BMI > 40). **c**, Strata plot with the same strata as in **b**, but here each individual's BMI value is shown on the y-axis. The sample is randomly thinned to 5% of the total size, and lateral spread within each stratum is applied, to make individual points visible, while red points correspond to individuals with severe obesity. Qualitatively similar patterns as these should be expected for PRSs corresponding to all reasonably heritable continuous or binary traits, with strength of patterns dependent on the predictive power of the PRS (here, the PRS explains ~5% of BMI in these data). BMI here could be considered analogous to the liability underlying a disease in the liability threshold model, and in this way plot **c** may be helpful in imagining the uncertainty in the true liability that underlies a given PRS value for a disease.

Choi et al. Nat Protocols. (2020) 15:2759-2772

Applications for PRS

- Test for GWAS association and quantify variance explained (GWAS heritability)
- Risk stratification (i.e. identifying people to later test for specific disease)
- Aid in clinical diagnosis
- Test for genetic overlap between traits (e.g. does a depression PRS predict cardiovascular disease?)
- Trait imputation when not measured (obviously imperfect and dependent on heritability)
- Personalized treatment (GWAS on treatment response are gaining power)

Overview

- Definition - Types of heritability
- Estimating GWAS heritability
  - Polygenic risk score (PRS)
- Examples of heritability and risk prediction in genetic studies

Heritability estimates for height

- Heritability from family and twin studies: 80%
- SNP-heritability (h^2_{SNP}): 48.3% (SE: 3.7%)
- GWAS-heritability (h^2_{GWAS}) :

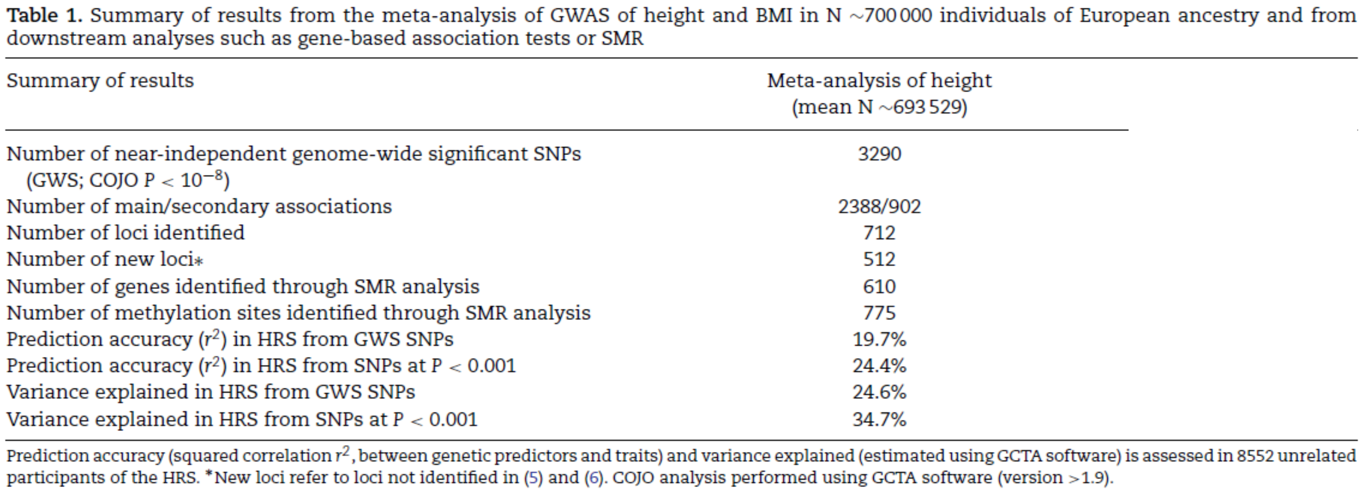
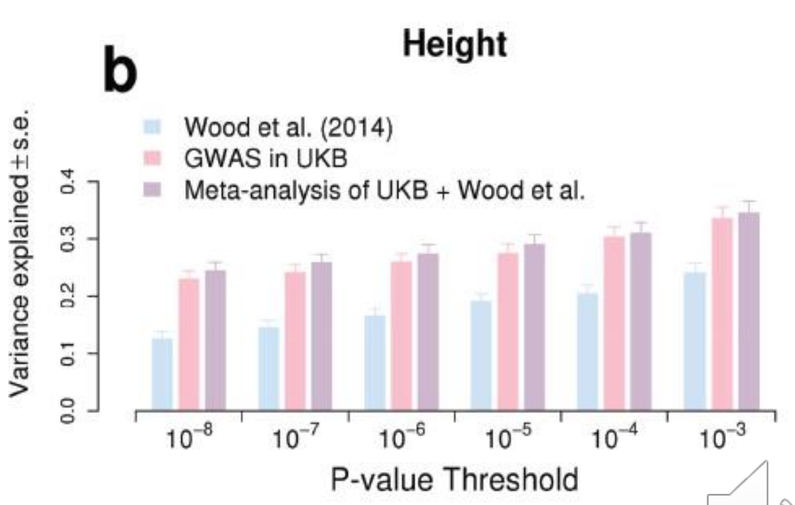


Table 1. Summary of results from the meta-analysis of GWAS of height and BMI in N ~700 000 individuals of European ancestry and from downstream analyses such as gene-based association tests or SMR

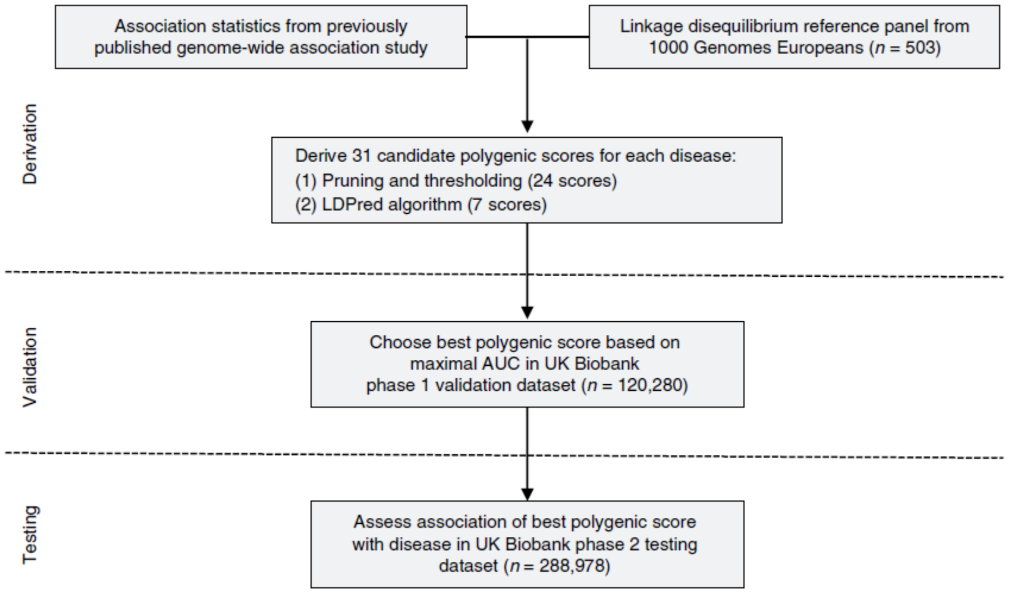
Summary of results	Meta-analysis of height (mean N ~693 529)
Number of near-independent genome-wide significant SNPs (GWS; COJO $P < 10^{-8}$)	3290
Number of main/secondary associations	2388/902
Number of loci identified	712
Number of new loci*	512
Number of genes identified through SMR analysis	610
Number of methylation sites identified through SMR analysis	775
Prediction accuracy (r^2) in HRS from GWS SNPs	19.7%
Prediction accuracy (r^2) in HRS from SNPs at $P < 0.001$	24.4%
Variance explained in HRS from GWS SNPs	24.6%
Variance explained in HRS from SNPs at $P < 0.001$	34.7%

Prediction accuracy (squared correlation r^2, between genetic predictors and traits) and variance explained (estimated using GCTA software) is assessed in 8552 unrelated participants of the HRS. *New loci refer to loci not identified in (5) and (6). COJO analysis performed using GCTA software (version >1.9).



Yengo et al. Hum Mole Genet. (2018) 27(20):3641-3649

Risk prediction using polygenic risk scores



Association statistics from previously published genome-wide association study
Linkage disequilibrium reference panel from 1000 Genomes Europeans (n = 503)
Derivation
Derive 31 candidate polygenic scores for each disease:
(1) Pruning and thresholding (24 scores)
(2) LDPred algorithm (7 scores)
Validation
Choose best polygenic score based on maximal AUC in UK Biobank phase 1 validation dataset (n = 120,280)
Testing
Assess association of best polygenic score with disease in UK Biobank phase 2 testing dataset (n = 288,978)

Fig. 1 | Study design and workflow. A GPS for each disease was derived by combining summary association statistics from a recent large GWAS and a linkage disequilibrium reference panel of 503 Europeans[34]. Then, 31 candidate GPSs were derived using two strategies: (1) 'pruning and thresholding' (that is, the aggregation of independent polymorphisms that exceeded a specified level of significance in the discovery GWAS); and (2) the LDPred computational algorithm[13], a Bayesian approach to calculate a posterior mean effect for all variants based on a prior (effect size in the previous GWAS) and subsequent shrinkage based on linkage disequilibrium. The seven candidate LDPred scores vary with respect to the tuning parameter ρ (that is, the proportion of variants assumed to be causal), as previously recommended[13]. The optimal GPS for each disease was chosen based on the AUC in the UK Biobank phase 1 validation dataset (n=120,280 Europeans) and subsequently calculated in an independent UK Biobank phase 2 testing dataset (n=288,978 Europeans).

Khera et al. Nat Genet. (2018) 50:1219-1224

Risk prediction using polygenic risk scores

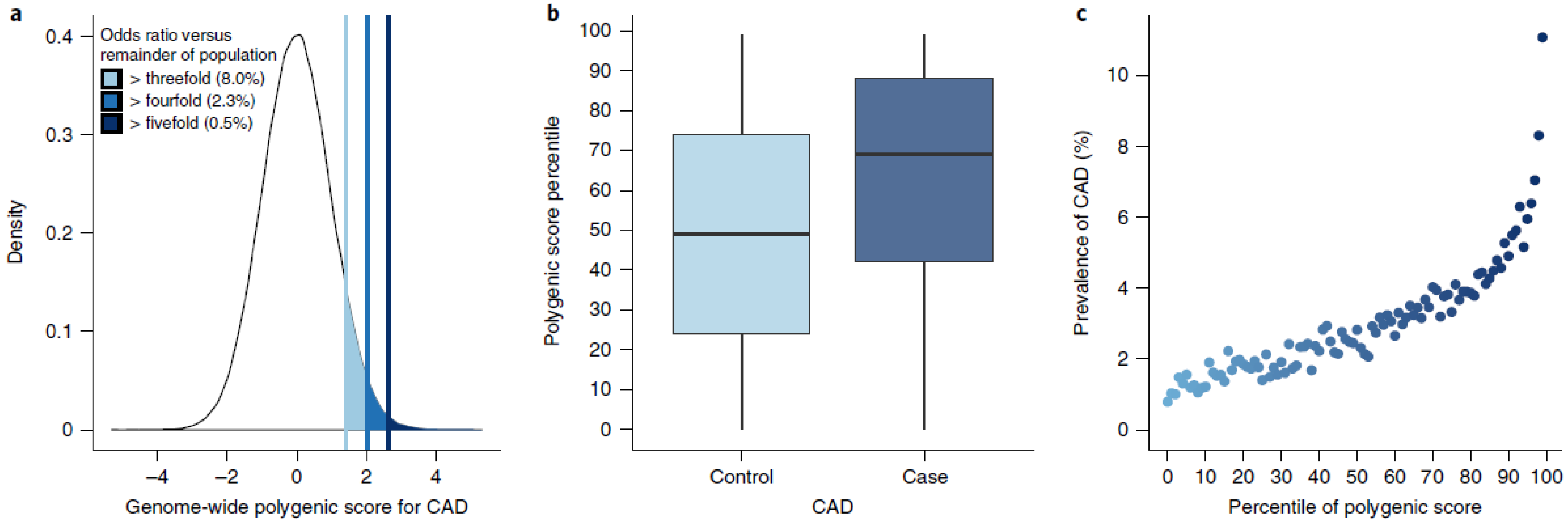


Fig. 2 | Risk for CAD according to GPS. a, Distribution of GPS_{CAD} in the UK Biobank testing dataset (*n* = 288,978). The *x* axis represents GPS_{CAD}, with values scaled to a mean of 0 and a standard deviation of 1 to facilitate interpretation. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. The odds ratio was assessed in a logistic regression model adjusted for age, sex, genotyping array, and the first four principal components of ancestry. **b**, GPS_{CAD} percentile among CAD cases versus controls in the UK Biobank testing dataset. Within each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the interquartile range, and the whiskers reflect the maximum and minimum values within each grouping. **c**, Prevalence of CAD according to 100 groups of the testing dataset binned according to the percentile of the GPS_{CAD}.

Khera et al. Nat Genet. (2018) 50:1219-1224

Summary

- Defined two heritability terms; SNP heritability and GWAS heritability.
- Discussed how to calculate a polygenic risk score
- Discussed the various applications.