Conventional Methods for PGS Prediction

Jian Zeng

j.zeng@uq.edu.au

Polygenic scores

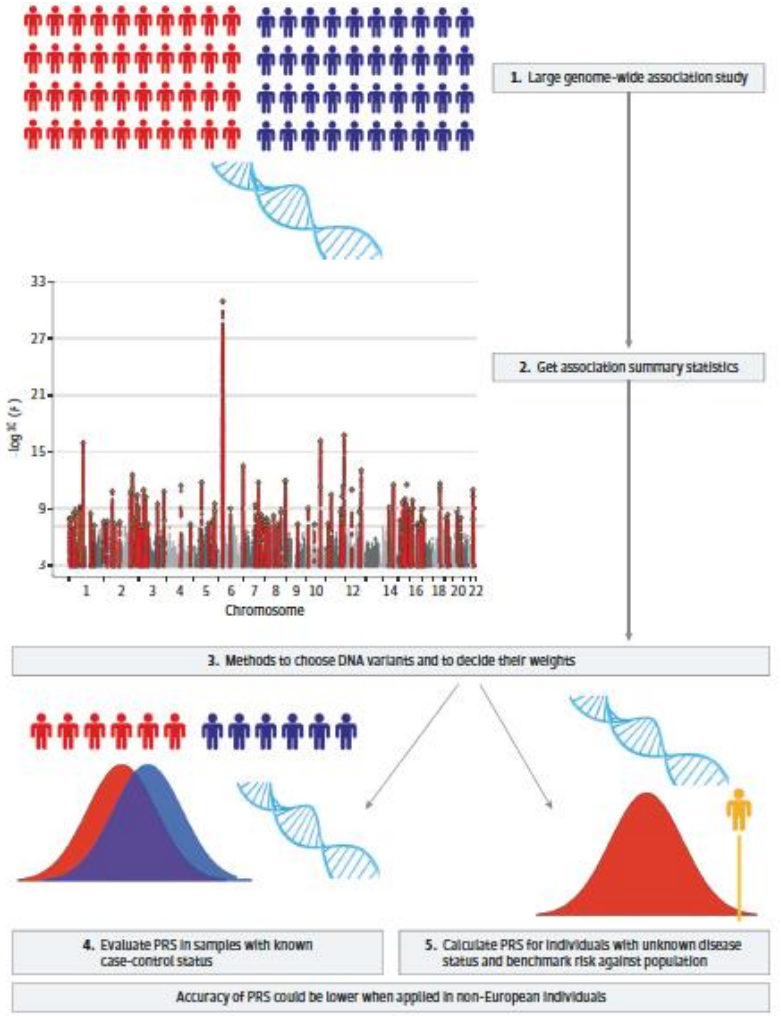
Polygenic score (PGS) is a weighted count of risk alleles

$$PGS = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

0, 1 or 2
Risk alleles

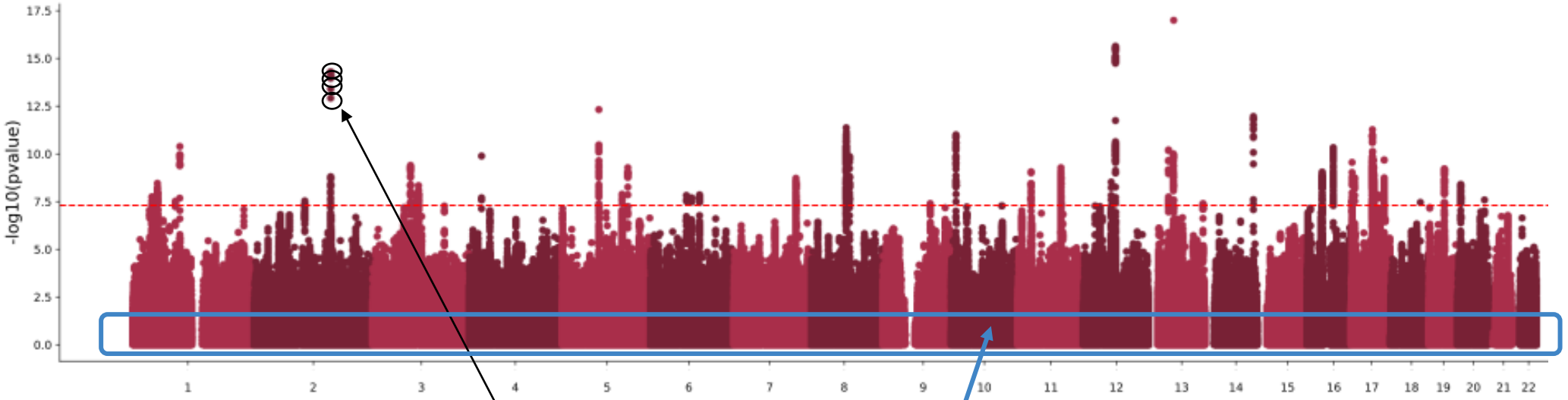
Which SNPs?

What weights?



- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.

SNP Weights



GWAS results give us $\hat{\beta}_j^{GWAS}$, not β_j. Two issues to consider when constructing $\sum_{j=1}^{n_{SNP}} \hat{\beta}_j^{GWAS} x_{ij}$:

1. For some SNPs, $\hat{\beta}_j^{GWAS}$ may be a very noisy estimate of β_j and/or β_j may be close to 0, so adding those SNPs will add more noise than signal

2. If we include all SNPs, we will overweight ("double-count") SNPs with high LD scores

From Aysu Okbay

Two solutions

$$\sum_{j=1}^{n_{SNP}} \hat{\beta}_j x_{ij}$$

Clumping and P-value thresholding (C+PT)

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Weights: Set equal to GWAS coefficients.

Loci: Selected by

1. using a **clumping** algorithm that ensures the included SNPs are all approximately independent of each other
2. omitting SNPs whose *P* value for association with the phenotype is above a certain **threshold**

Whole-genome regression approaches

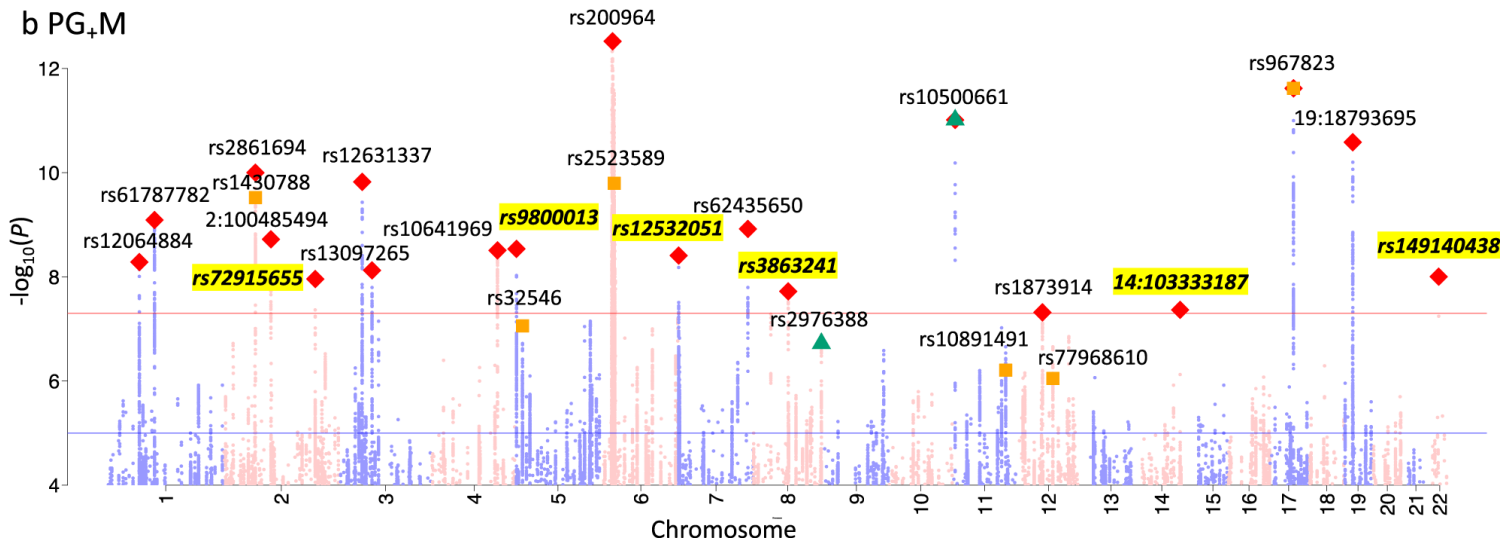
Include all SNPs but adjust the effect sizes for LD

Weights: Set to GWAS coefficients **adjusted for LD** → from a random-effect model regressing the phenotype on all SNPs

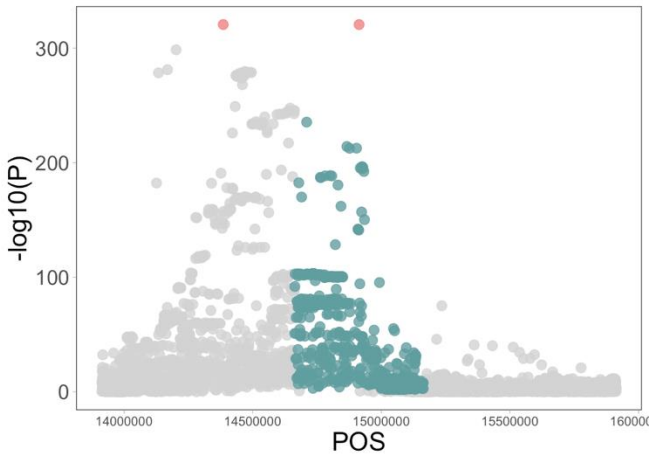
Loci: Include **all SNPs**, no LD-based pruning

Examples: BLUP (Meuwissen et al. 2001), LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019)

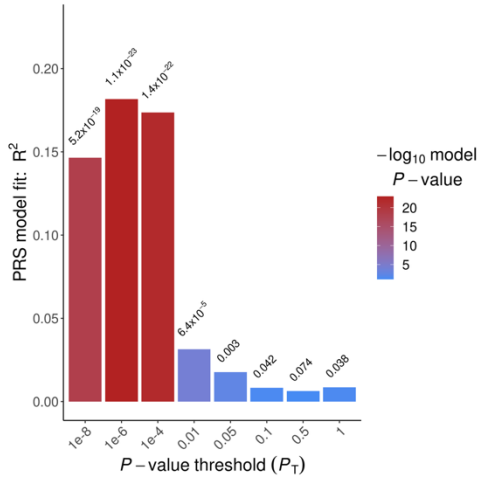
Clumping & P-value thresholding (C+PT, or P+T, C+T)



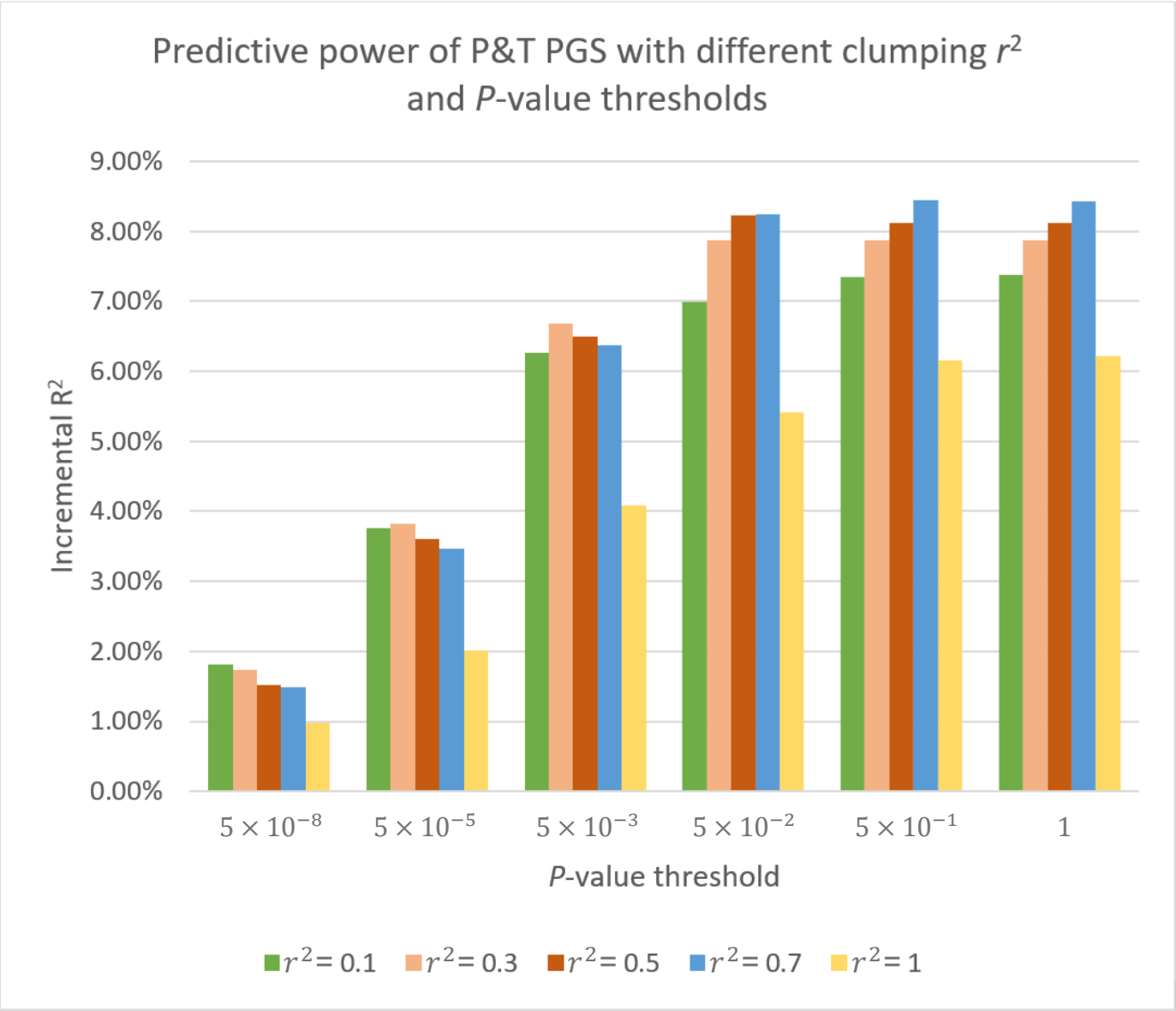
Step 1. Select most associated SNP in tower (LD-based clumping)



Step 2. Select on a p-value threshold in an independent tuning sample



Clumping & P-value thresholding (C+PT, or P+T, C+T)



- **Cohort:** Health and Retirement Study
- **Phenotype:** Educational attainment

From Aysu Okbay

Whole-genome regression approaches

Best Linear Unbiased Prediction (BLUP)

or Ridge Regression

Bayesian methods

- Fit all SNP effects as random
- Borrow information across SNPs
- Shrinkage estimation of SNP effects

Best Linear Unbiased Prediction (BLUP)

Linear mixed model

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

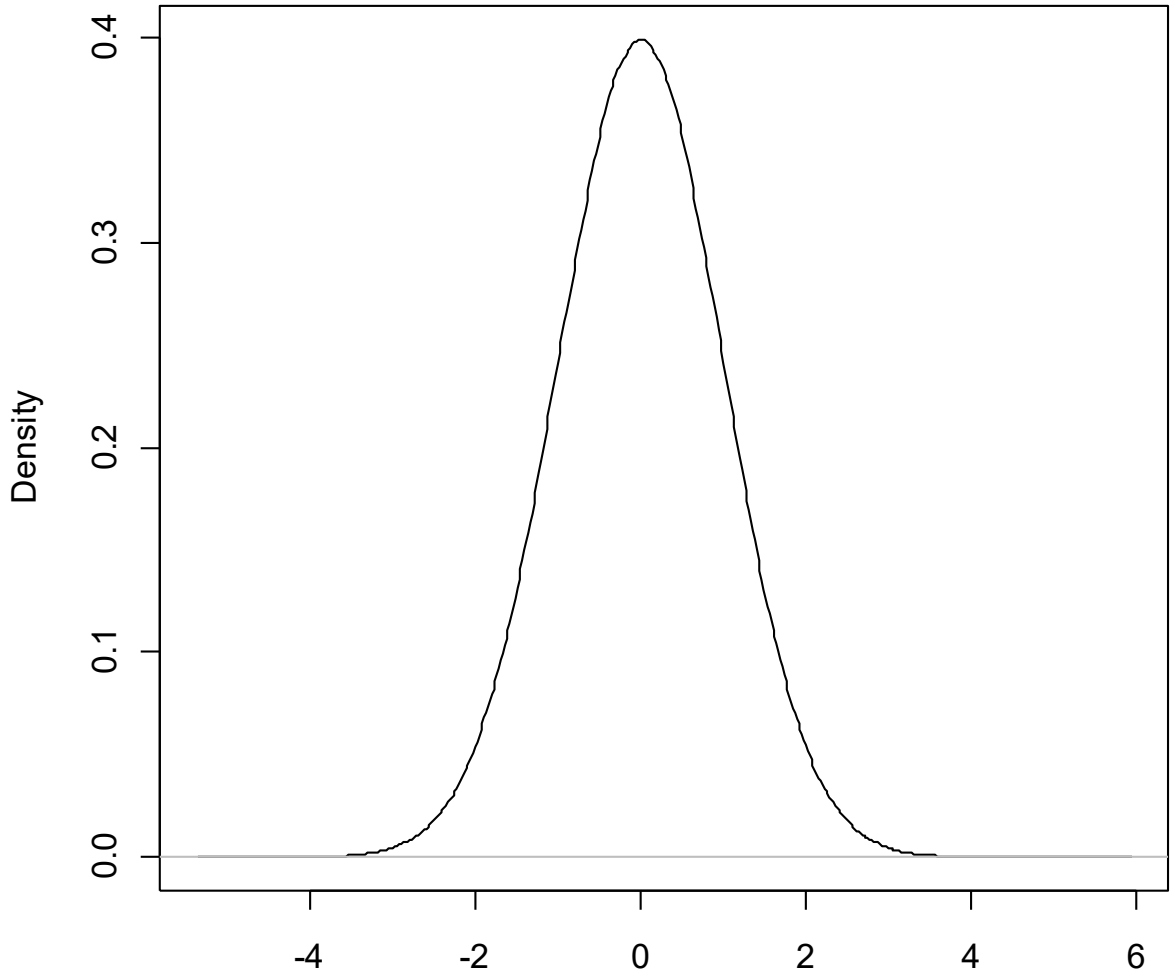
- $\mathbf{y}$ is a vector of *n* phenotypes,
- μ is the mean,
- $\mathbf{X}$ is an incidence matrix of individuals' genotypes for all SNPs,
- $\boldsymbol{\beta}$ are the random effects of the *m* SNPs,
- $\mathbf{e}$ is a vector of random residuals, $\mathbf{e} \sim N(0, \sigma_e^2)$

Assume SNP effects come from normal distribution with same variance

$\boldsymbol{\beta} \sim N(0, \sigma_\beta^2)$

Assumed distribution of SNP effects

$N(0, \sigma_\beta^2)$



Best linear unbiased prediction (BLUP)

To estimate random effects (Henderson 1975 & Robinson 1991).

Best: minimum mean square error within class of linear predictors

Linear: random variables $\boldsymbol{\beta}$ are linear functions of the data **y**

Unbiased: the average value of the estimate of $\boldsymbol{\beta}$ is equal to the average value of the quantity being estimated

Predictor: to distinguish random effects from fixed effect estimates

Best linear unbiased prediction (BLUP)

Linear mixed model

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

BLUP solutions

$$\begin{bmatrix} \hat{\mu} \\ \widehat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

I = identity matrix (dimensions *m* x *m*)

$\lambda = \sigma_e^2 \,/\, \sigma_\beta^2$

BLUP solutions

$$\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

Least Squares (LS) solutions

$$\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

Shrinkage

BLUP solutions

$$\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n'X} \\ \mathbf{X'1_n} & \mathbf{X'X} + \mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n'y} \\ \mathbf{X'y} \end{bmatrix}$$

$\lambda = \sigma_e^2 \,/\, \sigma_\beta^2$ is known as the shrinkage parameter

It shrinks LS estimates toward zero to an extent depending on the noise-signal ratio.

e.g., ignoring mean and other SNP $\hat{\beta}_1 = \frac{X_1'y}{X_1'X_1+\lambda} < \frac{X_1'y}{X_1'X_1}$ ← LS estimate

Shrinkage

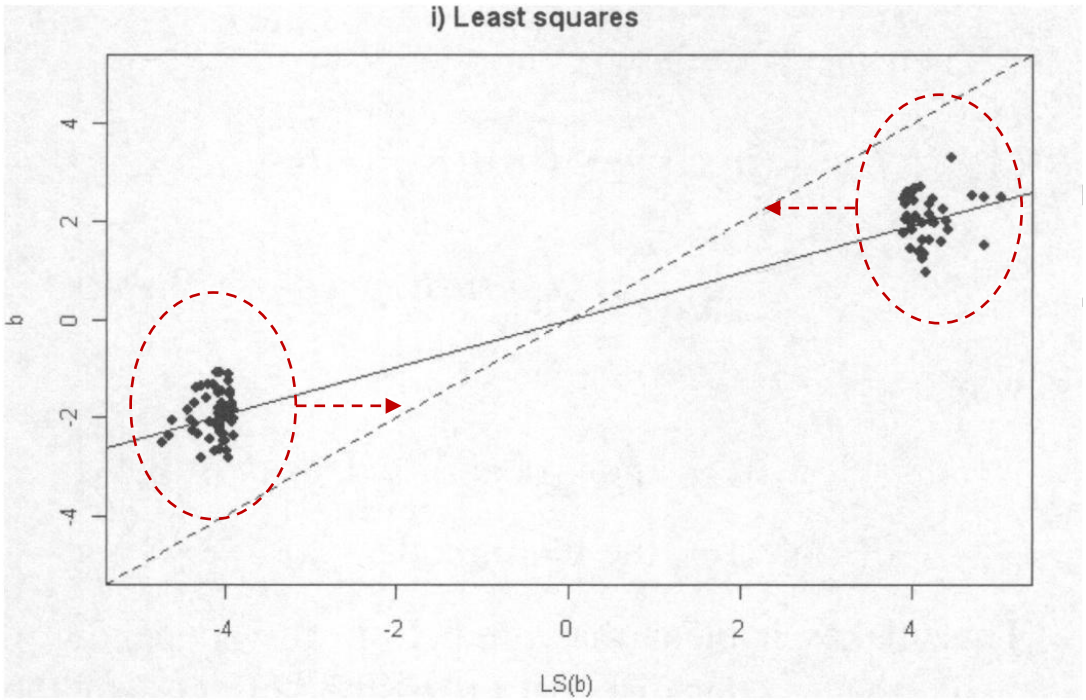
Shrinks LS estimates toward zero

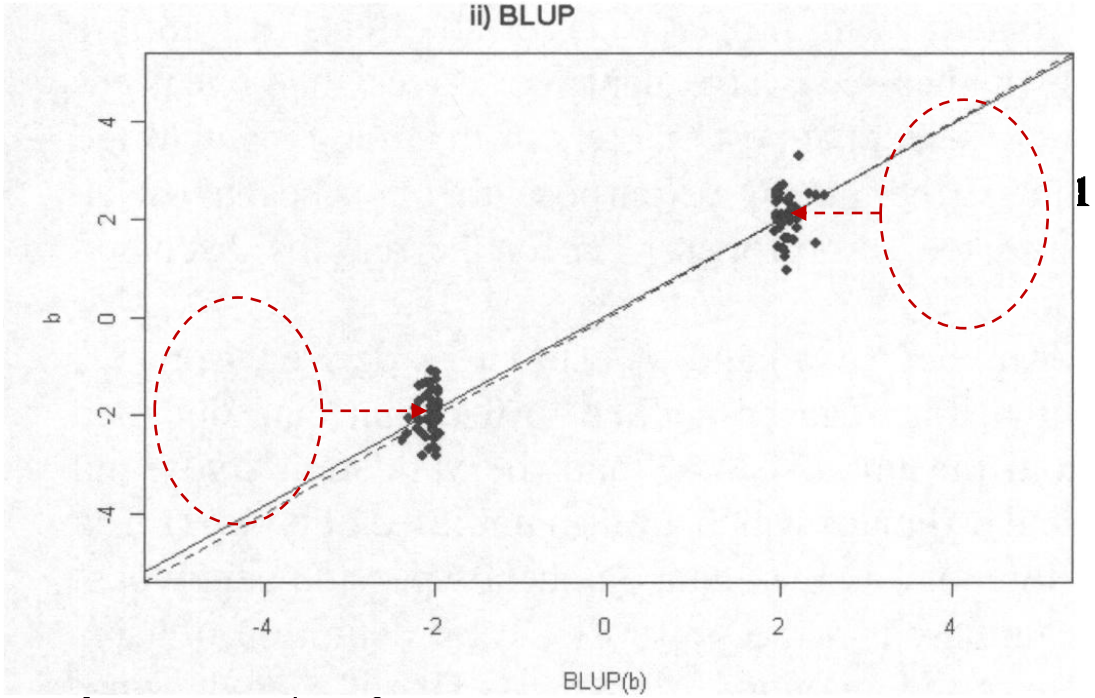
ronmental" factors. For example, height in humans involves many physiological processes and many genes but is also influenced by nongenetic factors such as nutrition and health care. These traits are called quanti-

Michael E. Goddard is Professor of Animal Genetic, Faculty of Land and Food Resources, University of Melbourne and Department of Primary Industries, Victoria, Australia. Naomi R. Wray is Professor of Psychiatric, Genetic Epidemiology and Queensland Statistical Genetics,

positions in the DNA sequence where the nucleotides can vary (e.g., G or T). Individuals carry pairs of homologous chromosomes and so have one of three genotypes at a G/T SNP—GG, GT or TT. Assays are now available that determine the genotype of an individual at 100,000 to over 1 million SNPs spread over all of the chromosomes of the species.

SNPs usually have no direct effect on a trait under study. However, any polymorphism that does affect the trait will be located on a chromosome close to





BLUP avoids selection bias!

Property of BLUP

ronmental" factors. For example, height in humans involves many physiological processes and many genes but is also influenced by nongenetic factors such as nutrition and health care. These traits are called quanti-

Michael E. Goddard is Professor of Animal Genetic, Faculty of Land and Food Resources, University of Melbourne and Department of Primary Industries, Victoria, Australia. Naomi R. Wray is Professor of Psychiatric, Genetic Epidemiology and Queensland Statistical Genetics,

positions in the DNA sequence where the nucleotides can vary (e.g., G or T). Individuals carry pairs of homologous chromosomes and so have one of three genotypes at a G/T SNP—GG, GT or TT. Assays are now available that determine the genotype of an individual at 100,000 to over 1 million SNPs spread over all of the chromosomes of the species.

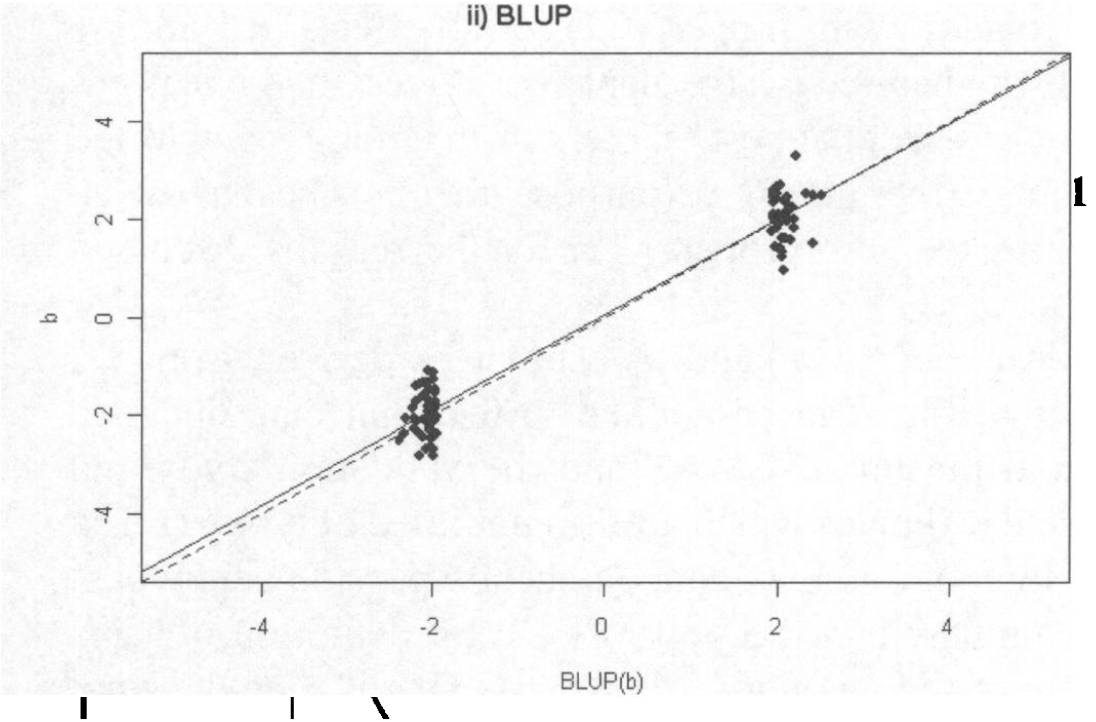
SNPs usually have no direct effect on a trait under study. However, any polymorphism that does affect the trait will be located on a chromosome close to

Unbiased: $E[\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_{\mathrm{BLUP}}] = \hat{\boldsymbol{\beta}}_{\mathrm{BLUP}}$

In contrast, for LS estimator: $E[\hat{\boldsymbol{\beta}}_{\mathrm{LS}} \mid \boldsymbol{\beta}] = \boldsymbol{\beta}$

Desirable property of a genetic predictor:

The regression of y on the predictor has an intercept of zero and a slope of one.



Calculate PGS with BLUP estimates

Let $\mathbf{Z}_i'$ is the genotypes of an individual to be predicted

SNP	1	2	3	4	5	6	7	8	9	10
Geno	Z_{i1}	Z_{i2}	Z_{i3}	Z_{i4}	Z_{i5}	Z_{i6}	Z_{i7}	Z_{i8}	Z_{i9}	Z_{i10}

$$PGS_i = \mathbf{Z}_i'\hat{\boldsymbol{\beta}}_{BLUP}$$ using all SNPs

SNP	1	2	3	4	5	6	7	8	9	10
Geno	Z_{i1}	Z_{i2}	Z_{i3}	Z_{i4}	Z_{i5}	Z_{i6}	Z_{i7}	Z_{i8}	Z_{i9}	Z_{i10}
$\hat{\boldsymbol{\beta}}_{BLUP}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$

$$= \sum_{j=1}^{10} Z_{ij} * \hat{\beta}_j$$

Where do we get λ from?

- If know σ_β^2, then know λ
- Can estimate total additive genetic variance (σ_g^2) and divide by number of segments, e.g. $\sigma_\beta^2 = \sigma_g^2/m$
- Assumes SNPs capture all of genetic variance!
- Estimate with REML
- Bayesian approach
- Cross validation

From Ben Hayes

Summary-data-based BLUP (SBLUP)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

BLUP solutions:

$$\widehat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]^{-1}\mathbf{X}'\boldsymbol{y}$$

$n\,\mathbf{R}$ $\qquad$ $n\,\mathbf{b}$

where $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$

Let

$\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ → LD matrix

$b_j = \frac{1}{n}\mathbf{X}_j'\mathbf{y}$ → GWAS effects

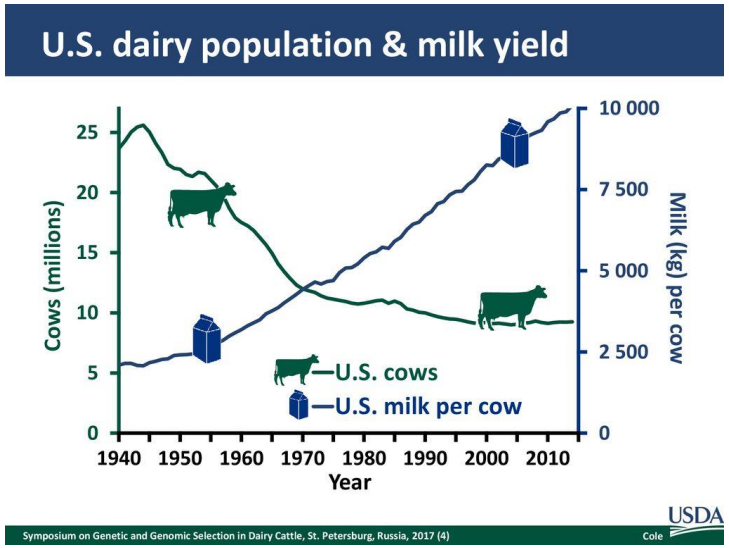
R (LD matrix), **b** (GWAS marginal effects) and n (sample size) are **sufficient statistics** for the estimation of $\boldsymbol{\beta}$.

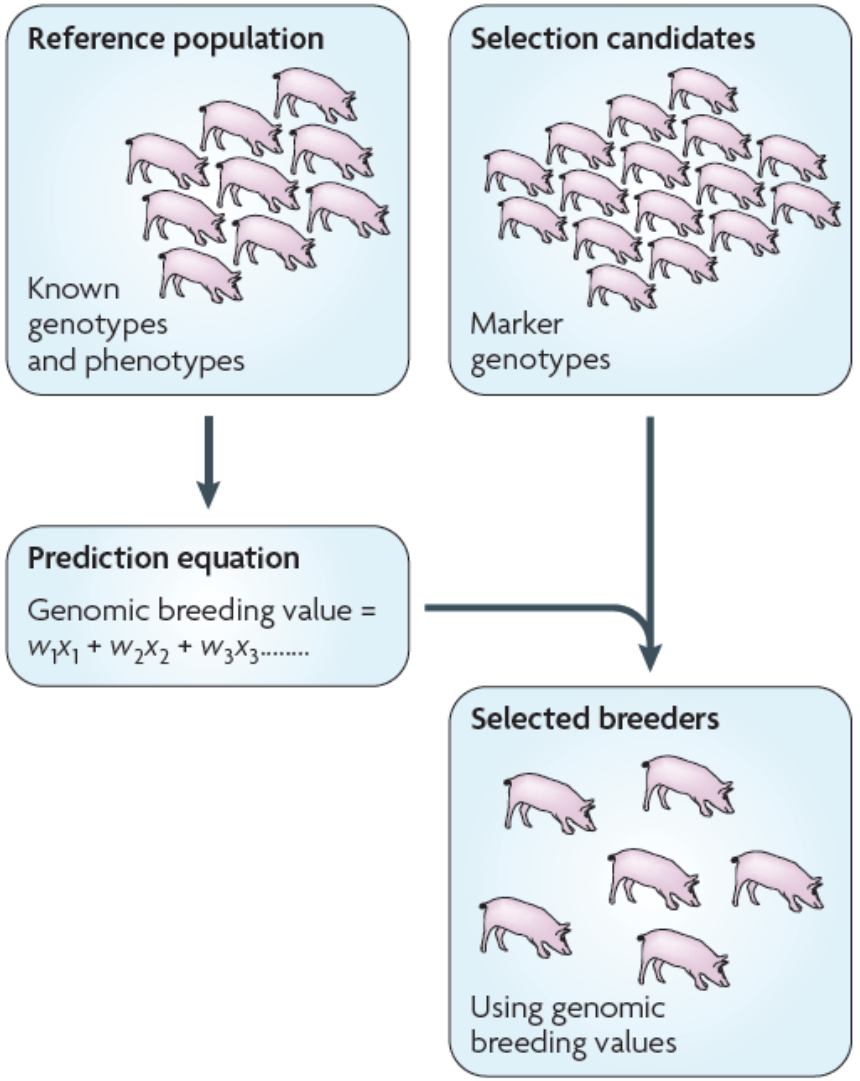
Examples of BLUP applications

Genomic selection in livestock

Use genome-wide SNPs to estimate the breeding value of selection candidates.

"Genomic selection" = "precision medicine" for animals





Examples of BLUP applications

Humans – Crohn's disease

Chen et al. 2017. BMC Medicine.

- Inflammatory Bowel Disease
- Affects 2 in every 1000 people (approx.)

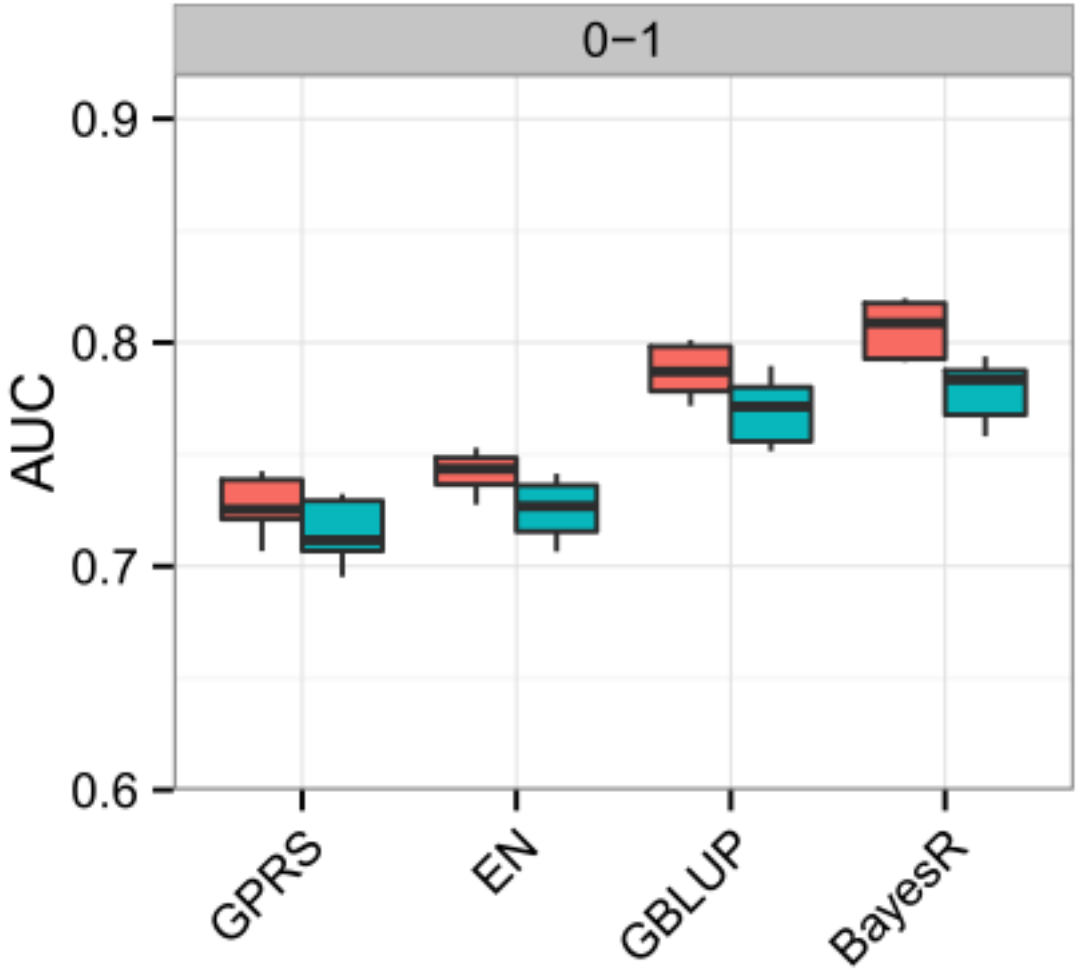
- 68,000 IBD patients and 29,000 healthy controls from 15 cohorts, European descent
- 909,763 GWAS SNPs or 123,437 SNPs on the custom designed Immunochip

- Prediction methods:
  - Genetic profile risk scores (GPRS) constructed using effects of all SNPs from GWAS
  - GBLUP
  - Elastic net (EN)
  - BayesR - Bayesian method that models SNP effects as a mixture of 4 normal distributions.

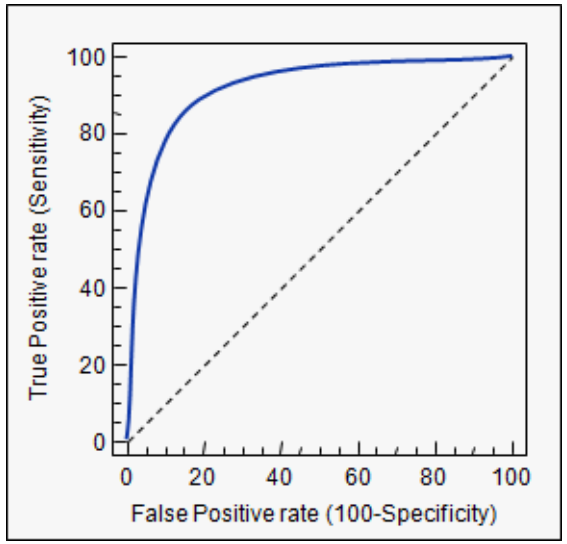
Examples of BLUP applications

Humans – Crohn's disease

Chen et al. 2017. BMC Medicine.



Assess value of predictions as "Area Under Curve" (AUC) from 5-fold cross-validation



Summary

- C+PT is a simple but commonly used method to calculate PGS
  - Involves SNP selection and requires an independent tuning sample
- BLUP simultaneously estimates all SNP effects as random
  - No need to prune SNPs on LD or select by p-value
  - Have nice properties, such as unbiasness & minimal prediction error variance, if the model correctly specified
  - Assumes normal distribution on SNP effects with equal variance
  - Need to specify the shrinkage parameter
  - Can work with GWAS summary statistics

Recommended reading

1. Choi SW, *et al*. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759–2772 (2020). (**General tutorial**)

2. Euesden J, *et al*. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015 May 1;31(9):1466-8. (**Popular tool implements the C+PT method**)

3. Goodard ME, *et al*. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Statist. Sci.* 24 (4) 517 - 529, November 2009. (**BLUP theory clearly explained**)

4. Maier RM, *et al*. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun* 9, 989 (2018). (**SBLUP methodology**)