THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Best Linear Unbiased Prediction (BLUP)

Jian Zeng

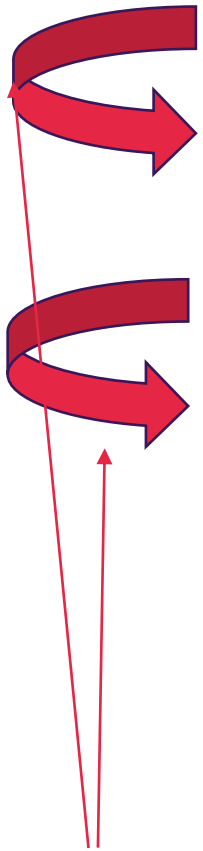
j.zeng@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Slides credit: Ben Hayes

Polygenic prediction



- **<u>Discovery</u>/Training/Derivation**
  - Estimate the effect sizes ($\hat{b}$) of SNPs on a trait (y) – GWAS
- **<u>Tunning</u>/Validation**
  - Further estimate some parameters (depends on methods; not all methods require it)
- **<u>Target</u>/Testing/Validation**
  - Build a polygenetic risk score (PRS) ($\hat{y}$):
  - Evaluate the prediction performance/accuracy

Should be independent; no overlap; out-of-sample prediction

GENOME-WIDE ASSOCIATION STUDIES — OPINION

Pitfalls of predicting complex traits from SNPs

Naomi R. Wray, Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard and Peter M. Visscher

Pitfall 1: No target sample – report R^2 in discovery sample

x: M markers for N samples

y from N(0,1) independently (null hypothesis)

1) Multiple linear regression of y on x (when M<N)

$E(R^2) = M/N$ variation "explained" by chance

2) Select m "best" markers out of M in total, and conduct multiple linear regression in the same dataset

$E(R^2) \gg m/N$ winner's curse

Winner's curse

ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel

~10 best markers selected from 2.5 million markers

Predicting phenotypes from genotypes

We used regression models to predict trait phenotypes from SNP genotypes and estimate the total variance explained by SNPs. The latter cannot be done by summing the individual contributions of the single marker effects because markers are not completely independent, and estimates of effects of single markers are biased when more than one locus affecting the trait segregates in the population. We derived gene-centred multiple regression models to estimate the effects of multiple SNPs simultaneously. In all cases 6–10 SNPs explain from 51–72% of the phenotypic variance and 65–90% of the genetic variance (Supplementary Tables 25 and 26 and Supplementary Figs 11–13). We also derived partial least square regression models using all SNPs for which the single marker effect was significant

"A cross-validated Bayesian prediction analysis using all genetic markers on the same data found that only 6% of phenotypic variation could be explained by the predictor."
(Wray et al., 2013. Nat. Rev. Genet.)

Pitfall 2: target sample overlapped with discovery sample

- Overlapping target and discovery sample
- Greater similarity between target and discovery sample (such as relatedness)
  - Cross-validation: not a pitfall, but to be aware

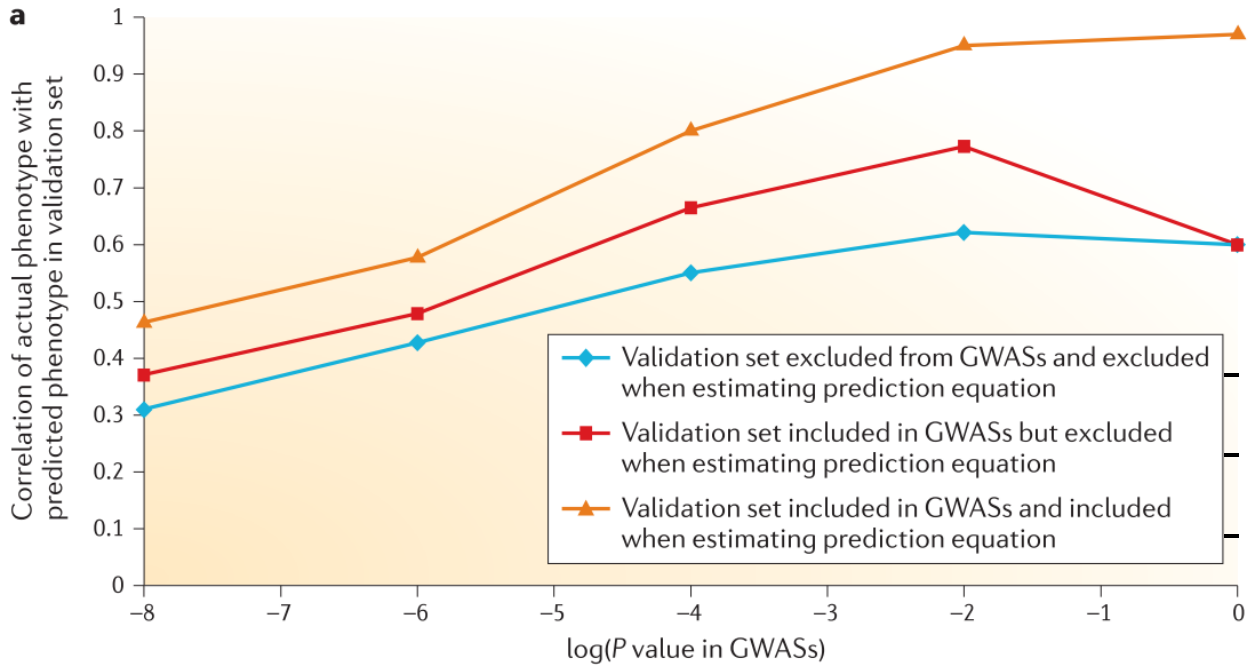
$$\mathrm{cov}(\hat{y}_i, y_i) = \mathrm{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij} b_j + e_i\}$$

$$= \sum_{j=1}^{m} \mathrm{var}(x_{ij})\hat{b}_j b_j + \sum_{j=1}^{m} x_{ij}\,\mathrm{cov}(\hat{b}_j, e_i)$$

If b estimated from the same data in which prediction is made, then the second term is non-zero

Pitfall 3: Less obvious non-independence

- Estimate SNP effects and/or select SNPs from total sample (discovery + target sample)
- Re-estimate effects in the target sample after selecting in the discovery sample



- Out-of-sample prediction
- Select SNPs in total sample
- Direct report R2 in the discovery sample

Think

What about selecting SNPs (P-value thresholding) in the target sample as we did in the 1st practical?

Is it an issue?

If so, how could we address it?

Polygenic score (PGS) methods

A weighted sum of the count of risk alleles

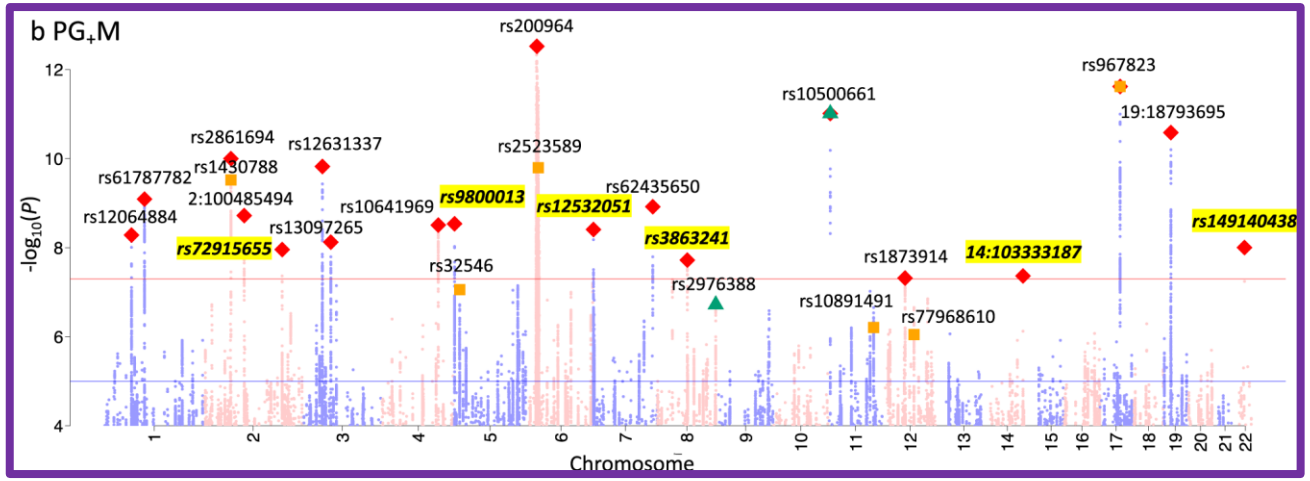
$$\text{PGS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold in a tuning sample



Polygenic score (PGS) methods

A weighted sum of the count of risk alleles

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

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold in a tuning sample

⇨ May neglect small effects (false negatives)

Requires additional tuning sample

Polygenic score (PGS) methods

Can we simultaneously use all SNPs without SNP selection?

Yes! But …

Cannot directly aggregate GWAS effects due to linkage disequilibrium (double counting)

No unique solutions from multiple regression (least squares) when #SNPs > #individuals (p > n problem)

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

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

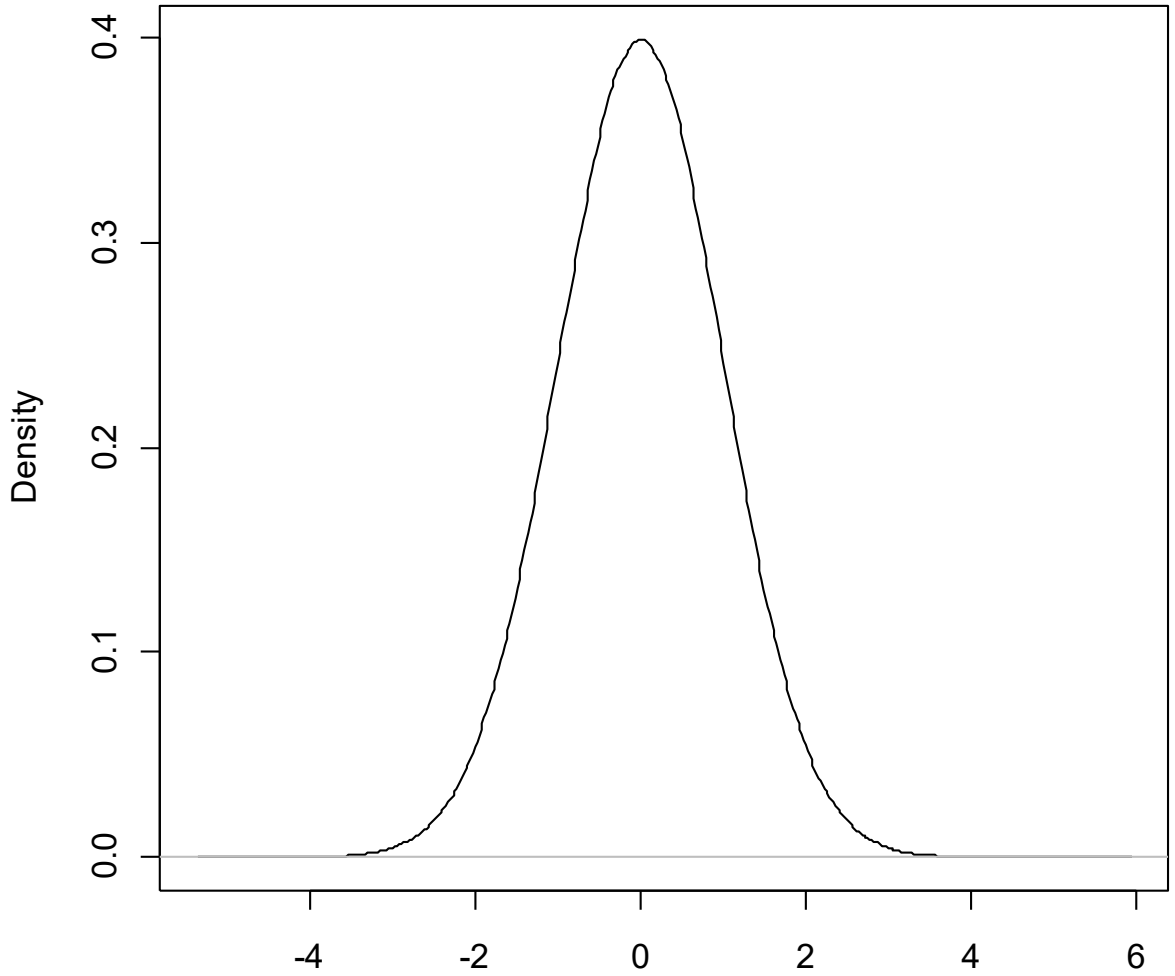
where

- $\mathbf{y}$ 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

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 $\mathbf{y}$

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

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

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

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

BLUP solutions

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

LS solutions

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

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

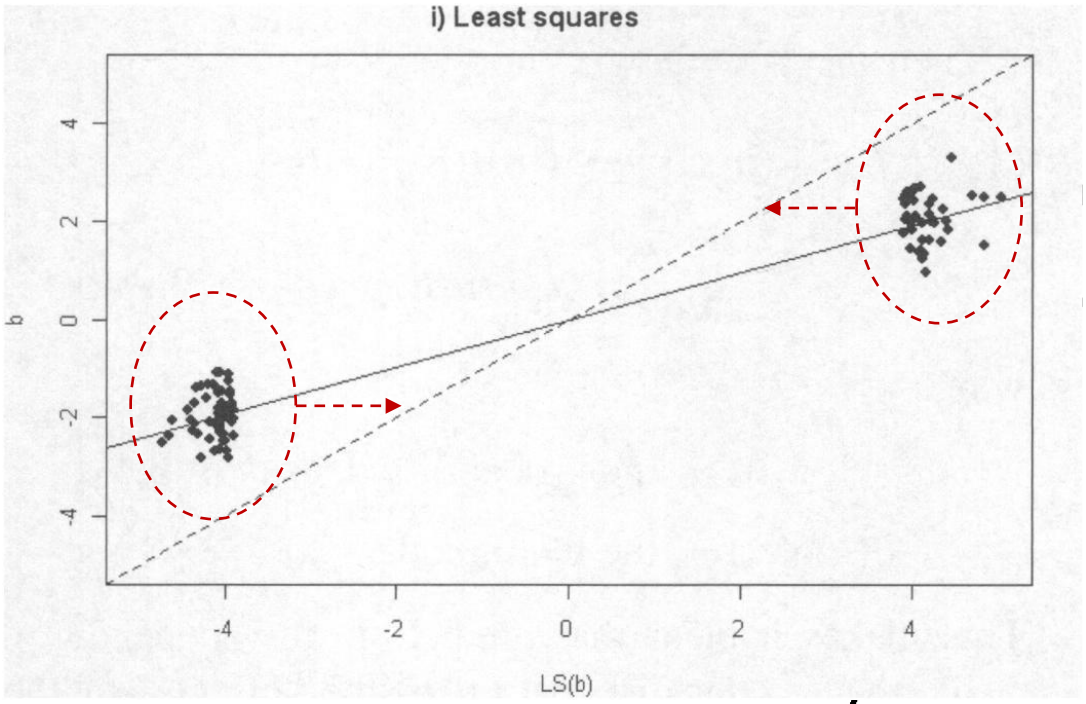
Shrinks LS estimates toward zero

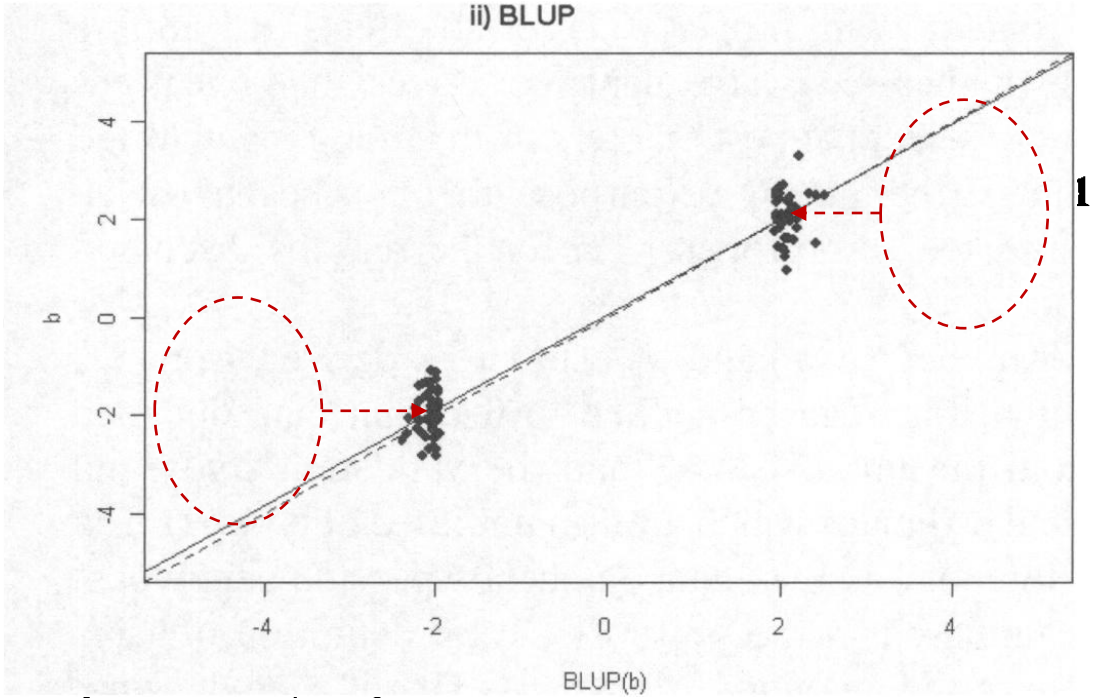
ronmental" factors. For example, height in humans involves many physiological processes and many genes but is also influenced by nongenetic factors such as nutrition and health care. These traits are called quanti-

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.

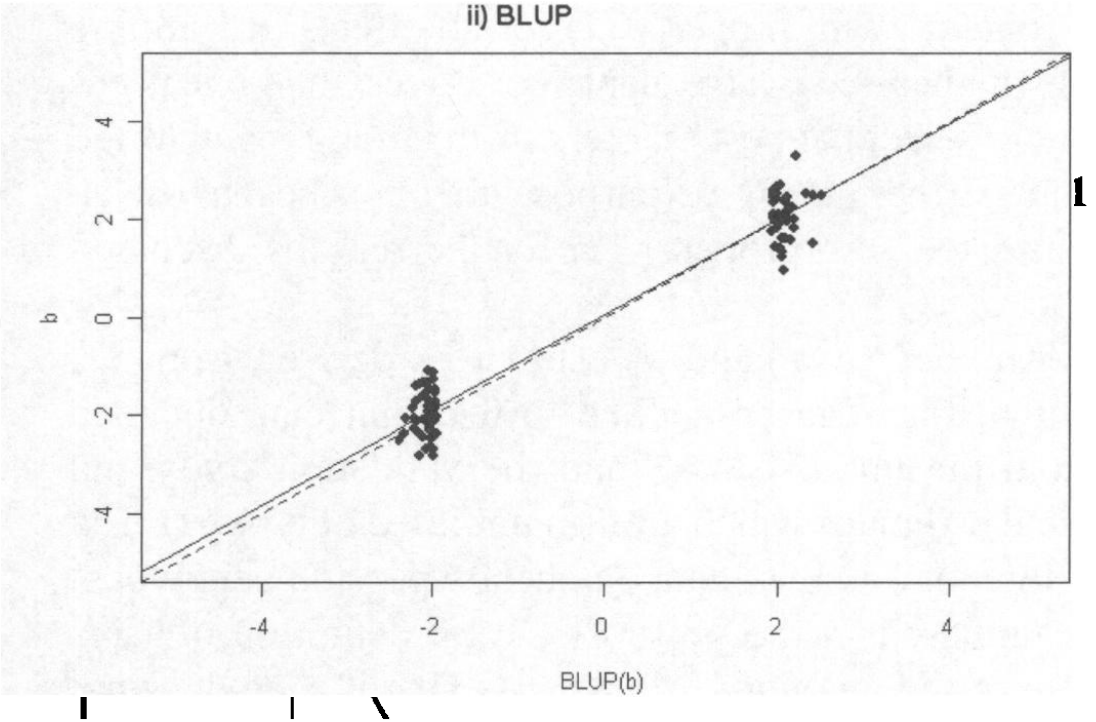
SNPs usually have no direct effect on a trait under study. However, any polymorphism that does affect the trait will be located on a chromosome close to

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}
$\hat{\boldsymbol{\beta}}_{BLUP}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$

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

Estimate λ

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

Summary-data-based BLUP (SBLUP)

Consider

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

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

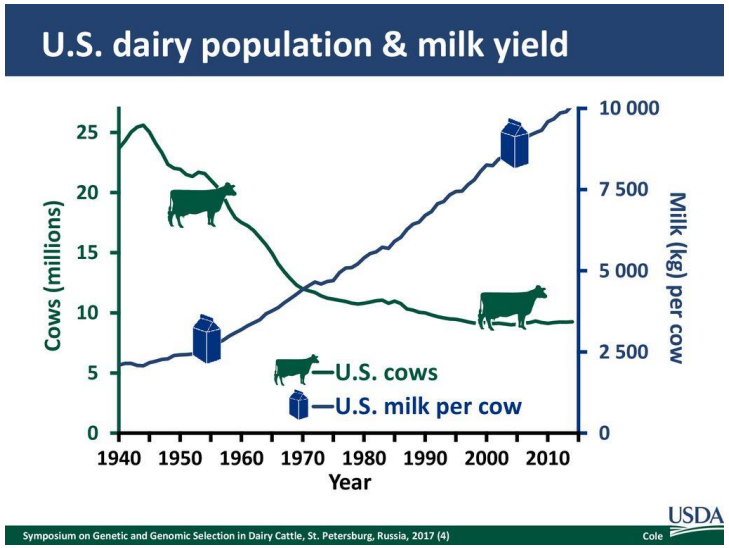
R (LD matrix), **b** (GWAS marginal effects) and n (sample size) are **sufficient statistics** for the estimation of $\boldsymbol{\beta}$.

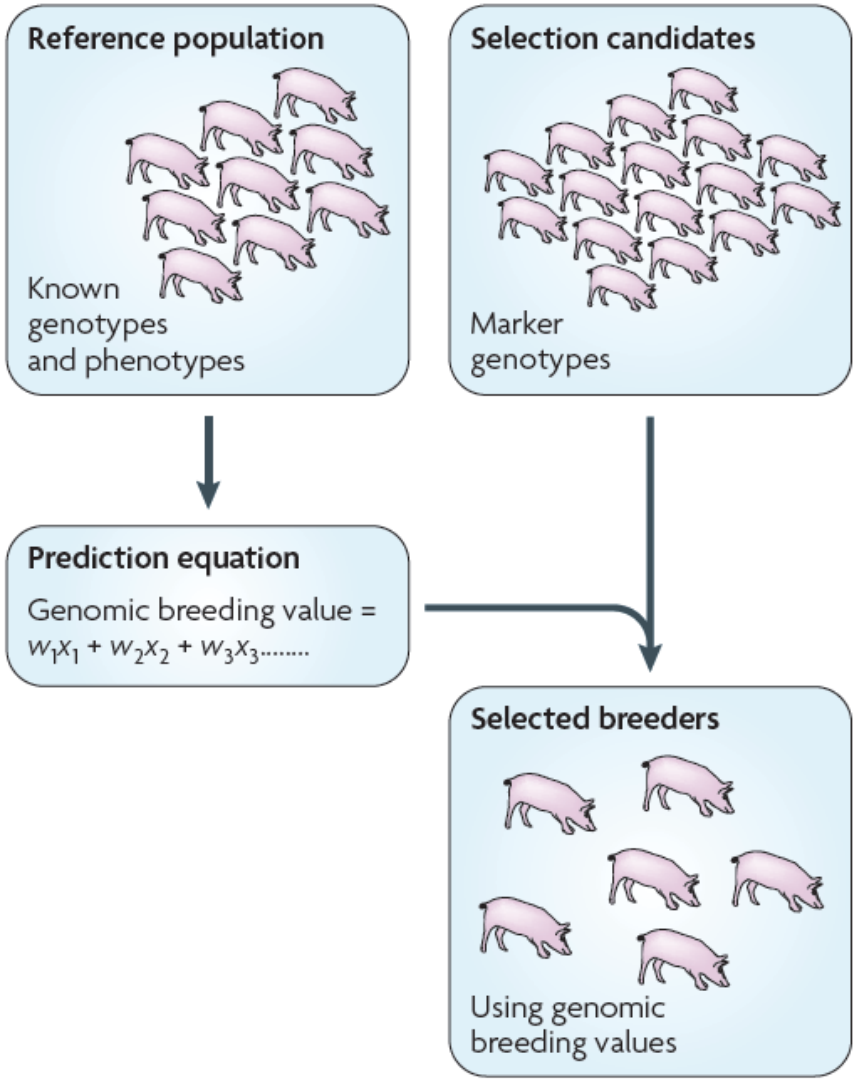
Examples of BLUP applications

Genomic selection in livestock

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

"Genomic selection" = "precision medicine" for animals





Humans – Crohn's disease

Chen et al. 2017. BMC Medicine.

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

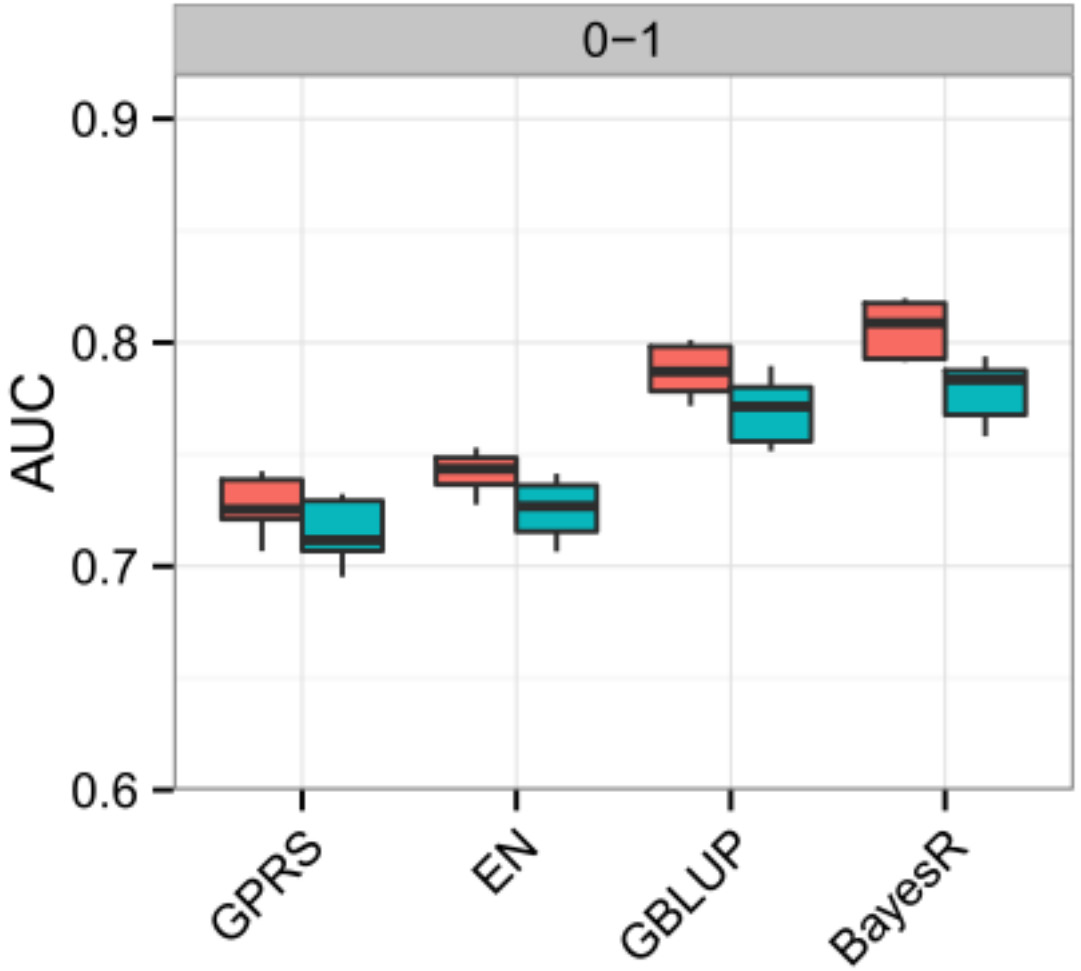
- 68,000 IBD patients and 29,000 healthy controls from 15 cohorts, European descent
- 909,763 GWAS SNPs or 123,437 SNPs on the custom designed Immunochip

- 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.

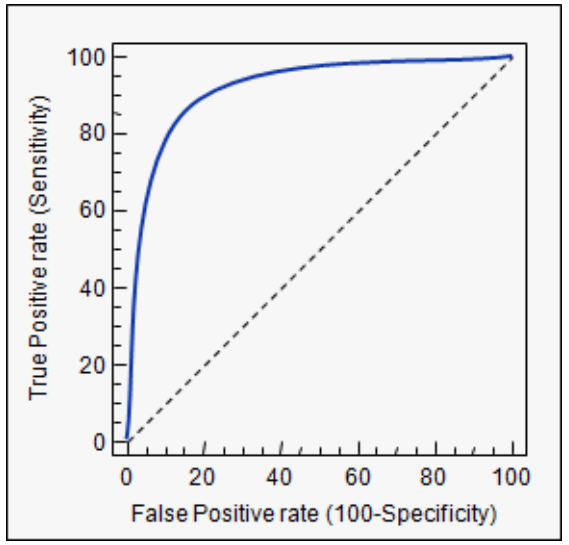
Examples of BLUP applications

Humans – Crohn's disease

Chen et al. 2017. BMC Medicine.



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



BLUP

- Simultaneously estimate all SNP effects as random
  - No need to prune on LD or select p-value threshold
  - No need to know causal variants or biological function
- Assumes normal distribution on SNP effects with equal variance
- Need to specify the shrinkage parameter
- Unbiased estimates of SNP effects
- Improved prediction accuracy in practice

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. Wray NR, *et al*. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet*. 14, 507–15 (2013). (**Pitfalls of interpretation**)

3. Goddard 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**)

Questions?

5 min break

Practical 2: BLUP

https://cnsgenomics.com/data/teaching/GNGWS26/module5/Practical2_BLUP.html

To log into your server, type command below in **Terminal** for Mac/Linux users or in **Command Prompt** or **PowerShell** for Windows users.

```
ssh username@hostname
```

And then key in the provided password.