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 score (PGS) methods

A weighted sum of the count of risk alleles

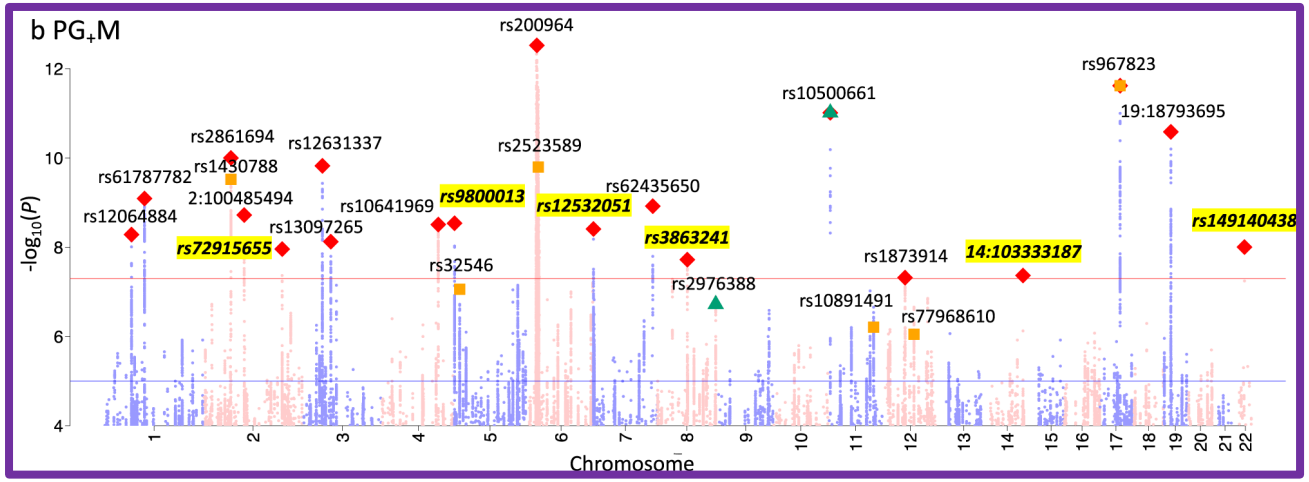
$$\text{PGS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

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



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?

Can we simultaneously use all SNPs?

Yes! But …

cannot aggregate GWAS effects

due to linkage disequilibrium (double counting)

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?

Estimate SNP effects with a multiple regression?

Yes!

But …

Least squares method

Linear model

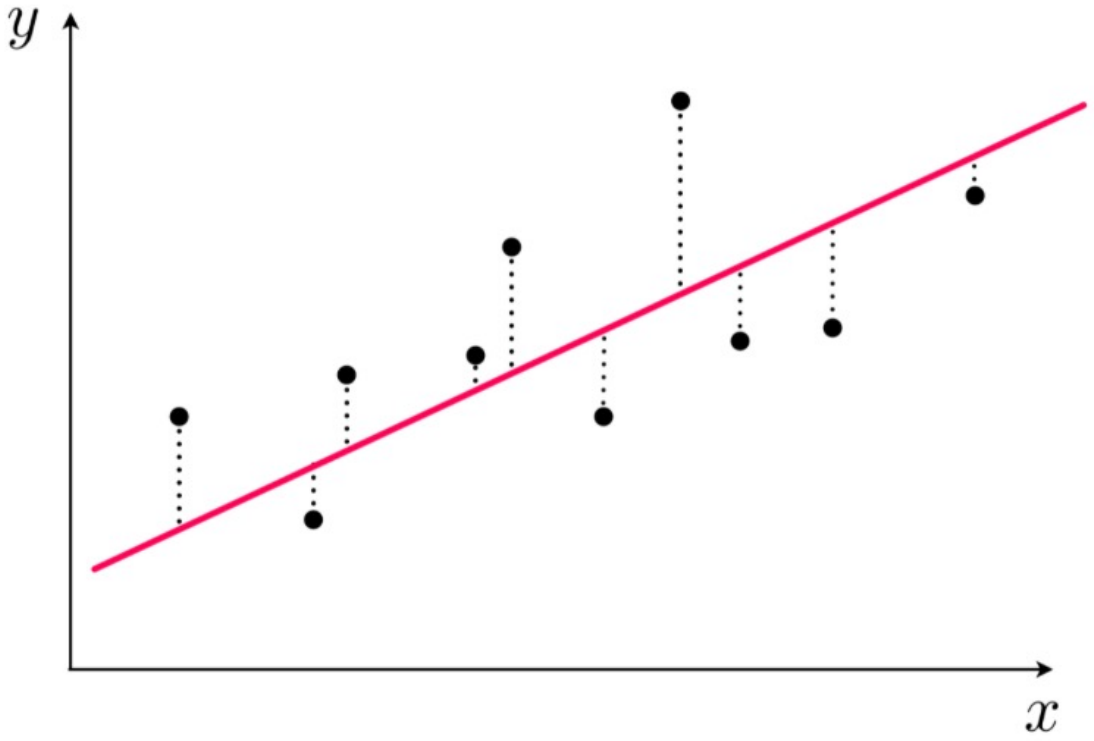
$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

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

Least squares method

Least squares (LS): minimising the sum of squares of the residuals.



Least squares method

Linear model

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

LS solutions

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

No unique solutions when #SNPs > #individuals

($p > n$ problem)

Linear mixed model

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

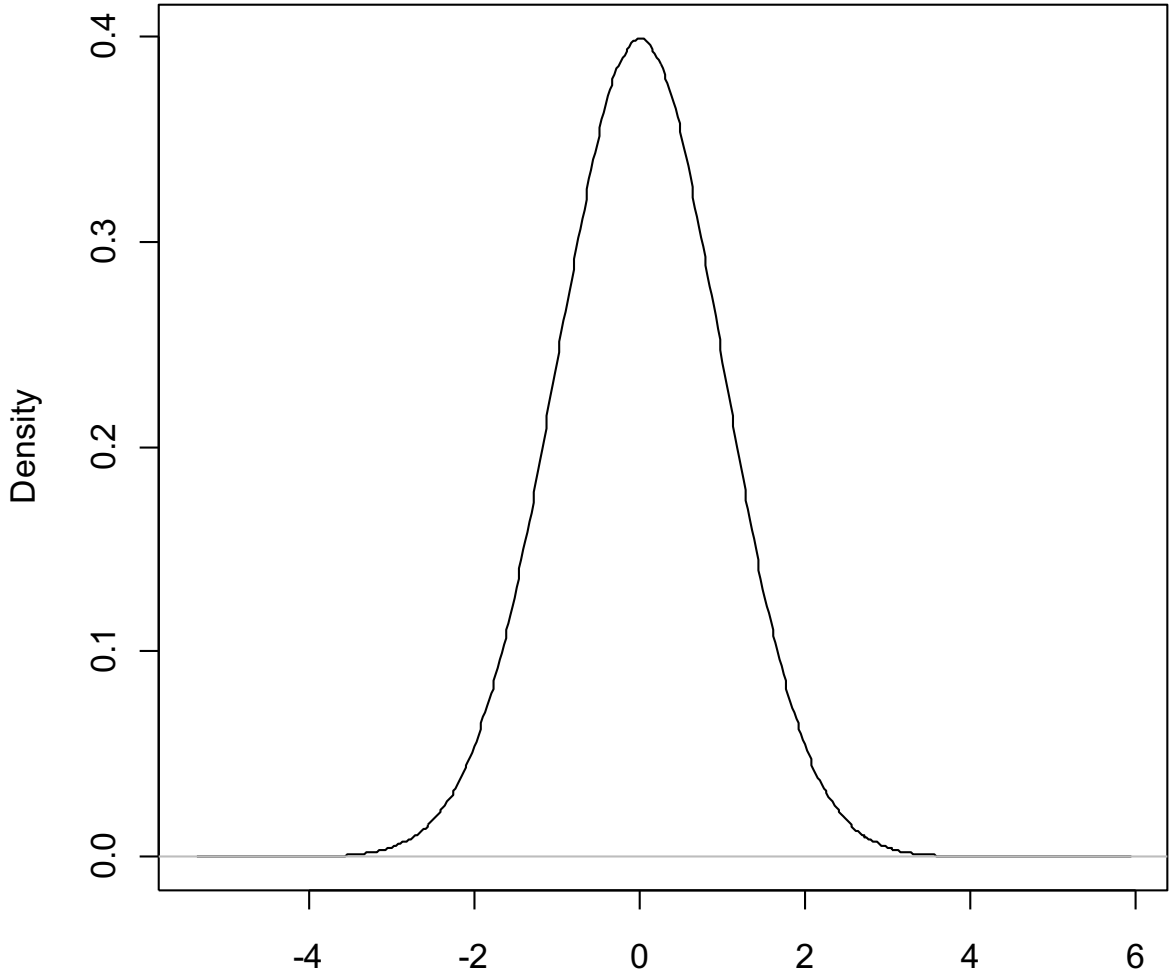
where

- **y** is a vector of *n* phenotypes,
- μ is the mean,
- **X** is an incidence matrix of individuals' genotypes for all SNPs,
- $\boldsymbol{\beta}$ are the random effects of the *m* SNPs,
- **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

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

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

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

BLUP solutions

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

LS solutions

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

Example

- 10 SNPs
- Only 5 phenotypes

		X									
Individual	y	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

Example

Let $1_n' = [1\ 1\ 1\ 1\ 1]$

Assume value of 1 for λ

		X									
Individual	y	1	2	3	4	5	6	7	8	9	10
1	0.19	0	0	0	0	0	0	1	2	0	2
2	1.23	1	0	0	1	1	1	2	1	0	1
3	0.86	1	0	0	1	0	0	1	1	1	1
4	1.23	1	1	1	1	0	1	2	1	1	1
5	0.45	0	1	1	1	1	1	2	1	0	1

BLUP solutions

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

Example

BLUP solutions

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

5	3	2	2	4	2	3	8	6	2	6	3.96
3	4	1	1	3	1	2	5	3	2	3	3.32
2	1	3	2	2	1	2	4	2	1	2	1.68
2	1	2	3	2	1	2	4	2	1	2	1.68
4	3	2	2	5	2	3	7	4	2	4	3.77
2	1	1	1	2	3	2	4	2	0	2	1.68
3	2	2	2	3	2	4	6	3	1	3	2.91
8	5	4	4	7	4	6	15	9	3	9	6.87
6	3	2	2	4	2	3	9	9	2	8	4.15
2	2	1	1	2	0	1	3	2	3	2	2.09
6	3	2	2	4	2	3	9	8	2	9	4.15

Example

BLUP solutions

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

5	3	2	2	4	2	3	8	6	2	6	3.96
3	4	1	1	3	1	2	5	3	2	3	3.32
2	1	3	2	2	1	2	4	2	1	2	1.68
2	1	2	3	2	1	2	4	2	1	2	1.68
4	3	2	2	5	2	3	7	4	2	4	3.77
2	1	1	1	2	3	2	4	2	0	2	1.68
3	2	2	2	3	2	4	6	3	1	3	2.91
8	5	4	4	7	4	6	15	9	3	9	6.87
6	3	2	2	4	2	3	9	9	2	8	4.15
2	2	1	1	2	0	1	3	2	3	2	2.09
6	3	2	2	4	2	3	9	8	2	9	4.15

Example

BLUP solutions

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

5	3	2	2	4	2	3	8	6	2	6	3.96
3	4	1	1	3	1	2	5	3	2	3	3.32
2	1	3	2	2	1	2	4	2	1	2	1.68
2	1	2	3	2	1	2	4	2	1	2	1.68
4	3	2	2	5	2	3	7	4	2	4	3.77
2	1	1	1	2	3	2	4	2	0	2	1.68
3	2	2	2	3	2	4	6	3	1	3	2.91
8	5	4	4	7	4	6	15	9	3	9	6.87
6	3	2	2	4	2	3	9	9	2	8	4.15
2	2	1	1	2	0	1	3	2	3	2	2.09
6	3	2	2	4	2	3	9	8	2	9	4.15

Example

BLUP solutions

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

5	3	2	2	4	2	3	8	6	2	6	3.96
3	4	1	1	3	1	2	5	3	2	3	3.32
2	1	3	2	2	1	2	4	2	1	2	1.68
2	1	2	3	2	1	2	4	2	1	2	1.68
4	3	2	2	5	2	3	7	4	2	4	3.77
2	1	1	1	2	3	2	4	2	0	2	1.68
3	2	2	2	3	2	4	6	3	1	3	2.91
8	5	4	4	7	4	6	15	9	3	9	6.87
6	3	2	2	4	2	3	9	9	2	8	4.15
2	2	1	1	2	0	1	3	2	3	2	2.09
6	3	2	2	4	2	3	9	8	2	9	4.15

Example

BLUP solutions

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

5.96	-0.46	-0.04	-0.04	-0.81	-0.31	-0.01	-1.01	-1.19	-0.50	-1.19		3.96
-0.46	0.65	0.11	0.11	-0.11	0.08	-0.06	-0.06	0.11	-0.18	0.11		3.32
-0.04	0.11	0.72	-0.28	-0.03	0.04	-0.11	-0.11	0.03	-0.07	0.03		1.68
-0.04	0.11	-0.28	0.72	-0.03	0.04	-0.11	-0.11	0.03	-0.07	0.03		1.68
-0.81	-0.11	-0.03	-0.03	0.83	-0.09	-0.05	-0.05	0.17	-0.09	0.17		3.77
-0.31	0.08	0.04	0.04	-0.09	0.68	-0.12	-0.12	0.09	0.24	0.09		1.68
-0.01	-0.06	-0.11	-0.11	-0.05	-0.12	0.76	-0.24	0.05	0.07	0.05		2.91
-1.01	-0.06	-0.11	-0.11	-0.05	-0.12	-0.24	0.76	0.05	0.07	0.05		6.87
-1.19	0.11	0.03	0.03	0.17	0.09	0.05	0.05	0.83	0.09	-0.17		4.15
-0.50	-0.18	-0.07	-0.07	-0.09	0.24	0.07	0.07	0.09	0.68	0.09		2.09
-1.19	0.11	0.03	0.03	0.17	0.09	0.05	0.05	-0.17	0.09	0.83		4.15

Example

BLUP solutions

Mean	0.47
SNP1	0.29
SNP2	-0.05
SNP3	-0.05
SNP4	0.08
SNP5	-0.02
SNP6	0.13
SNP7	0.13
SNP8	-0.08
SNP9	0.11
SNP10	-0.08

"Smear" the effect over SNPs in LD

PGS prediction with BLUP

Now we want to predict PGS of a group of young individuals without phenotypes

$$\mathbf{PGS} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

We have the $\widehat{\boldsymbol{\beta}}$, and we can get $\mathbf{X}$ from their genotypes (after genotyping)……

Young individuals	X									
1	1	1	1	1	1	1	2	1	0	1
2	1	0	0	1	1	1	1	1	0	1
3	1	0	0	1	1	1	2	1	0	1
4	1	0	0	1	1	2	2	1	0	1
5	0	0	0	0	0	0	1	2	0	2

PGS prediction with BLUP

$$\mathbf{PGS} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

$\mathbf{X}$										$\widehat{\boldsymbol{\beta}}$	PGS
1	1	1	1	1	1	2	1	0	1	0.29	0.48
1	0	0	1	1	1	1	1	0	1	-0.05	0.45
1	0	0	1	1	1	2	1	0	1	-0.05	0.58
1	0	0	1	1	2	2	1	0	1	0.08	0.71
0	0	0	0	0	0	1	2	0	2	-0.02	-0.19
										0.13	
										0.13	
										-0.08	
										0.11	
										-0.08	

Shrinkage

BLUP solutions

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

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

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

Shrinkage

BLUP solutions

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

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

Ignoring mean and other SNP

$\hat{\beta}_1 = \frac{X_1'y}{X_1'X_1+\lambda}$

= (0*0.19+1*1.23+1*0.86+1*1.23+ 0*0.45)/(3+1)

		X
Individual	Y	1
1	0.19	0
2	1.23	1
3	0.86	1
4	1.23	1
5	0.45	0

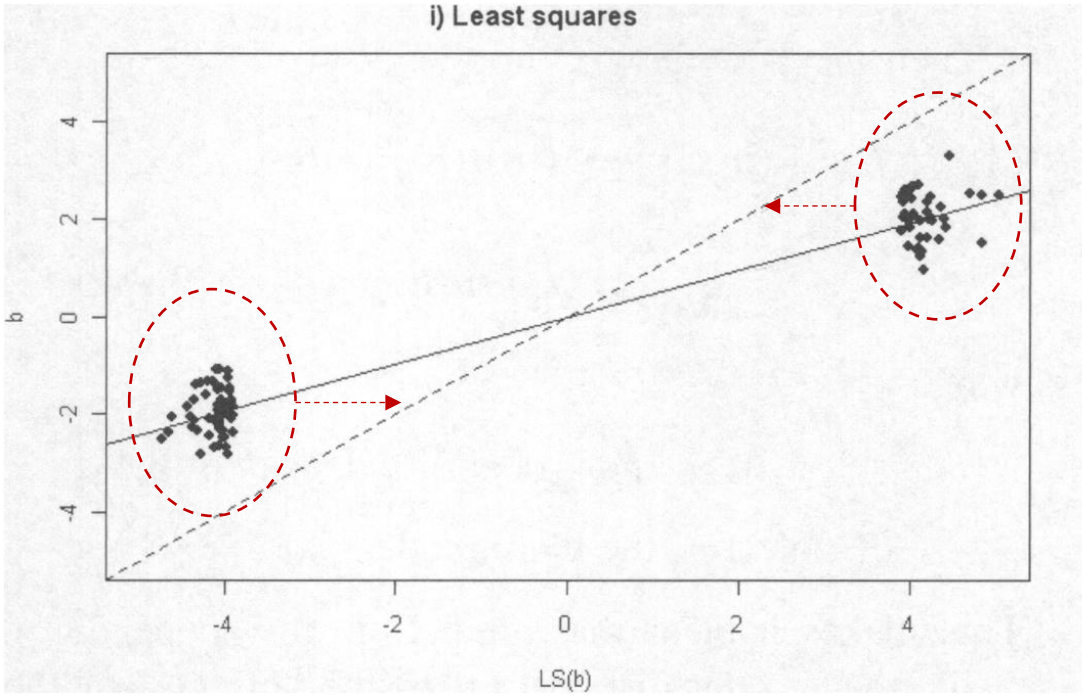
Shrinkage

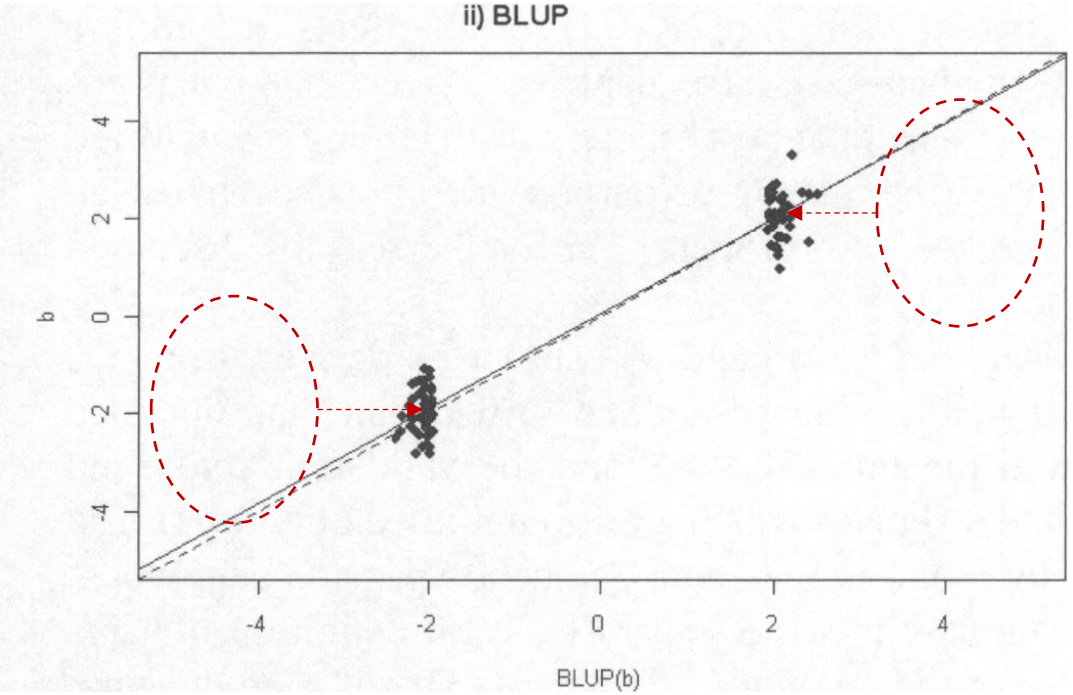
Shrinks LS estimates toward zero

Statistical Science
2009, Vol. 24, No. 4, 517–529
DOI: 10.1214/09-STS306

Estimating Effects and Making Predictions from Genome-Wide Marker Data

Michael E. Goddard, Naomi R. Wray, Klara Verbyla and Peter M. Visscher





BLUP avoids selection bias!

Property of BLUP

Statistical Science
2009, Vol. 24, No. 4, 517–529
DOI: 10.1214/09-STS306

Estimating Effects and Making Predictions from Genome-Wide Marker Data

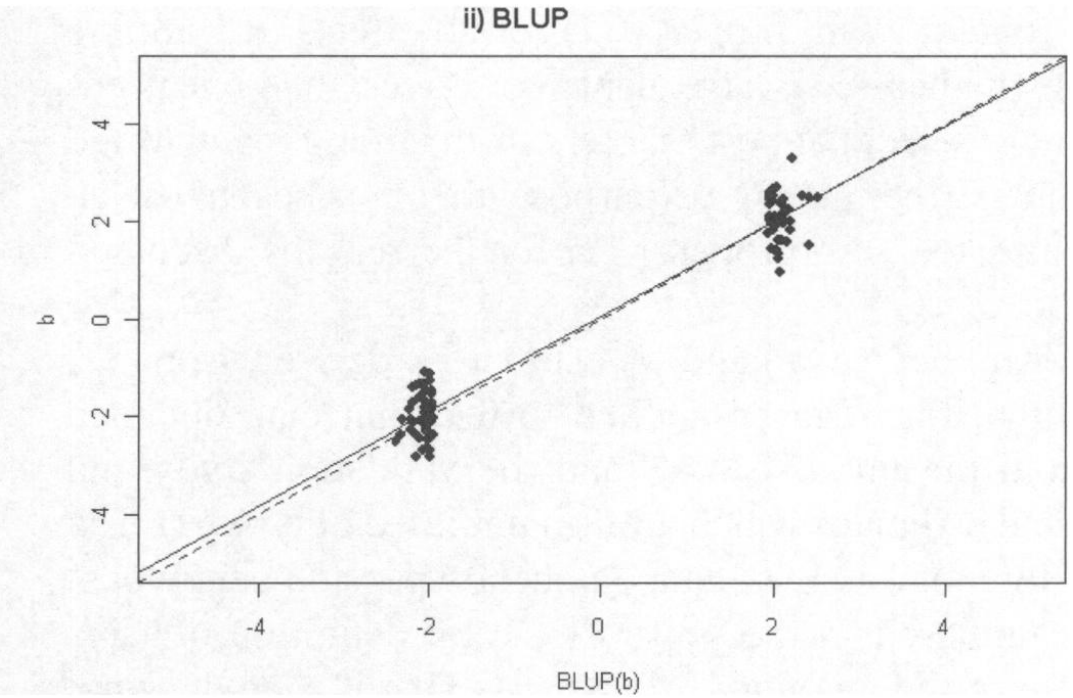
Michael E. Goddard, Naomi R. Wray, Klara Verbyla and Peter M. Visscher

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

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

Desirable property of a genetic predictor:

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



Estimate λ

Where do we get λ from?

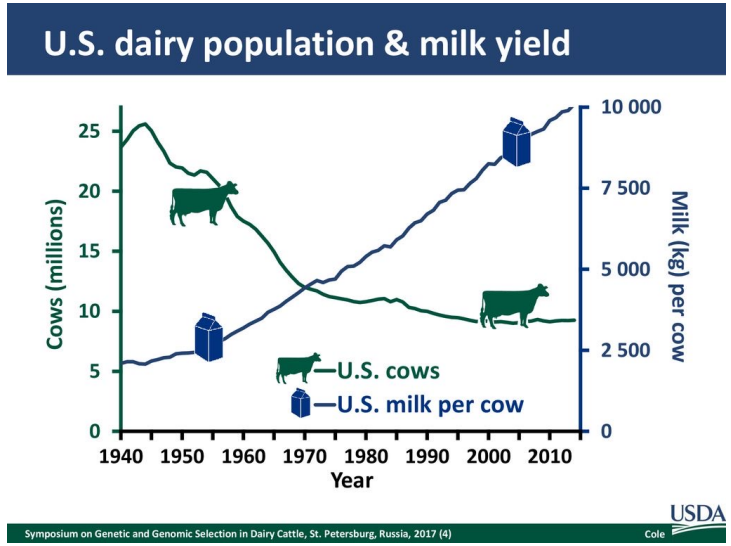
- If know σ_β^2, then know λ.
- Can estimate total additive genetic variance (σ_g^2) and divide by number of segments, e.g. $\sigma_\beta^2 = \sigma_g^2/m$
- Assumes SNPs capture all of genetic variance!
- Estimate with REML
- Bayesian approach
- Cross validation

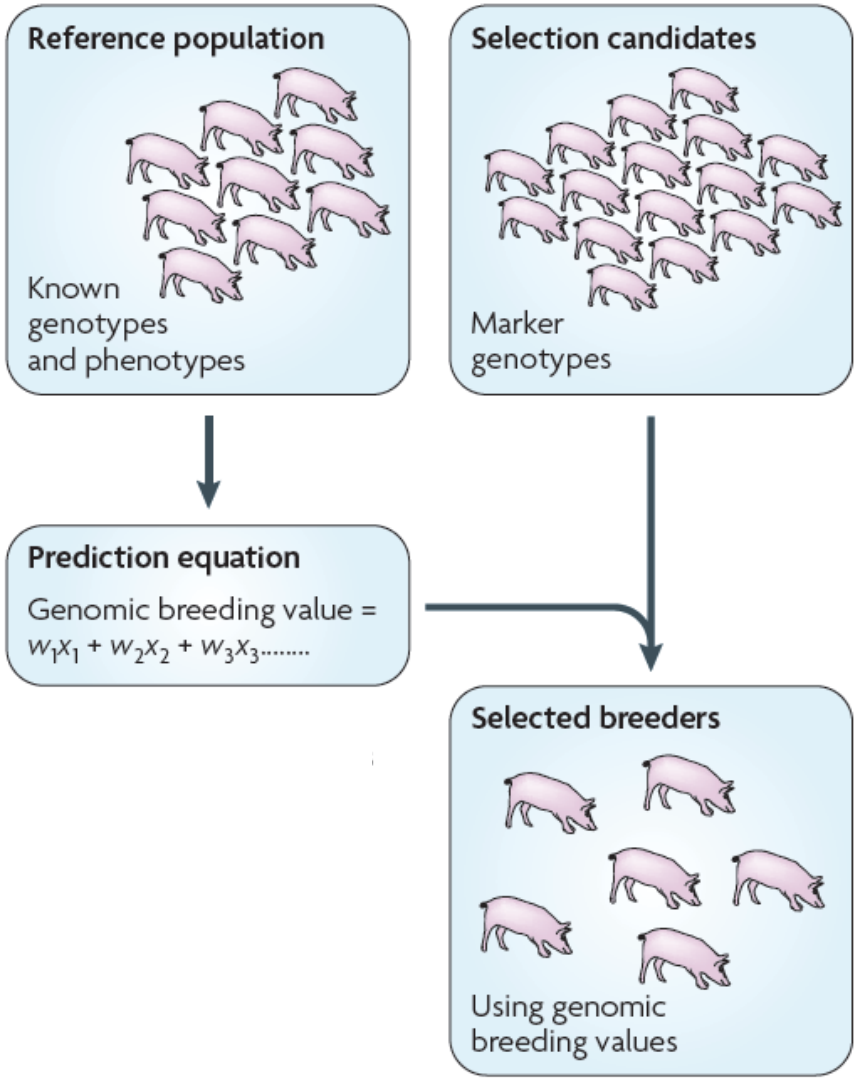
Examples of BLUP applications

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.)

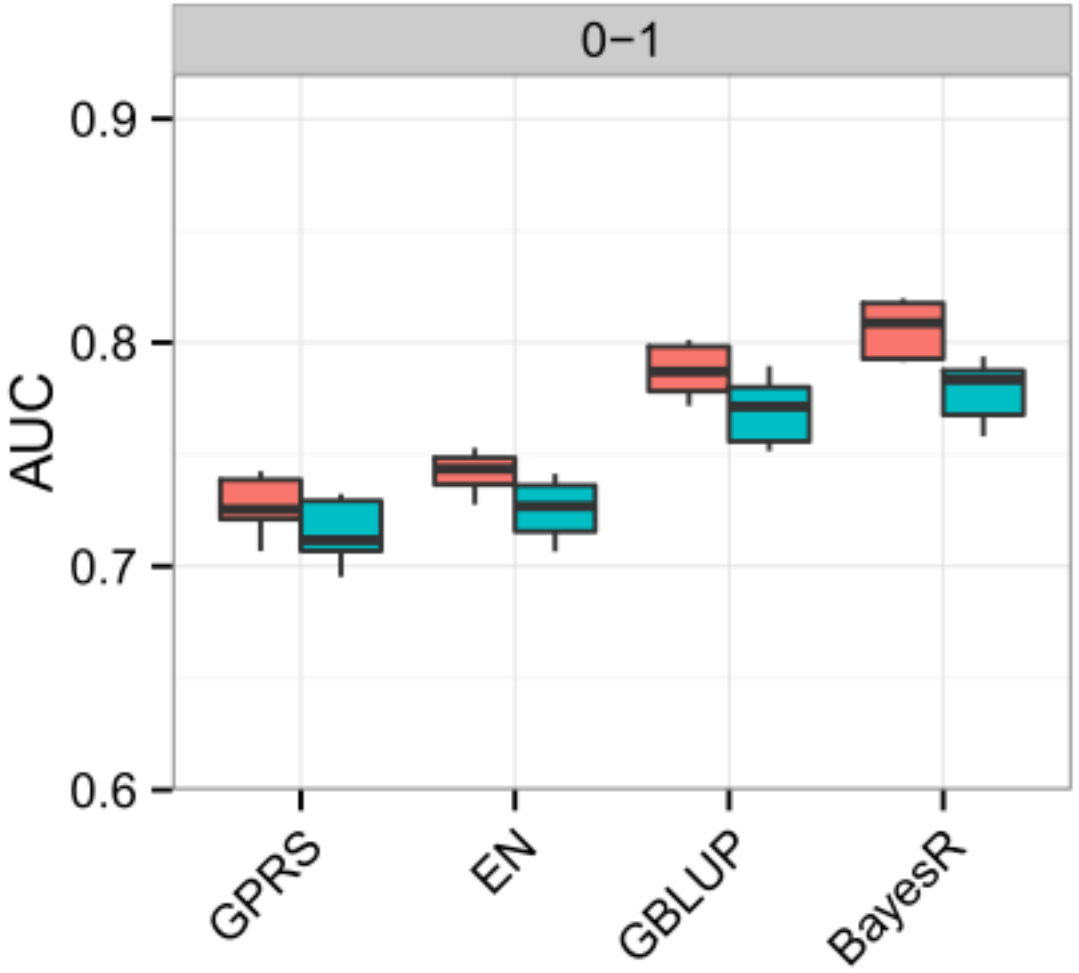
- 68,000 IBD patients and 29,000 healthy controls from 15 cohorts, European descent
- 909,763 GWAS SNPs or 123,437 SNPs on the custom designed Immunochip

- 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.

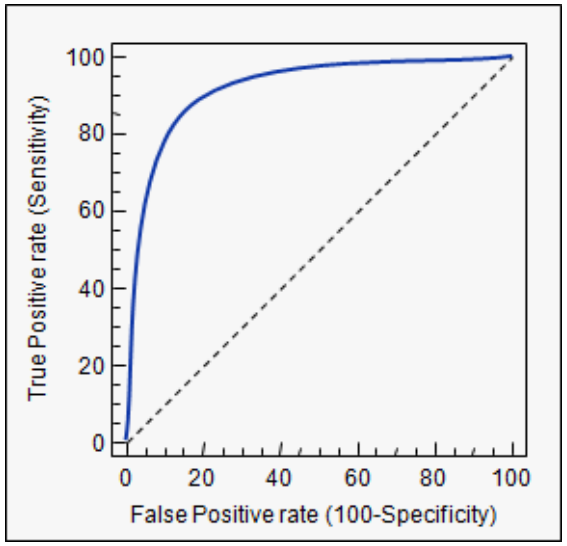
Examples of BLUP applications

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

Questions?

Practical 3: BLUP

https://cnsgenomics.com/data/teaching/GNGWS25/module5/Practical3_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.