

Best Linear Unbiased Prediction (BLUP)

G & G Winter School 2022

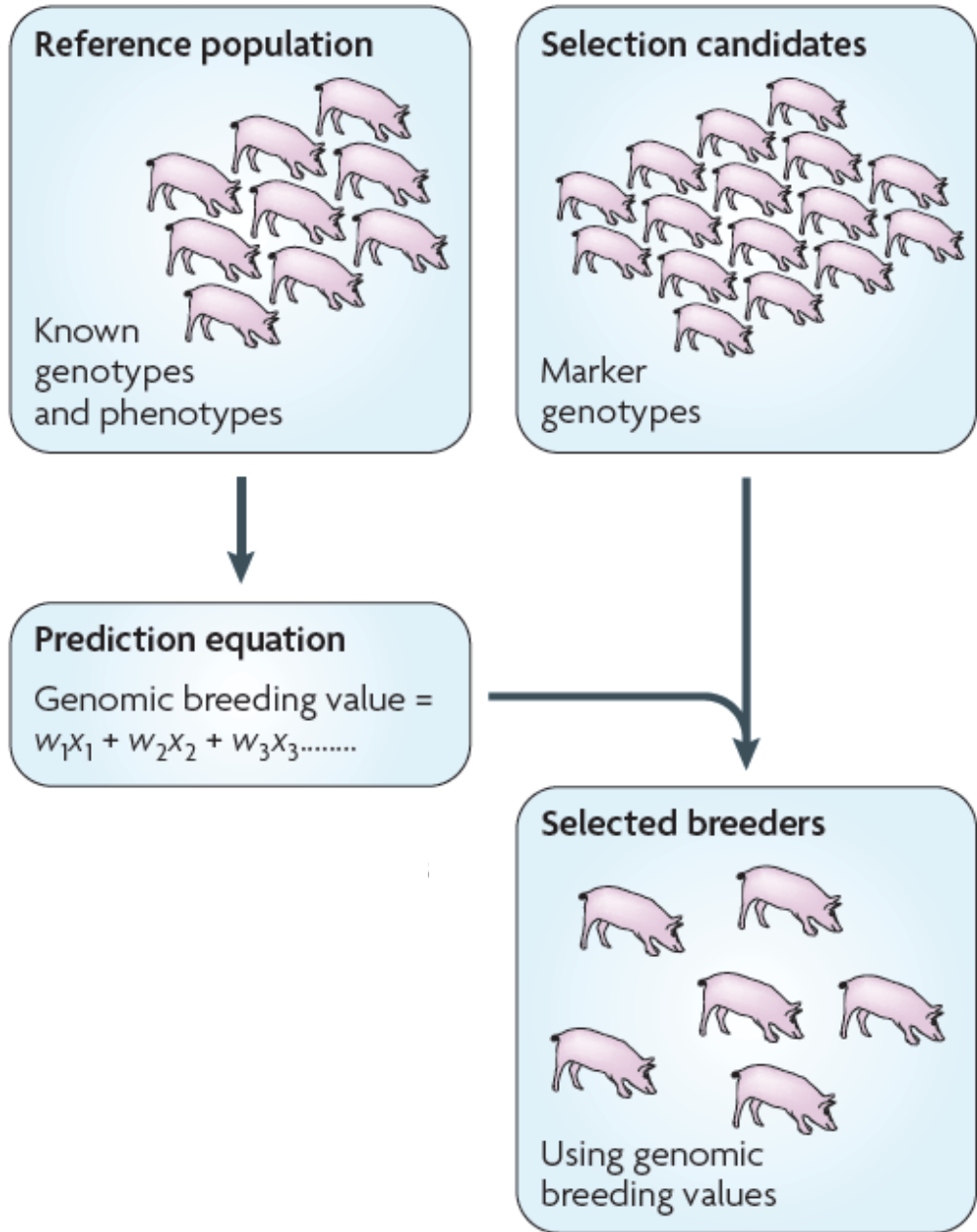
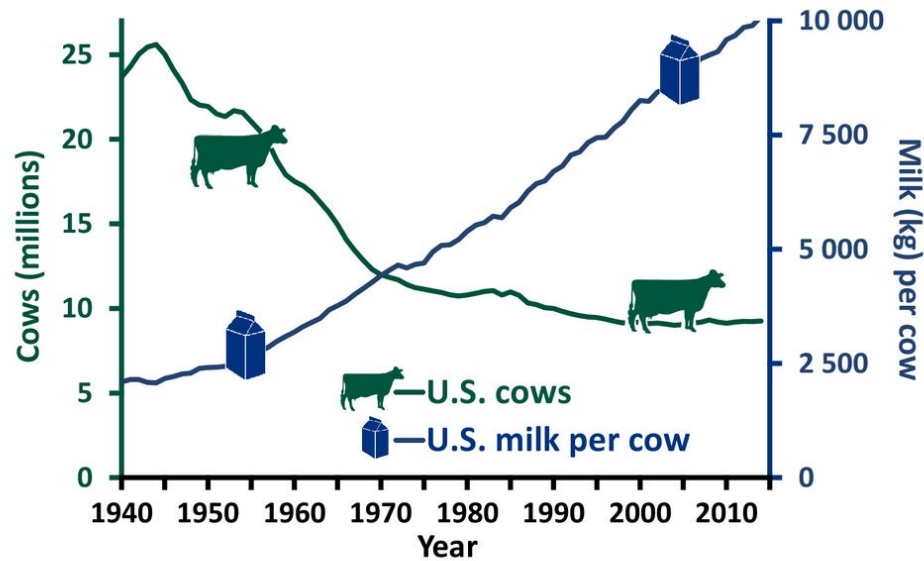
Jian Zeng



Slide courtesy of Prof. Ben Hayes

Genomic selection

U.S. dairy population & milk yield



Genomic prediction

- First step is to predict the SNP effects in a reference population
- Number of effects \ggg than number of records
- E.g. 50,000 SNPs
- From \sim 2000 records?
- Need methods that can deal with this

Linear mixed model

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

Where

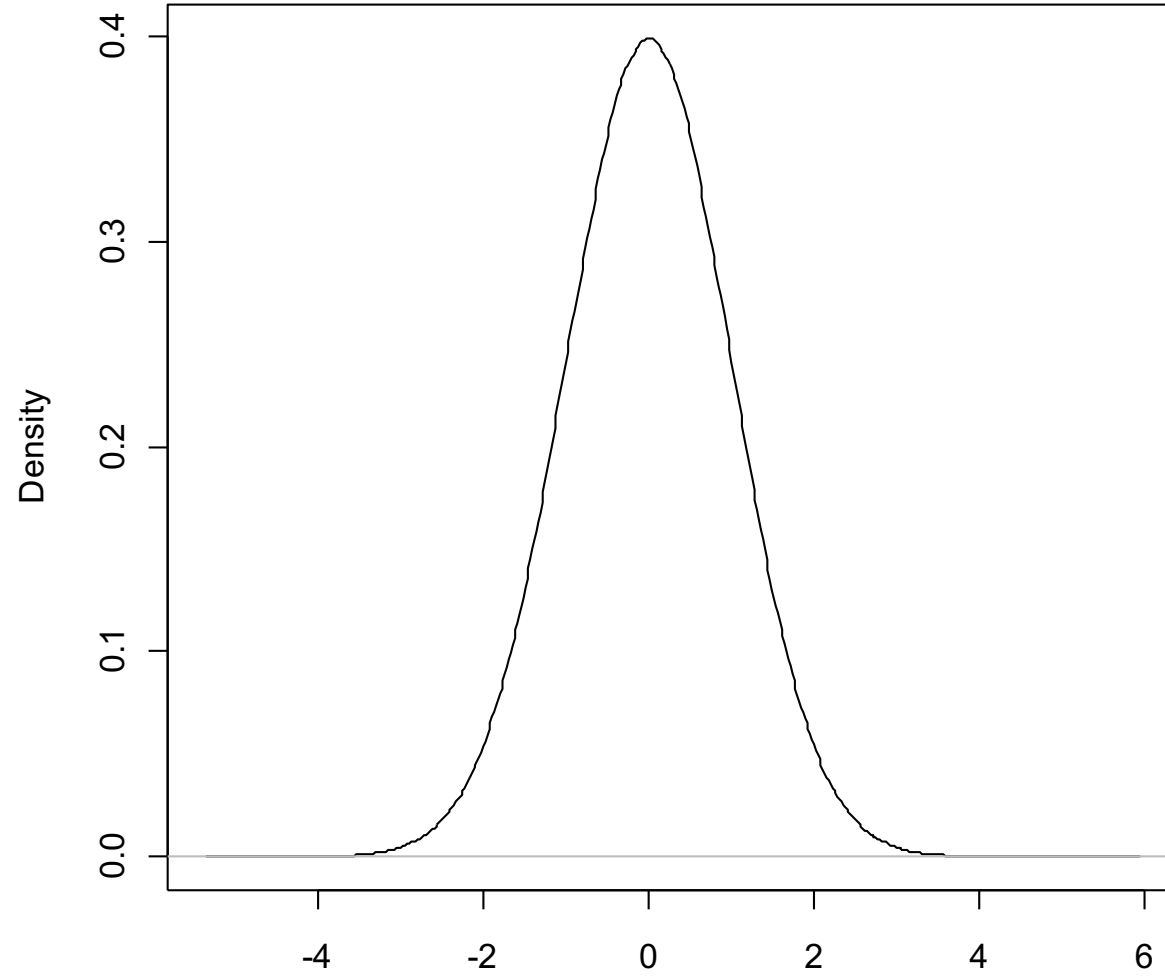
- \mathbf{y} is a vector of n phenotypes,
- μ is the mean,
- \mathbf{X} is an incidence matrix of individuals' genotypes for all SNPs,
- $\boldsymbol{\alpha}$ are the effects of the m SNPs,
- \mathbf{e} is a vector of random residuals.

Assume SNP effects come from normal distribution with same variance

$$\boldsymbol{\alpha} \sim N(0, \sigma_\alpha^2), \mathbf{e} \sim N(0, \sigma_e^2)$$

Linear mixed model

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



Best linear unbiased prediction

To estimate random effects (Henderson 1950 & Robinson 1991).

Best: minimum mean square error within class of linear predictors

Linear: random variables α are linear functions of the data \mathbf{y}

Unbiased: the average value of the estimate of α is equal to the average value of the quantity being estimated

Predictor: to distinguish random effects from fixed effect estimates

Genomic prediction with BLUP

Linear mixed model

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

BLUP solutions

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

$$\lambda = \sigma_e^2 / \sigma_\alpha^2$$

\mathbf{I} = identity matrix (dimensions $m \times m$)

Genomic prediction with BLUP

Example:

$$y_i = \mu + \sum_{j=1}^{10} X_{ij}\alpha_j + e_i$$

- A “simulated” data set
- Single chromosome, with 10 markers
- Phenotypes “simulated”
- overall mean of 1
- an effect for SNP 1 of 2 allele of 1
- normally distributed error term with mean 0 and variance 1.

Genomic prediction with BLUP

Example:

Animal	Y	X									
		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

- 10 SNPs
- Only 5 phenotypic records.

Genomic prediction with BLUP

Example:

		X									
Animal	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

Assume value of 1 for λ

$$\mathbf{1}_n' = [1 \ 1 \ 1 \ 1 \ 1]$$

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

Genomic prediction with BLUP

Example:

		X									
Animal	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

Assume value of 1 for λ

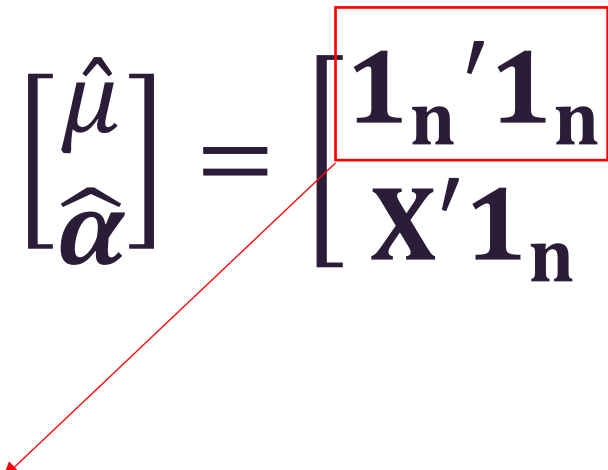
$$\mathbf{1}_n' = [1 \ 1 \ 1 \ 1 \ 1]$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 5 & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I} \lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

Genomic prediction with BLUP

Example:

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

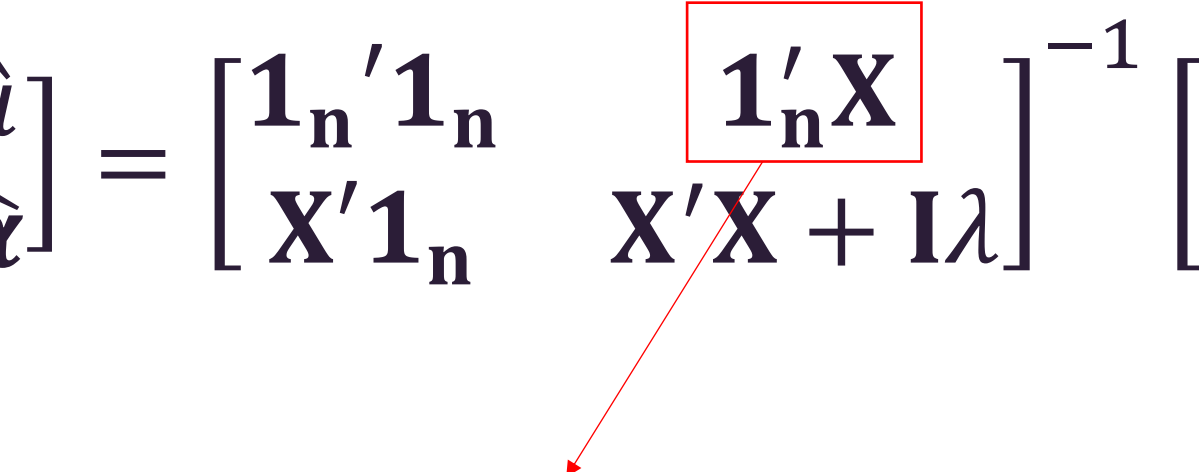


5	3	2	2	4	2	3	8	6	2	6		3.96
3	4	1	1	3	1	2	5	3	2	3		3.32
2	1	3	2	2	1	2	4	2	1	2		1.68
2	1	2	3	2	1	2	4	2	1	2		1.68
4	3	2	2	5	2	3	7	4	2	4		3.77
2	1	1	1	2	3	2	4	2	0	2		1.68
3	2	2	2	3	2	4	6	3	1	3		2.91
8	5	4	4	7	4	6	15	9	3	9		6.87
6	3	2	2	4	2	3	9	9	2	8		4.15
2	2	1	1	2	0	1	3	2	3	2		2.09
6	3	2	2	4	2	3	9	8	2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

Genomic prediction with BLUP

Example:

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



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

Genomic prediction with BLUP

Example:

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

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

Genomic prediction with BLUP

Example:

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

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

Genomic prediction with BLUP

Example:

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

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

Genomic prediction with BLUP

Example:

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

Genomic prediction with BLUP

Now we want to predict breeding value of a group of young animals without phenotypes (GEBV = genomic estimated breeding values).

$$\mathbf{GEBV} = \mathbf{X}\hat{\boldsymbol{\alpha}}$$

We have the $\hat{\boldsymbol{\alpha}}$, and we can get \mathbf{X} from their genotypes (after genotyping).....

Progeny	X										
	1	1	1	1	1	1	2	1	0	1	
	2	1	0	0	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

Best linear unbiased prediction

To estimate random effects (Henderson 1950 & Robinson 1991).

Best: minimum mean square error within class of linear predictors

Linear: random variables α are linear functions of the data y

Ignoring mean and other SNP

$$\hat{\alpha}_1 = \frac{x_1'y}{x_1'x_1 + \lambda}$$
$$= (0*0.19 + 1*1.23 + 1*0.86 + 1*1.23 + 0*0.45) / (3 + 1)$$

		X									
Animal	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

Genomic prediction with BLUP

An equivalent model (**GBLUP**)

If there are many QTLs whose effects are normally distributed with constant variance,

Then genomic selection equivalent to replacing the expected relationship matrix with the realised or genomic relationship matrix (**G**) estimated from DNA markers in normal BLUP equations.

- G_{ij} = proportion of genome that is IBD between individuals i and j

Genomic prediction with BLUP

An equivalent model

Rescale \mathbf{X} to account for allele frequencies

$$w_{ij} = \frac{(x_{ij} - 2p_i)}{\sqrt{2p_i(1 - p_i)}}$$

Then breeding values are $\mathbf{g} = \mathbf{W}\boldsymbol{\alpha}$ ($\mathbf{GEBV} = \mathbf{X}\hat{\boldsymbol{\alpha}}$)

And

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m}$$

Then

$$\text{Var}(\mathbf{g}) = \mathbf{G} \sigma_g^2$$

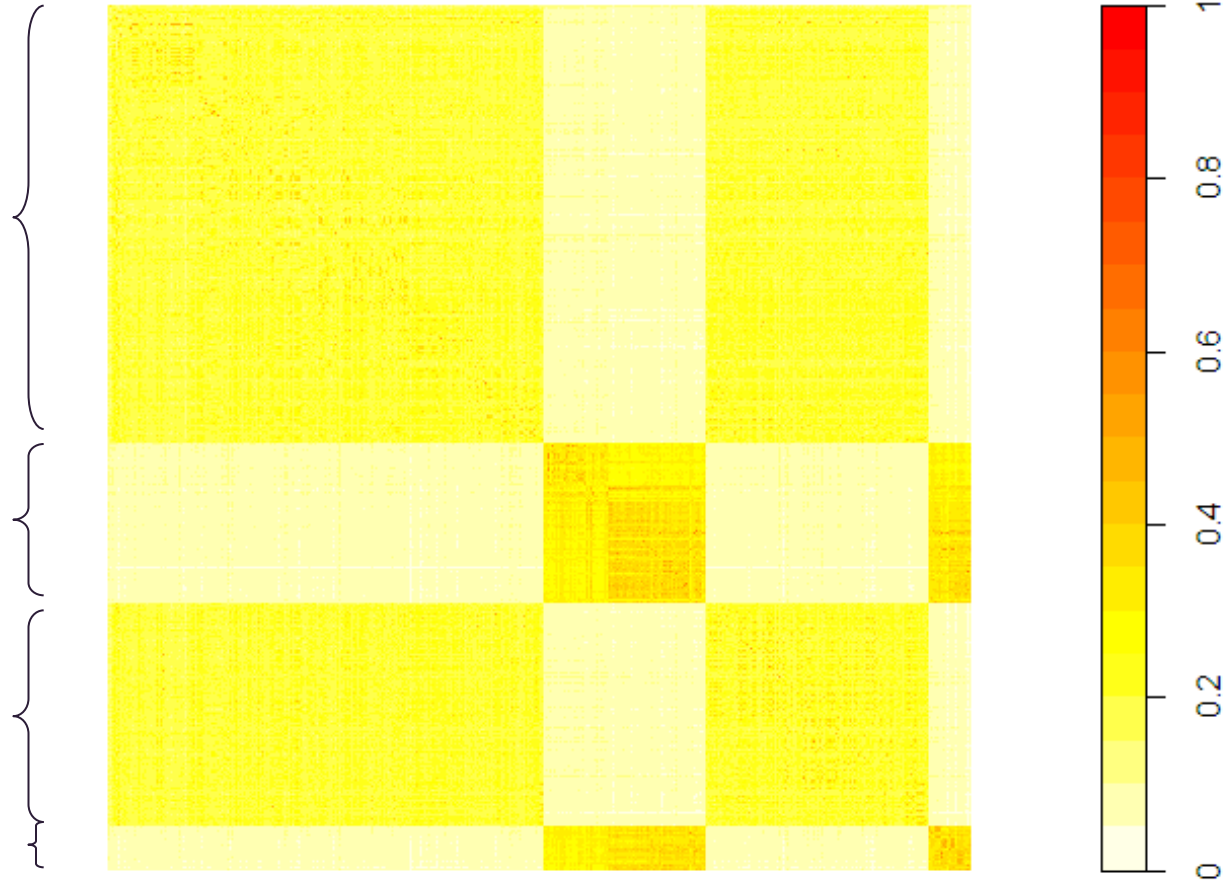
G matrix

Holstein reference $n = 781$

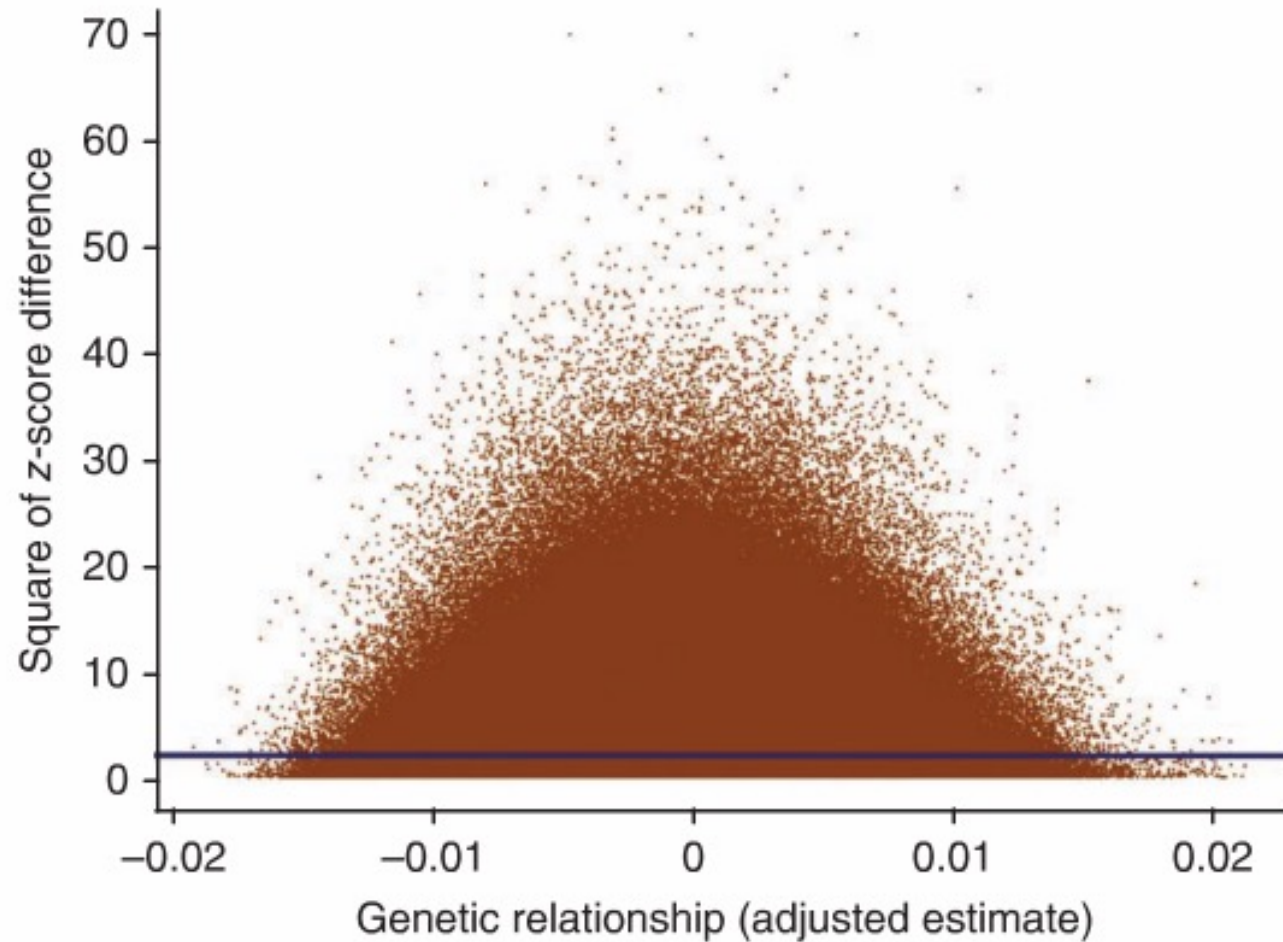
Jersey reference $n = 287$

Holstein validation $n = 400$

Jersey validation $n = 77$



G off-diagonals in unrelated human population



Genomic prediction with BLUP

An equivalent model

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z} \mathbf{g} + \mathbf{e}$$

where

$$\text{Var}(\mathbf{g}) = \mathbf{G} \sigma_g^2$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{Z} \\ \mathbf{Z}' \mathbf{1}_n & \mathbf{Z}' \mathbf{Z} + \mathbf{G}^{-1} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{Z}' \mathbf{y} \end{bmatrix}$$

Genomic prediction with BLUP

- Z matrix maps the phenotypic records onto the genetic values
- e.g. 5 individuals with the first 3 having records

Z

1	0	0	0	0
0	1	0	0	0
0	0	1	0	0

Genomic prediction with BLUP

Model 1 - SNP-BLUP

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

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{X} \\ \mathbf{X}' \mathbf{1}_n & \mathbf{X}' \mathbf{X} + \mathbf{I} \frac{\sigma_e^2}{\sigma_\alpha^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{X}' \mathbf{y} \end{bmatrix}$$

$$\text{GEBV} = \mathbf{X}\hat{\alpha}$$

Model 2 - GBLUP

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}$$

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n' \mathbf{1}_n & \mathbf{1}_n' \mathbf{Z} \\ \mathbf{Z}' \mathbf{1}_n & \mathbf{Z}' \mathbf{Z} + \mathbf{G}^{-1} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1}_n' \mathbf{y} \\ \mathbf{Z}' \mathbf{y} \end{bmatrix}$$

Genomic prediction with BLUP

An equivalent model

Why use model 2 (GBLUP)?

- If number of markers $\gg \gg$ large than number of animals, more computationally efficient
- Calculate accuracy of GEBV from inverse coefficient matrix (amount of data in estimate!)

– Prediction error variance $PEV_i = C^{ii} \sigma_e^2$

– Accuracy $r_i = \sqrt{1 - PEV_i / \sigma_g^2}$

$$\mathbf{C} = \begin{bmatrix} \mathbf{1}'_n \mathbf{1}_n & \mathbf{1}'_n \mathbf{Z} \\ \mathbf{Z}' \mathbf{1}_n & \mathbf{Z}' \mathbf{Z} + \mathbf{G}^{-1} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}^{-1}$$

Very useful – can calculate how well we predict for individuals without their own phenotype (e.g., young calves, people)

Genomic prediction with BLUP

Moving from GBLUP to SNP-BLUP

called Backsolving for SNP effects

$$\hat{\alpha} = \mathbf{X}' \mathbf{G}^{-1} \hat{\mathbf{g}} / m$$

Can use in alternative form of GWAS