Genomic Selection in Animal Breeding

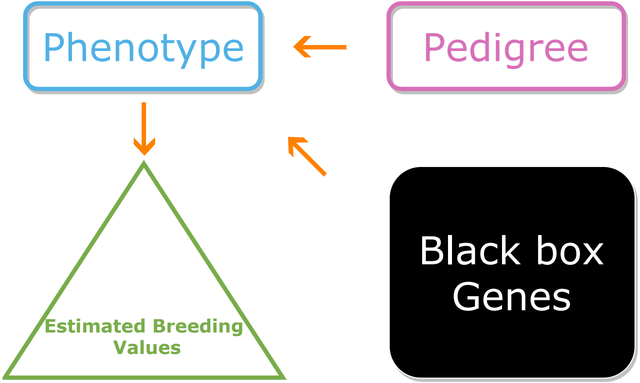
Hongding Gao

Natural Resources Institute Finland (LUKE)

Aug. 23rd, 2024

Traditional genetic prediction

- Successful but Limited



Recap: Animal model

- Pedigree BLUP (PBLUP)

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Playing Lego

$$\left[\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{array}\right] \left[\begin{array}{c} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{array}\right]$$

Limitations

- Slow genetic progress for low heritability traits (reproduction, health traits)
- Young animals (no own records, no offspring)
- Long generation intervals (e.g. dairy cattle 5 yr)
- High cost for progeny testing

QTL

- QTL: Quantitative Trait Locus
- QTL: A locus/region of DNA which is correlated with variation of a quantitative trait in the phenotypes
- QTL mapping: identify the position of genes or markers that influence the trait

Genetic markers

- A fragment of DNA that is associated with a certain location within the genome
- Many types: Microsatellite, RFLP, AFLP, Single nucleotide polymorphism (SNP), . . .

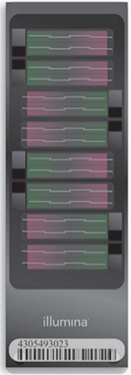
Revolution of genotyping technology

- Highly dense molecular markers covering the whole genome
  - Single nucleotide polymorphisms (SNPs)
  - High throughput genotyping technology (SNP chips in 2007)
  - Available for most livestock species
  - Genotyping costs continue to decrease

Various SNP chips

BovineLD
v2.0 (3k)

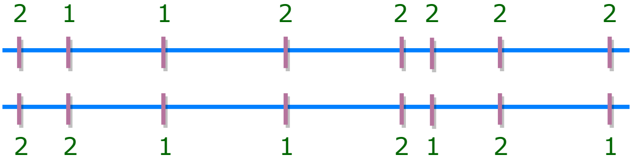
BovineSNP50
v3 (50k)



BovineHD
(777k)

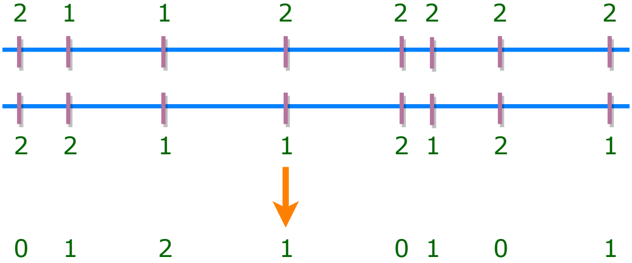
SNP information

- SNPs have 2 alleles (diallelic)



Recode SNP data

- Minor allele frequency (MAF): the frequency of less frequent allele at each locus in the population
- SNPs are commonly coded based on the counts of minor allele
- 0, 2 homozygote; 1 heterozygote
- Assume 1 is the minor allele for the example below



One more example

- Assume lower case letter is the minor allele in that locus

	Locus 1	Locus 2	Locus 3	Locus 4
Ani1	Ag	AA	GG	Ct
Ani2	AA	AA	Ga	Ct
Ani3	AA	tt	GG	CC

One more example

	Locus 1	Locus 2	Locus 3	Locus 4
Ani1	1	0	0	1
Ani2	0	0	1	1
Ani3	0	2	0	0

SNP data

```
10001112200200121110111121111011110011211000201220022220111
1202101200211122110021112001111001011011010220011002201101
12002011010202221211221020100111000112202212222112021120120
2010020220200002110001120201122111211102201111000021220200
0221012020002211220111012100111211102112110020102100022000
2201000201100002202211022112101121110122220012112122200200
0200202020122211002222222002212111121002111120011011101120
0202220001112011010211121211102022100211201211001111102111
2110211122000101101110202200221110102011121111011202102102
1211011022122001211011211012022011002220021002110001110021
1021101110002220020221212110002220102002222121221121112002
0110202001222222112212021211210110012110110200220002001002
0001111011001211021212111201010121202210101011111021102112
2111111212111210110120011111021111011111220121012121101022
202021211222120222002121210121210201100111222121101
```

Genomic selection

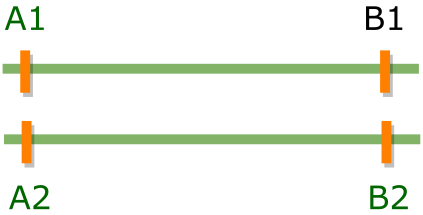
- First proposed by Meuwissen et al., 2001
- All SNPs fitted simultaneously
- Innovating & Boosting the breeding cycle in animal and plant

Linkage Disequilibrium (LD)

- Alleles present at the two loci are not independent

Linkage Disequilibrium (LD)

- A1 is more often associated with B1 than A2 with B1
- $freq(A1B1) \neq freq(A1)freq(B1)$



A1B1

A1B2

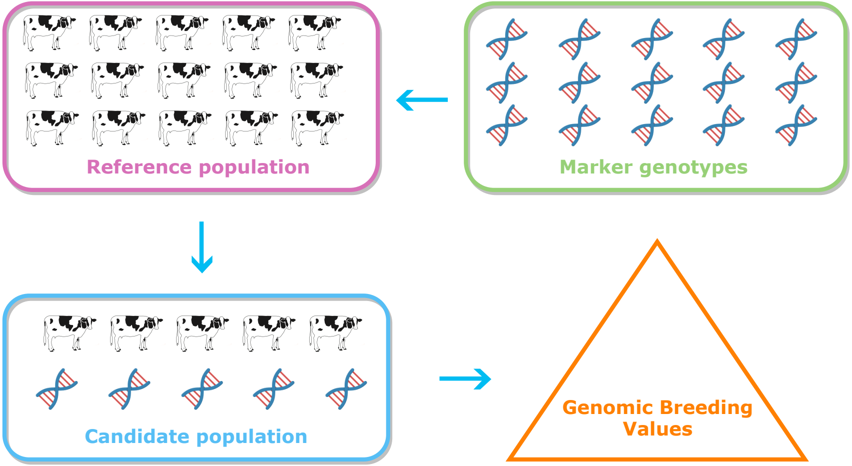
A2B1

A2B2

Linkage Disequilibrium (LD)

- Genomic selection requires LD
- SNPs are in LD with the QTLs across the whole population
- Assume all QTLs in the genome can be traced by SNPs
- SNP density must be sufficiently high to ensure that all QTLs are in LD with at least a marker

Genomic selection



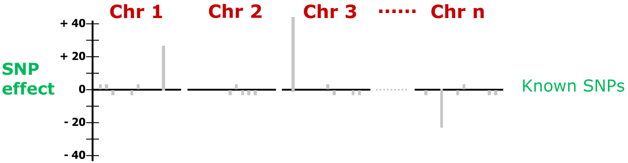
Genomic breeding values

- Sum of single marker effects

$$\mathbf{GEBV} = \sum_{i=1}^{m} \mathbf{M}_i \hat{\mathbf{g}}_i$$

where
m is the number of markers
$\mathbf{M}_i$ is the SNP code at locus i
$\hat{\mathbf{g}}_i$ is SNP effect at locus i

Genomic breeding values



1 + 1 - 1 - 1 + 1 + 25 - 1 + 1 - 1 - 1 - 1 + 42 + 1 - 1 - 1 - 1 - 1 - 22 - 1 + 1 - 1 - 1 = +38

Benefits of genomic selection

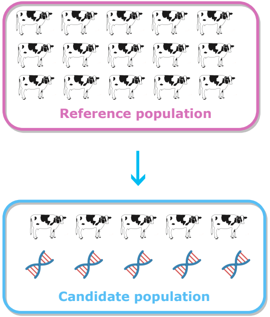
- Double the genetic progress (Schaeffer, 2006)
  - Increase accuracy of selection
  - Reduce generation interval
- Reduce costs in the breeding plan by 92% (Schaeffer, 2006)
  - Reduce the scale for progeny testing (dairy cattle)

Benefits of genomic prediction

- Accuracy exceeds 0.8 and gains in reliability up to 48% in US Holsteins (Wiggans et al., 2011)
- Accuracy increases by 50% in pig breeding (Knol et al., 2016)
- Accuracy increases by 20%-50% in poultry breeding (Wolc et al., 2015; Wang et al., 2013)
- Promising in plant breeding, e.g. genetic gain 1.4 to 2.7 times higher than phenotypic selection in wheat (Battenfield et al., 2016)

Factors affecting accuracy of GP

- Prediction models
- Size of reference population
- Heritability of trait
- Marker density
- Population structure
- Variance components
- ...



Reference population

Candidate population

Two equivalent models

- SNP-BLUP: marker based model
  - Estimated marker effects
  - Better when no. of animals $>$ no. of markers
- GBLUP: breeding value based model
  - Estimated breeding values directly
  - Better when no. of animals $<$ no. of markers

Response variables

- Deregressed proofs (DRP)
  - Derived from EBV which is a regressed variable
- Daughter yield deviations (DYD)
  - Average of the daughter's actual performance adjusted for fixed and non-genetic random effects and genetic effects of the daughters' dams
- Estimated breeding values (EBV)

SNP-BLUP

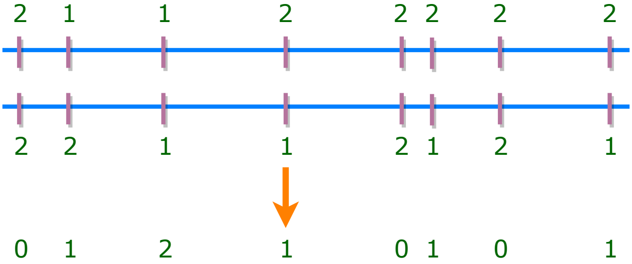
- We assume markers can explain all the genetic variance
- We assume marker effects are identically and independently distributed

$$\mathbf{y} = \mathbf{Xb} + \sum_{i}^{m} \mathbf{M}_i \mathbf{g}_i + \mathbf{e}$$

where
M is a n (no. of animals) by m (no. of markers) matrix
g is a vector of marker effects
e is a vector of random residuals

SNP coding

- **M** is a marker covariates matrix containing SNP codes
- SNPs are commonly coded based on counts of minor allele
- 0, 2 homozygote; 1 heterozygote



Centering

- To set the mean value of allele effects equal to 0

$$\mathbf{Z} = \mathbf{M} - \mathbf{P}$$

where

$\mathbf{P}$ is a matrix with column j equal to $2p_j$

p_j is the allele frequency of the second allele at locus j

SNP-BLUP

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

$$\mathbf{g} \;\sim\; N(\mathbf{0}, \mathbf{I}\sigma_g^2) \quad \mathbf{e} \;\sim\; N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

where
σ_g^2 is the SNP variance for each SNP
σ_e^2 is the residual variance

MME for SNP-BLUP

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (\mathbf{I}\sigma_g^2)^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where

$$\alpha = \sigma_e^2/\sigma_g^2$$

$$\sigma_g^2 = \frac{\sigma_a^2}{\sum_{i=1}^{m} 2p_i q_i}$$

Direct genomic value (DGV)

- An "old" terminology for GEBV
- Means GEBV calculated only by marker information

$$\mathbf{DGV} = \mathbf{a} = \mathbf{Z}\hat{\mathbf{g}}$$

SNP-BLUP with polygenic effect

- We assume markers cannot explain all the genetic variance
- We assume marker effects are identically and independently distributed

$$\mathbf{y} = \mathbf{Xb} + \sum_{i}^{m} \mathbf{M}_i \mathbf{g}_i + \mathbf{Wu} + \mathbf{e}$$

where
M is a n (no. of animals) by m (no. of markers) matrix
g is a vector of marker effects
u is a vector of polygenic effects
W is a design matrix linking records to animals

SNP-BLUP with polygenic effect

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{i}^{m} \mathbf{M}_i \mathbf{g}_i + \mathbf{W}\mathbf{u} + \mathbf{e}$$

where

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}\sigma_g^2) \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$$

Let ω equal to the proportion of total genetic variance cannot be explained by markers, then

$$\sigma_u^2 = \omega\sigma_a^2 \quad \sigma_g^2 = \frac{(1-\omega)\sigma_a^2}{\sum_{i=1}^{m} 2p_i q_i}$$

PBLUP ⇒ GBLUP

- You can understand **GBLUP** is a "improved" version of traditional **BLUP**

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z'y} \end{bmatrix}$$