Genomic Prediction

Modern Statistical Approaches for Biological Data
29.09.2020

A/Prof Hans Daetwyler and Dr Zibei Lin
Hans.Daetwyler@agriculture.vic.gov.au

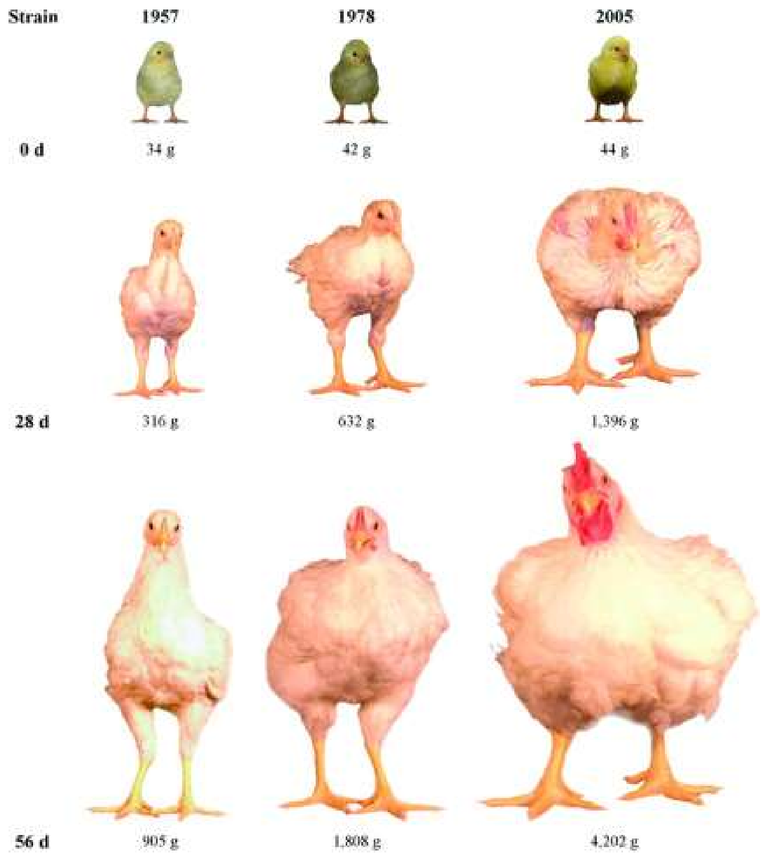
(some slides adapted from course by Prof Ben Hayes, University of Queensland)

Outline

- Introduction to Genomic Prediction
- Methods
- Validation principles
- Limitations
- Examples

Quantitative Traits and Selection

- Dramatic changes in phenotypes due to selection
- Many traits affected by large number of mutations
  - Quantitative trait loci (QTL)
- Variance explained by individual markers will be small
- Genomic prediction -> Use large numbers of DNA markers to simultaneously track all QTL
- Increase efficiency of selection



Unselected Commercial

Zuidhof et al., 2014. Poultry Sci 93:2970-2982

Methods to ‘Genetically’ Evaluate Individuals

- Phenotypic Selection
  - Low tech
  - Simple to implement
  - Works best when heritability is higher
  - Must observe phenotype
  - Still widely used in plant breeding

$y = Xb + Zu + e$
- $V(u)=I$
- Individuals are assumed independent

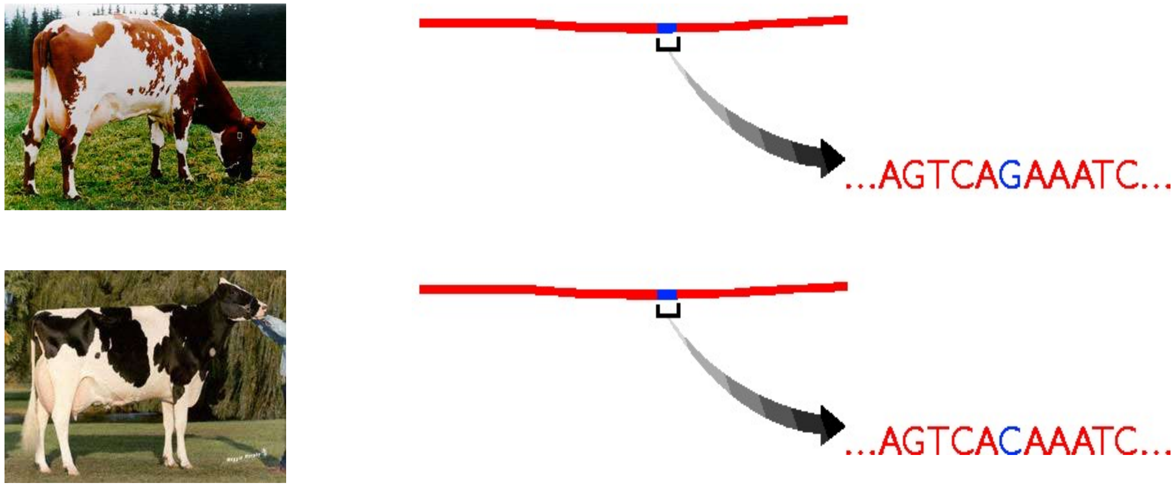
- Pedigree Breeding
  - Can predict performance based on relatives
    - Juvenile = parent average
    - Requires pedigree recording
  - Observed phenotypes increase accuracy
  - Info on Mendelian sampling term from own records and progeny
  - Efficiency less dependent on h^2 than phenotypic selection
  - More inbreeding than phenotypic selection at low h^2 (BLUP)

$y = Xb + Zu + e$
- $V(u)=A$
- Covariance of lines from pedigree relationship matrix (A)

The Genetic Marker Revolution

- As a result of sequencing animal and plant genomes, have a huge amount of information on variation in the genome
  - at the DNA level

- Most abundant form of variation are Single Nucleotide Polymorphisms (SNPs)



The Genomic Revolution

- Genotyping solutions available for most species

- SNP arrays
  - Accurate genotypes at specific positions
- Genotyping by (re)-sequencing
  - Targeted and untargeted approaches
  - Not quite as accurate but more flexible than chips
- Cost?
  - ~ $15-100 AUD for 50,000+ markers

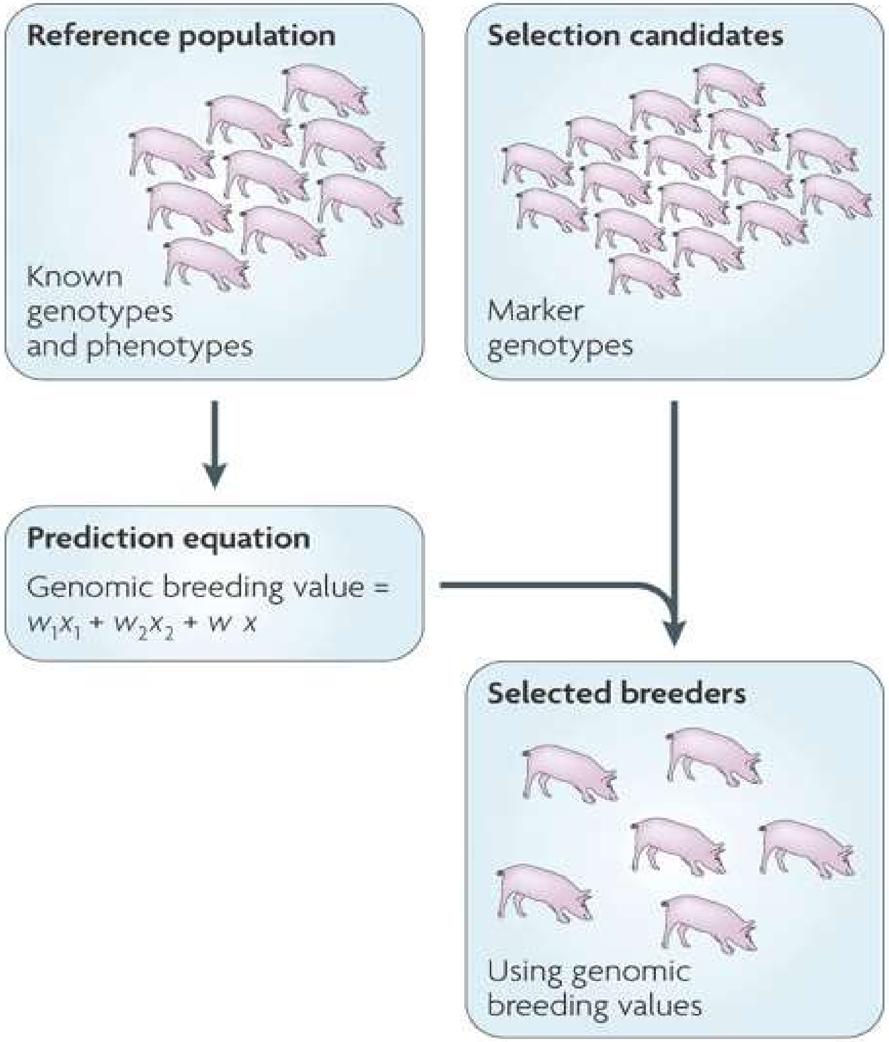
Methods to Genetically Evaluate Individuals

- Genomic Prediction
  - – Predict performance based on reference population (relatives?)
    - Predict young individuals with only genotypes
      - – Decrease generation interval
  - – Requires genotyping
  - – Observed phenotypes increase accuracy
  - – Info on Mendelian sampling term from all individuals in reference

$y = Xb + Zu + e$

- $V(u)=G$
- Covariance of lines from genomic relationship matrix (G)

Genomic Prediction



Reference population

Known genotypes and phenotypes

Selection candidates

Marker genotypes

Prediction equation

Genomic breeding value = $w_1x_1 + w_2x_2 + w\ x$

Selected breeders

Using genomic breeding values

Michael E. Goddard & Ben J. Hayes

Nature Reviews Genetics **10**, 381-391 (June 2009)

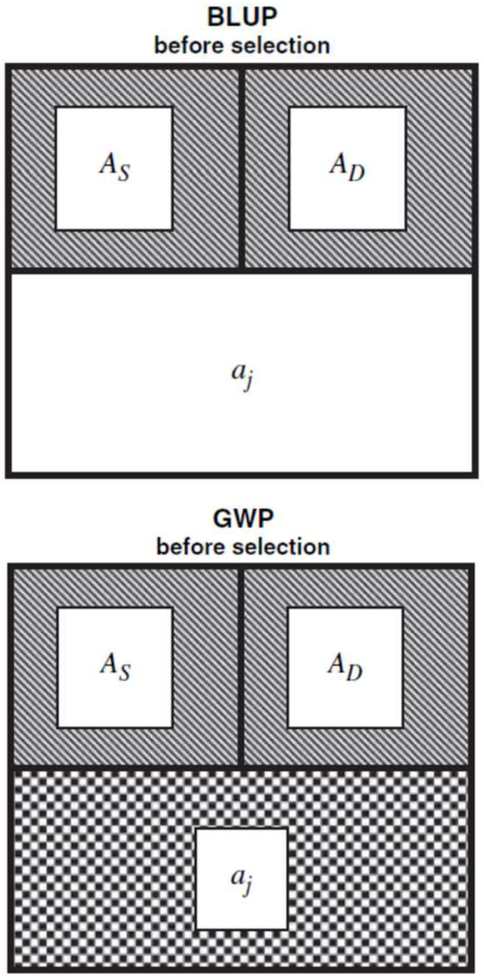
Why makes genomic prediction different to pedigree breeding?

The Mendelian Sampling Term

An individuals breeding value has two components

- 50% due to parent average component
  - Prediction at birth is the average of two parents breeding value

- 50% due to Mendelian sampling component
  - Individual’s deviation from parent average breeding value
  - Sampling of parental alleles
  - Reason for differences in:
    - a pair of full sibs
    - a pair of F2 in a bi-parental
  - Cannot predict at birth/seed using pedigree alone
  - Genetic gain driven by
    - Accuracy of and time taken to estimate of Mendelian sampling term
  - Genomic prediction (GWP) provides information on which alleles received from parents



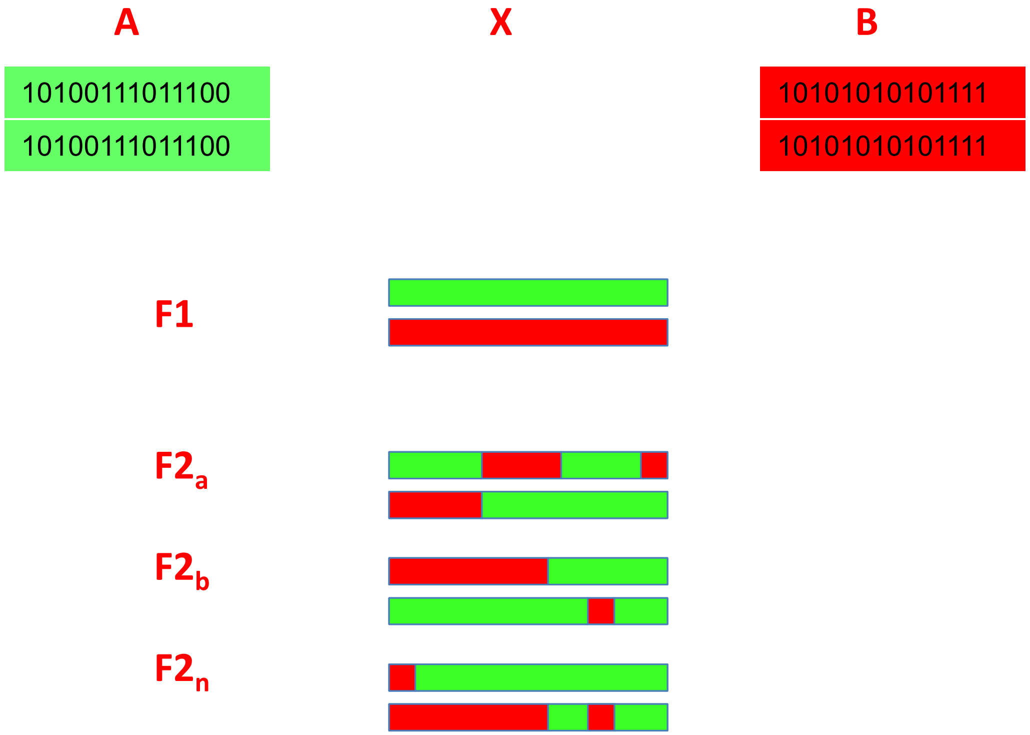
BLUP
before selection

GWP
before selection

Daetwyler et al., 2007. J Anim Breed Genet 124: 369-376

What Mendelian sampling looks like in a inbred bi-parental cross

- Diploid genetics
  - Each individual has two gametes
  - If individual is inbred these gametes are the same



A X B

10100111011100
10100111011100

10101010101111
10101010101111

F1

F2$_a$

F2$_b$

F2$_n$

Factors affecting genomic prediction accuracy

$$r = \sqrt{\frac{N_P h^2}{N_P \ h^2 + M_e}}$$

- Reference population size (Np)
- Heritability (h^2)
- Number of effective chromosome segments (Me)
  - Effective population size
    - Linkage disequilibrium
  - Genome length

- Number of QTL (if few)

- Dense genetic markers

Daetwyler et al., 2010. Genetics 185:1021-1031

Genomic Prediction

- Genomic selection exploits linkage disequilibrium
  - Assumption is that markers are correlated with mutations (QTL) and have same effect across whole population

- Justified assumption as we now have dense marker maps

- Trace whole genome with markers
  - Capture all mutations = all genetic variance

- Genomic selection avoids bias in estimation of effects due to multiple testing, as all effects fitted simultaneously

Genomic Prediction Methods

- Mixed linear models
  - Often referred to as best linear unbiased prediction (BLUP) methods
  - Two equivalent models: SNP BLUP and GBLUP (Habier et al., 2007. Genetics 177:2389-2397)

- Bayesian models
  - More flexible assumption on marker variances than BLUP
  - Utilise Gibbs sampling

De los Campos et al., 2013. Genetics. 193: 327-345

Genomic prediction with BLUP

- SNP BLUP model

$$\mathbf{y} = \mu\mathbf{1_n} + \sum_{i=1}^{p}\mathbf{X_i g_i} + \mathbf{e}$$

$$\begin{bmatrix}\hat{\mu}\\ \hat{\mathbf{g}}\end{bmatrix} = \begin{bmatrix}\mathbf{1_n{}'1_n} & \mathbf{1_n{}'X}\\ \mathbf{X'1_n} & \mathbf{X'X}+\mathbf{I}\dfrac{\sigma_e^2}{\sigma_g^2}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{1_n{}'y}\\ \mathbf{X'y}\end{bmatrix}$$

$$\mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

- GBLUP model

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

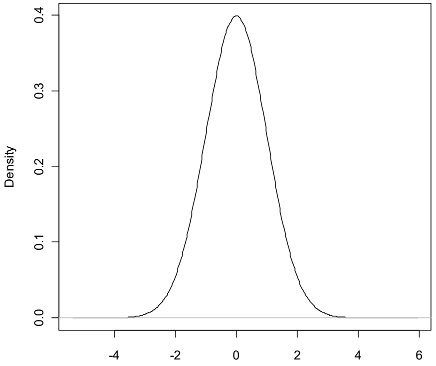
$$\begin{bmatrix}\hat{\mu}\\ \hat{\mathbf{v}}\end{bmatrix} = \begin{bmatrix}\mathbf{1_n{}'1_n} & \mathbf{1_v{}'Z}\\ \mathbf{Z'1_n} & \mathbf{Z'Z}+\mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_v^2}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{1_n{}'y}\\ \mathbf{Z'y}\end{bmatrix}$$

SNP BLUP

- BLUP = best linear unbiased prediction (SNP-BLUP)
- Model:

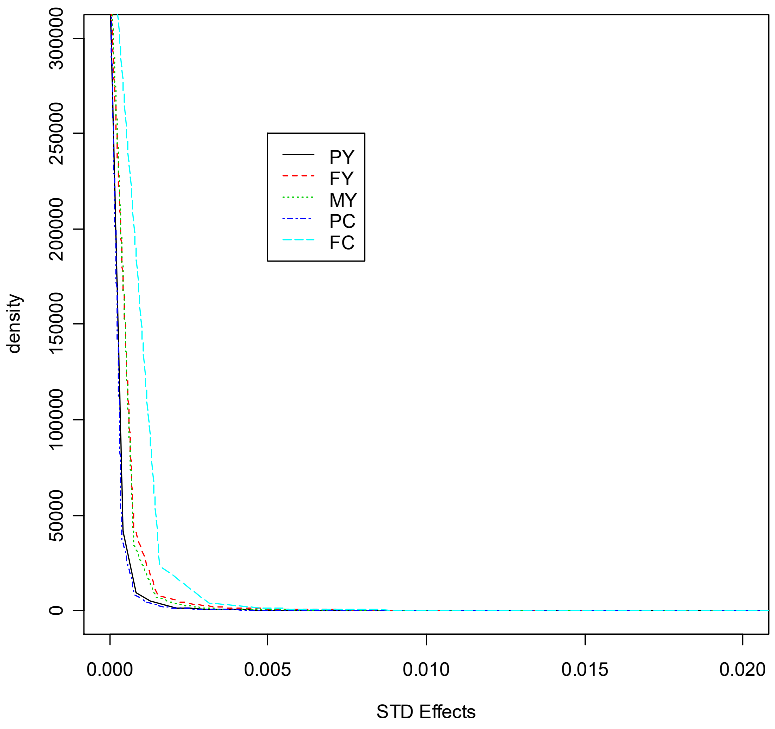
$$\mathbf{y} = \mu \mathbf{1_n} + \sum_{i=1}^{p} \mathbf{X_i g_i} + \mathbf{e}$$

- In BLUP we assume all SNP effects come from normal distribution with same variance
  - E(**g**) ~ N(0,σ_g^2)



Alternative prior assumptions for SNP effects

- BLUP assumes normally distributed QTL effects
- Does not match prior knowledge of distributions of QTL effects for some traits



Alternative prior assumptions for SNP effects

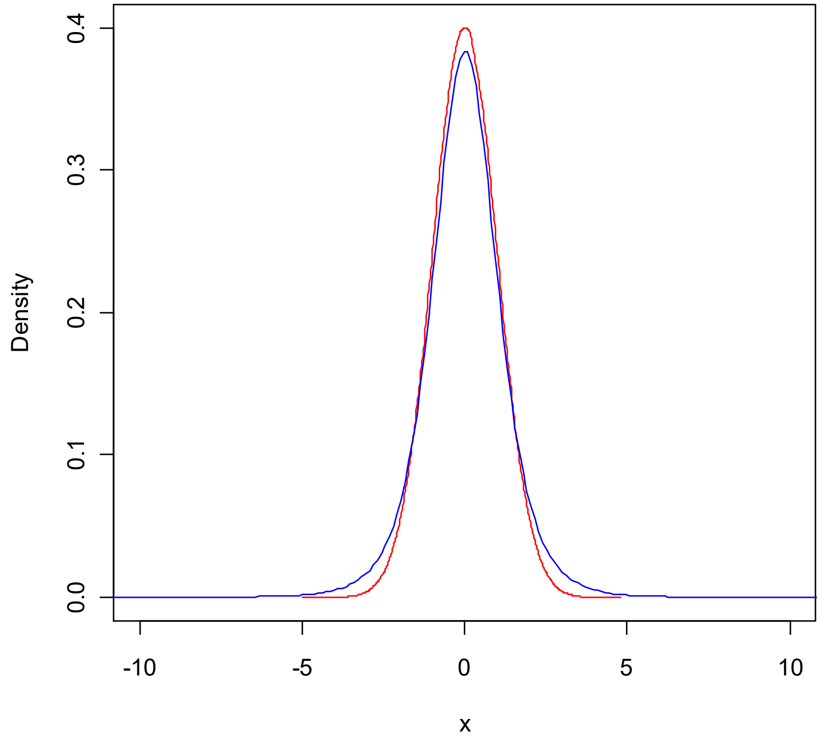
- Students t distribution?
  - BayesA
- Many zero effects and a proportion Students t distribution?
  - BayesB
- Many zero effect and rest normal distribution
  - BayesCpi
- Double exponential effects
  - BayesianLASSO
- Multiple normal distributions
  - BayesMulti, BayesR

De los Campos et al., 2013. Genetics. 193: 327-345

Bayesian Methods

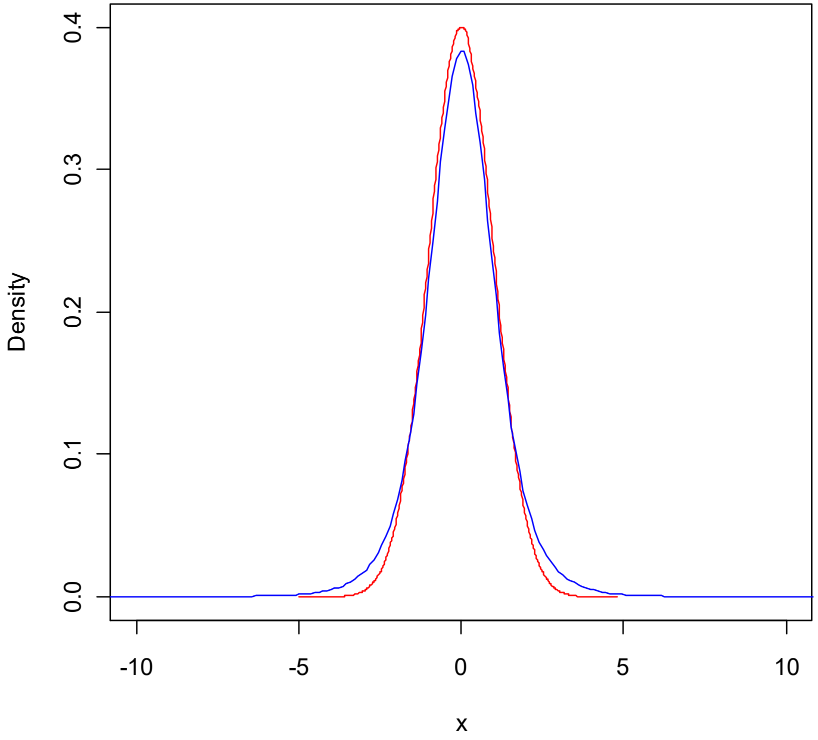
- For some traits prior knowledge suggests t-distribution of effects

- How to incorporate this into our predictions?



Bayesian Methods

- The **t distribution** can be presented as a two level hierarchical model
- Allow different variances for markers
- Assume a distribution of these variances
- Computationally easier to deal with than original form

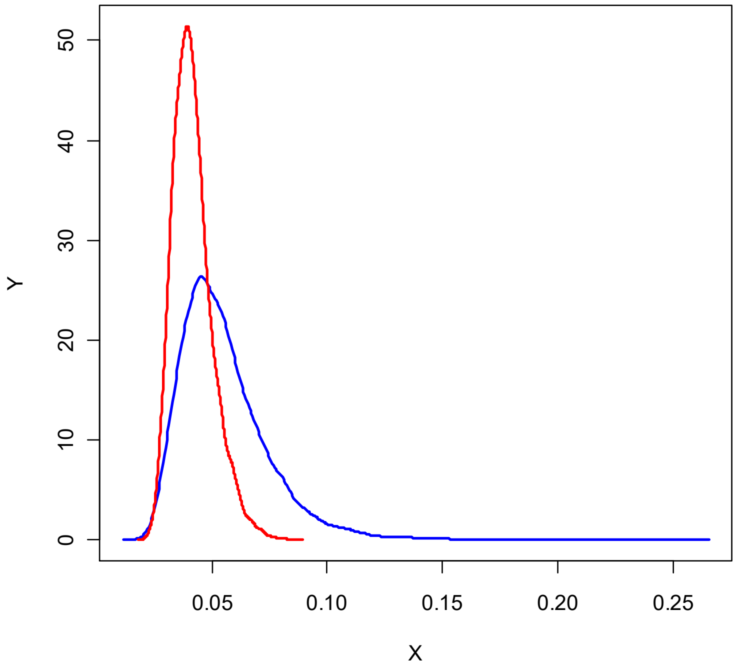


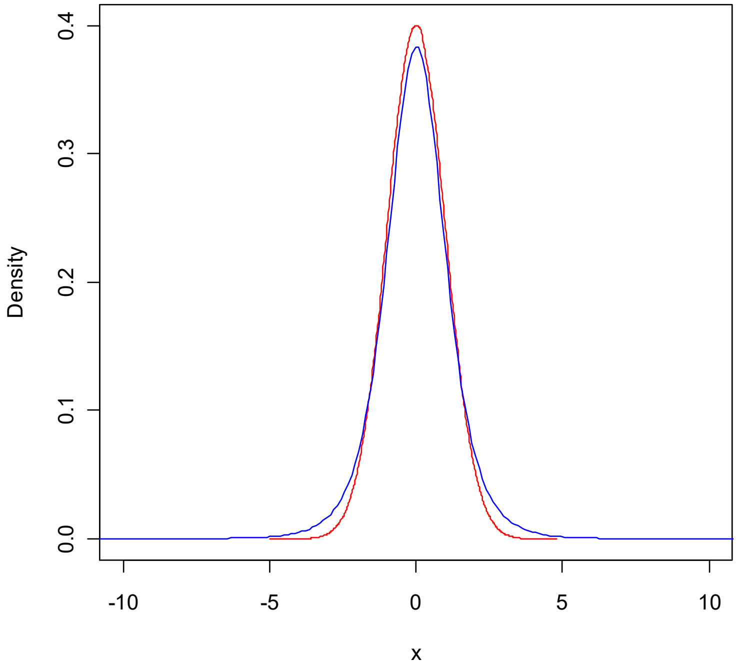
Bayesian methods

- Now lets allow different variances of marker effects

$$\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}}_1 \\ . \\ \hat{\mathbf{g}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{1_n'1_n} & \mathbf{1_n'X_1} & . & \mathbf{1_n'X_p} \\ \mathbf{X_1'1_n} & \mathbf{X_1'X_1} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{g1}^2} & . & \mathbf{X_1'X_p} \\ . & . & . & . \\ \mathbf{X_p'1_n} & \mathbf{X_p'X_1} & . & \mathbf{X_p'X_p} + \mathbf{I}\dfrac{\sigma_e^2}{\sigma_{gp}^2} \end{bmatrix}^{\mathbf{-1}} \begin{bmatrix} 1_n'y \\ X_1'y \\ . \\ X_p'y \end{bmatrix} \qquad \mathbf{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

Distribution of σ_{gj}^2 → Distribution of g_j





Bayesian methods

- Now lets allow different variances of marker effects
- Two levels of models
  - Data

$$P(\mathbf{g},\mu \mid y) \propto P(y \mid \mathbf{g},\mu)P(\mathbf{g},\mu)$$

  - Variances of marker effects

$$P(\sigma_{gi}^{2} \mid g_{i}) \propto P(g_{i} \mid \sigma_{gi}^{2})P(\sigma_{gi}^{2})$$

Bayesian methods – A word on priors

- Bayesian methods utilise priors
- A prior reflects the existing knowledge about the parameter to be estimated
- Priors affect results
  - The stronger the prior, the more the influence

Bayesian methods

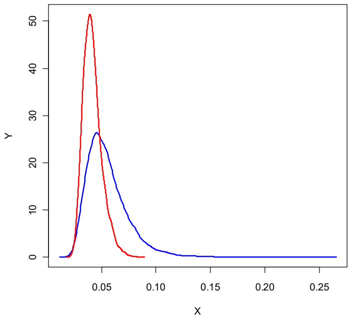
- Variances of chromosome segments

$$P(\sigma_{gi}^2 \mid g_i) \propto P(g_i \mid \sigma_{gi}^2)P(\sigma_{gi}^2)$$

- Prior?

$$S^2 / \chi_v^2$$

- We can choose *v* (degrees of freedom) and *S*^{2} (scale factor) so that the prior reflects our knowledge that there are many QTL of small effect and few of large effect



Meuwissen et al., 2001. Genetics. 157: 1819-1829

Bayesian methods

- Variances of chromosome segments

$$P(\sigma_{gi}^{2} \mid \mathbf{g_i}) \propto P(\mathbf{g_i} \mid \sigma_{gi}^{2})P(\sigma_{gi}^{2})$$

- Posterior?

$$\chi^{-2}_{(4.012+n_i,0.002+\mathbf{g_i}'\mathbf{g_i})}$$

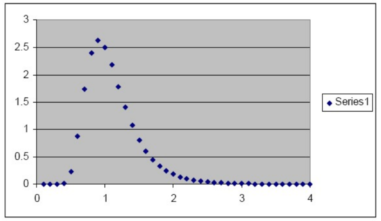
- But posterior cannot be estimated directly, dependent on $\mathbf{g_i}$!

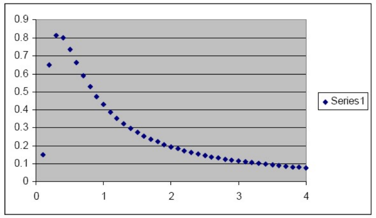
Bayesian methods

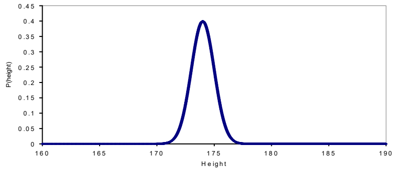
- Solution is to use Gibbs sampling
  - Draw samples from the posterior distributions of parameters conditional on all other effects
  - The average of these samples can be used as the estimates of the parameters

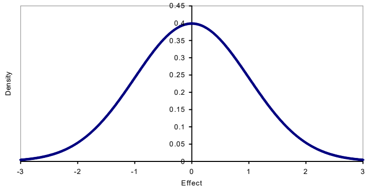
Bayesian methods

- Gibbs sampling scheme
  - Parameters to estimate and their posteriors
  - P(σ_{gi}^2|g_i) $\quad \chi^{-2}_{(4.012+n_i\,,0.002+\mathbf{g_i}\,'\mathbf{g_i}\,)}$
  - P(σ_e^2|**e**) $\quad \chi^{-2}_{(n-2,\mathbf{e'e})}$
  - P(μ|**y**,**e**,**g**,σ_e^2) $\quad N\left(\frac{1}{n}\left(\mathbf{1_n^{'}y}-\mathbf{1_n^{'}Xg}\right),\sigma_e^2/n\right)$
  - P(g_{ij}|**y**,μ,**g**≠ij,σ_{gi}^2,σ_e^2) $\quad N\left(\frac{\mathbf{X_{ij}^{'}y}-\mathbf{X_{ij}^{'}Xg_{(ij=0)}}-\mathbf{X_{ij}^{'}1_n}\mu}{\mathbf{X_{ij}^{'}X_{ij}}+\sigma_e^2/\sigma_{gi}^2},\sigma_e^2/\left(\mathbf{X_{ij}}'\mathbf{X}_{ij}+\sigma_e^2/\sigma_{gi}^2\right)\right)$





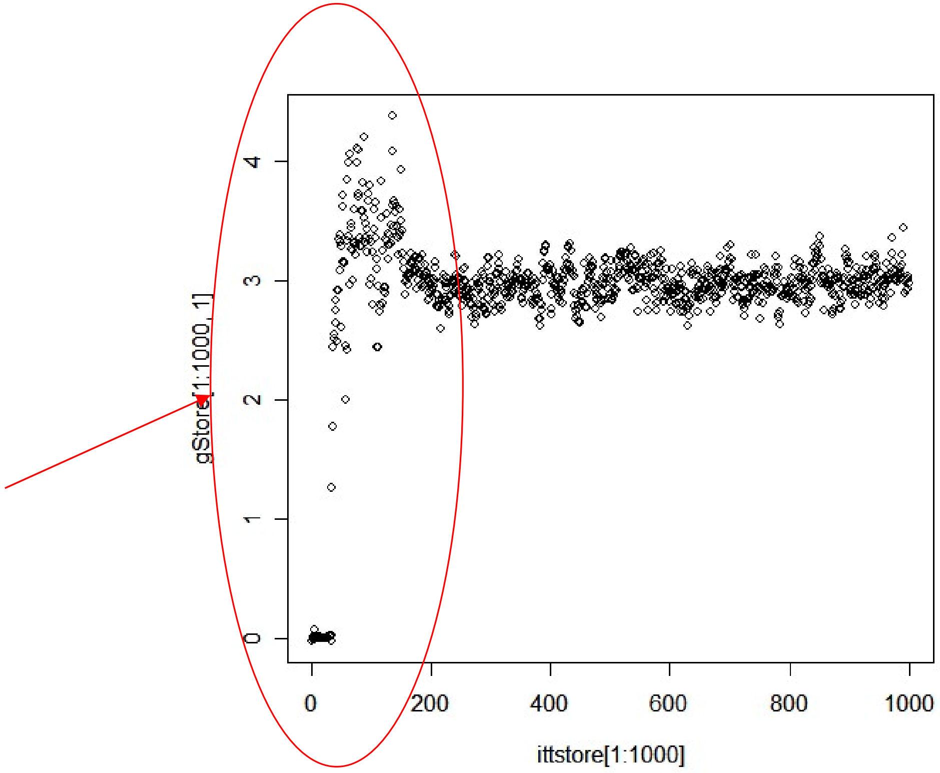




Bayesian methods

- Gibbs chain for 1000 cycles
  - $P(g_1|y,\mu,\mathbf{g}{\neq}1,\sigma_{g1}^2,\sigma_e^2)$

“Burn in”



Validation of genomic selection

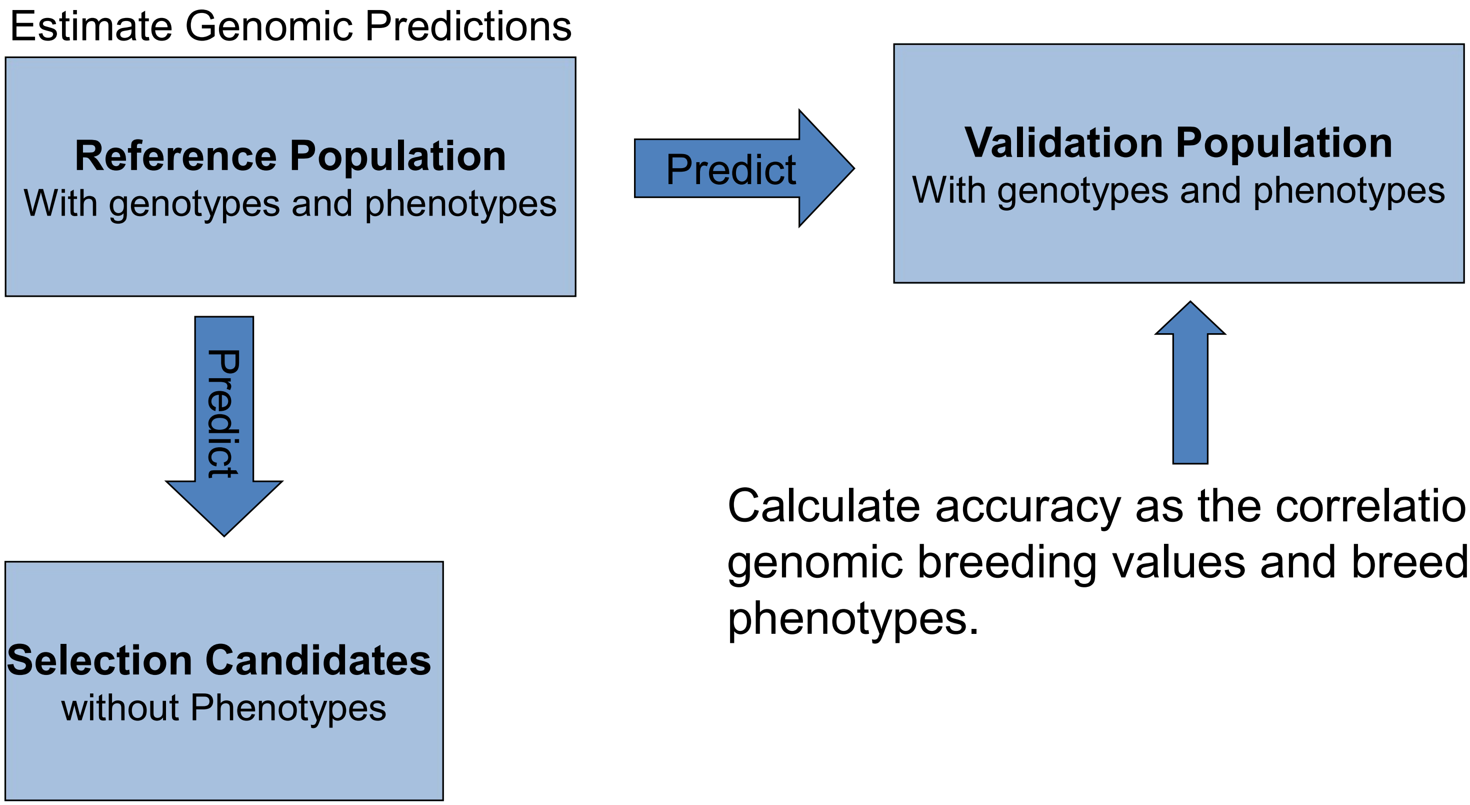
- Aim of genomic selection
  - predict (young) selection candidates without phenotypes

- How to test or validate predictions?

- Test predictions in a population sample that is similar to selection candidates

- Key principle of validation
  - Independence of reference and validation populations

Validation – Accuracy of genomic prediction



Estimate Genomic Predictions

Reference Population
With genotypes and phenotypes

Predict

Validation Population
With genotypes and phenotypes

Predict

Selection Candidates
without Phenotypes

Calculate accuracy as the correlation between genomic breeding values and breeding values or phenotypes.

Prediction Accuracy

- Most commonly used:
  - r = Pearson correlation(GEBV,phenotypes)
  - Gives accuracy of a group of individuals
  - Correlations have a standard error which depends on sample size and the magnitude of the correlation
    - An approximation of this standard error was given by Fisher (see Fisher z transform)
    - SE ~ 1/sqrt(N-3)
    - For example with 31 individuals
      - SE = 1/sqrt(31-3) = 0.189
- Individual accuracy
  - Calculated using the prediction error variance from the diagonal of the coefficient matrix (GBLUP)

Two main ways to validate

- 1st way: Independent set of individuals
  - Breeding values or phenotypes
  - Dairy bull progeny test (e.g. Daughter trait deviations)
  - Large progeny groups or many clones (plants)
  - Different population
  - Step 1: Estimate marker effects in reference population
  - Step 2: Predict highly accurate individuals and calculate accuracy

- 2nd way: ‘Classic’ cross-validation
  - Step 1: Divide dataset into n subsets of individuals
  - Step 2: Predict each subset using all other subsets
  - Step 3: Calculate accuracy in each subset and take mean across all subsets

Validation - Independence

- Always ask yourself this question:
  - If the validation individuals were selection candidates what data would be available?
  - Then only use that data for reference!

- Independence of ‘data’, not independence in relationship

Independence

- Validation individuals are not used in the reference population

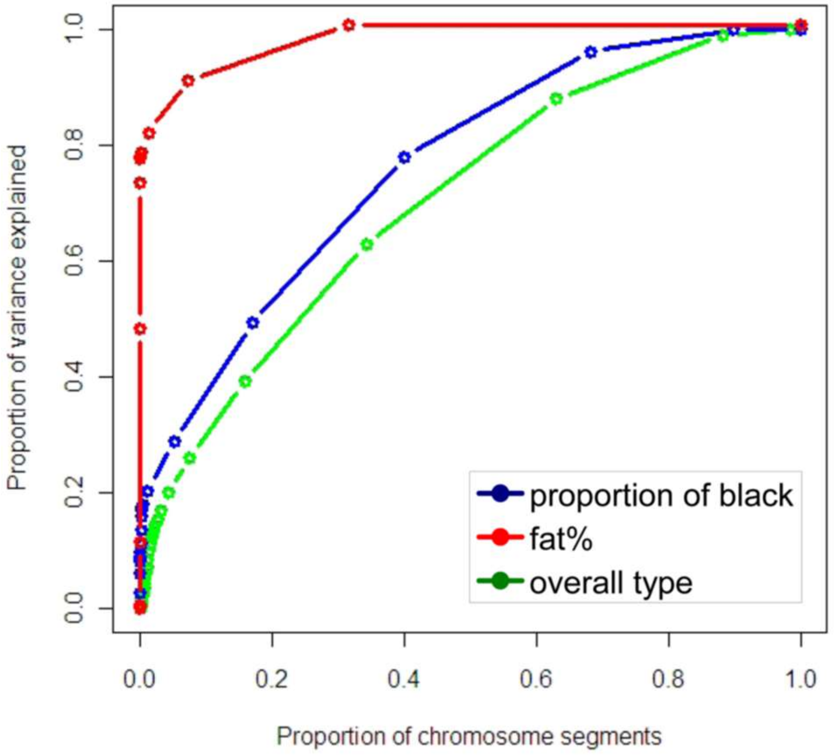
- Validation phenotypes do not contribute to observed variables of reference pop
  - E.g. excluded when calculating estimated breeding values

- Validation individuals do not have contemporaries of same age in reference

Target of prediction

- Validation population should be similar to selection candidates
- Similar relationship to reference as selection candidates
  - Same number of generations removed
  - Same breeds
  - Same population
- Same SNP density
  - Consider imputation error

Cattle: Performance of genomic prediction methods



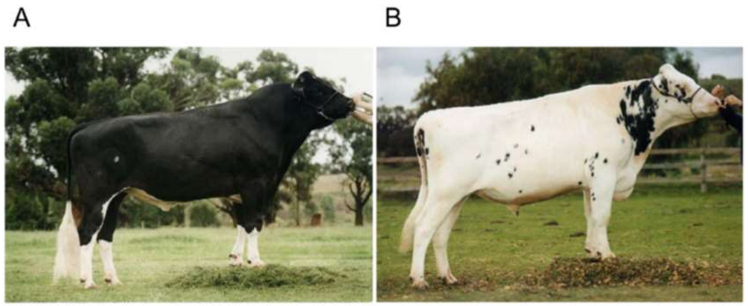


Figure 1. Proportion of black phenotype. Bull with 95% black (A) and bull with 5% black (B).
doi:10.1371/journal.pgen.1001139.g001

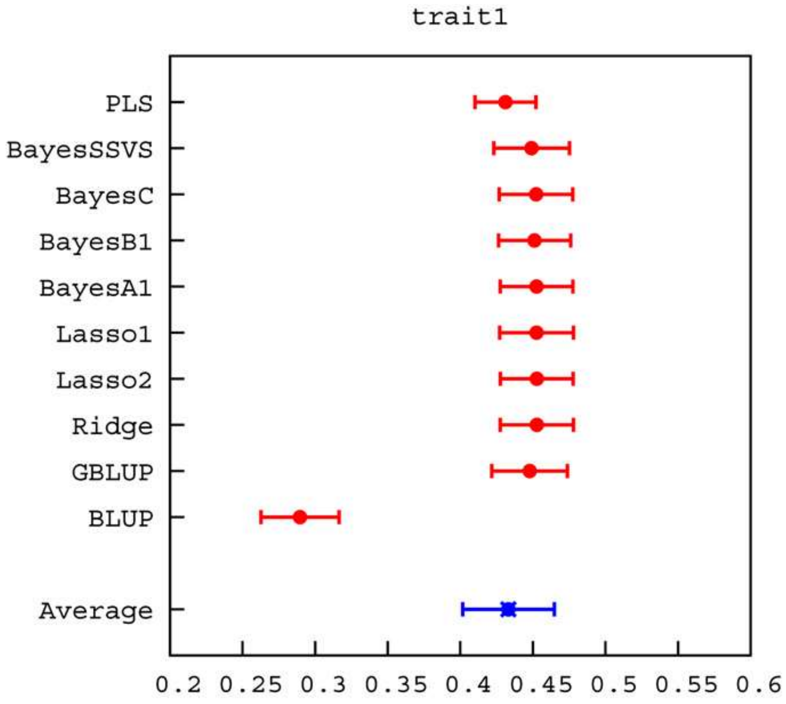
- 1200 Australian Holstein bulls

In traits with large QTL effects BayesA performed better than GBLUP

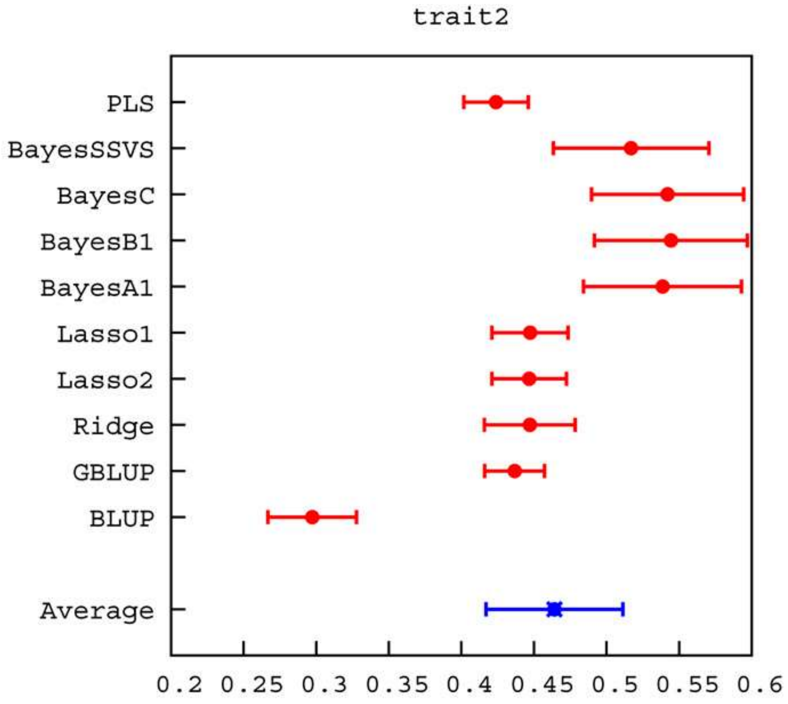
	Overall Type	Proportion Black Coat	Milk Fat %
GBLUP	0.42	0.46	0.63
BayesA	0.38	0.59	0.73

Hayes et al., 2010. PLoS Genetics 6: e1001139

Performance of genomic prediction methods



Many Mutations



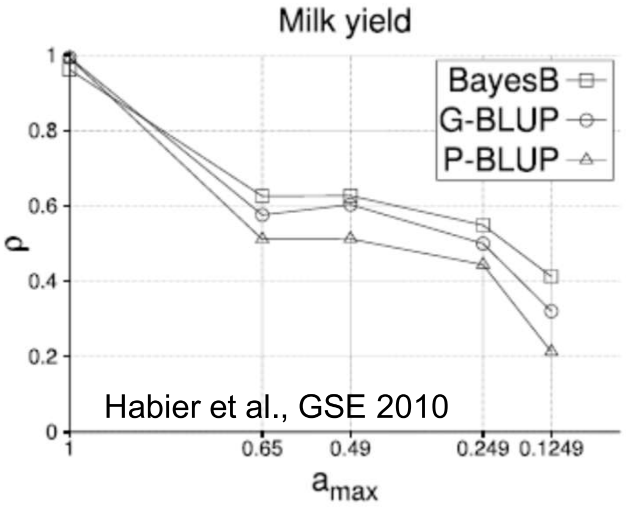
Fewer Mutations

- Many mutations, most methods perform the same
- Fewer mutations, methods that can differentially shrink marker effects are better

Daetwyler et al., 2013. Genetics. 193:347-365

Limitations of genomic selection

- Accuracy strongly related to relationship to reference population
  - Accuracy decreases as relationship decreases
  - Decay across generations
  - Lower accuracy across breeds
  - Low accuracy into novel germplasm

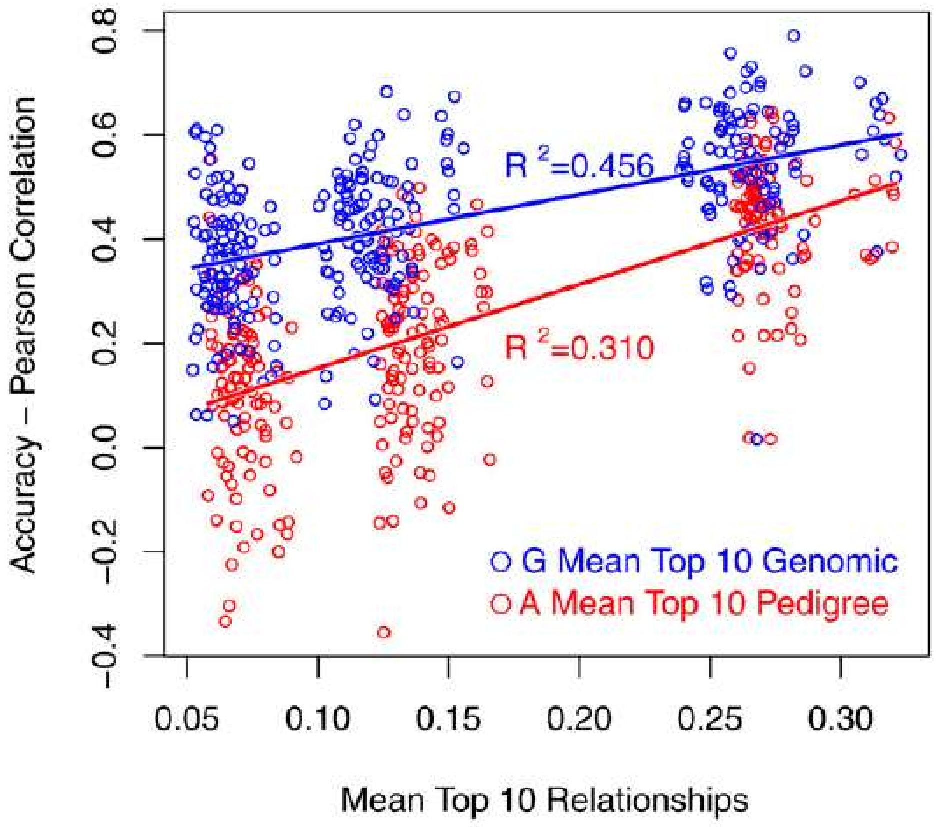


Habier et al., GSE 2010

- Accuracy into new environments low
  - Genotype-by-environment interactions

Influence of relationships on prediction accuracy

- Relationship of validation to reference important contributor to accuracy



Daetwyler et al., 2013. Genetics. 193:347-365

Reference population design

- Which individuals/lines?
- The relationship of the reference population to the selection candidates affects accuracy of GEBV
- Need individuals close to those being predicted in reference
- At the same time, as diverse as possible so that many individuals/lines can be accurately predicted

Optimal breeding program design

- Predict GEBV with good accuracy in selection candidates with only a DNA sample

- Achieve higher accuracy earlier in life

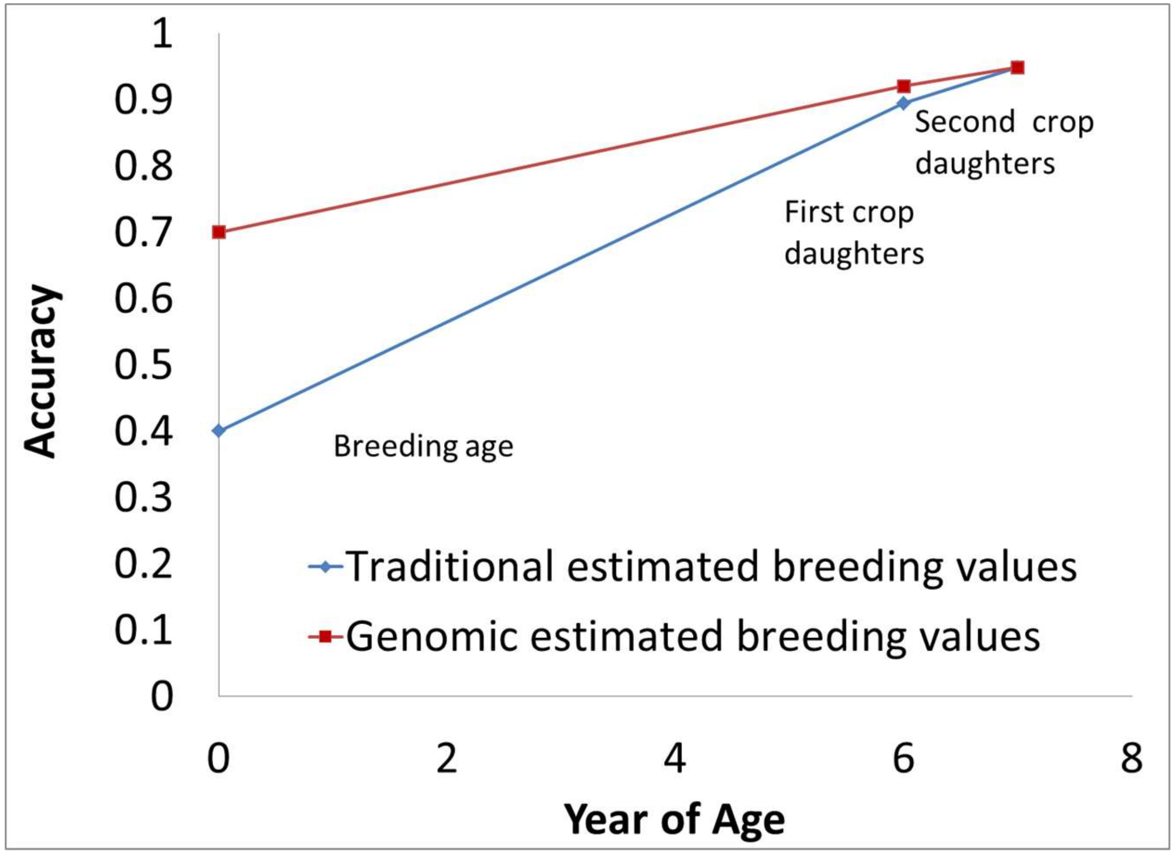
- How does this change the optimal breeding program design?

- Breed from individuals as early as possible

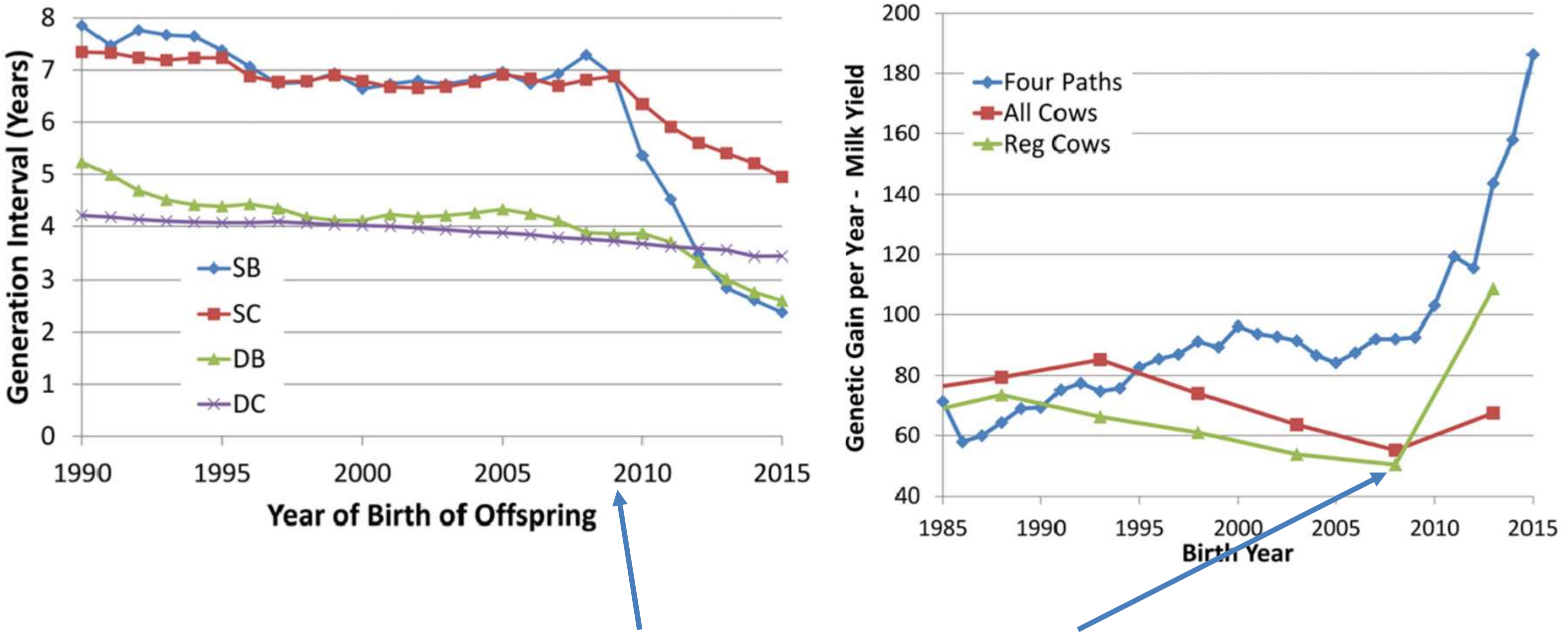
Genomic selection: dairy cattle

$$\Delta G = \frac{ir\sigma_g}{L}$$

ΔG genetic change
i selection intensity
r selection accuracy
σg genetic std deviation
L generation interval



Genetic Gain: US Dairy Cattle

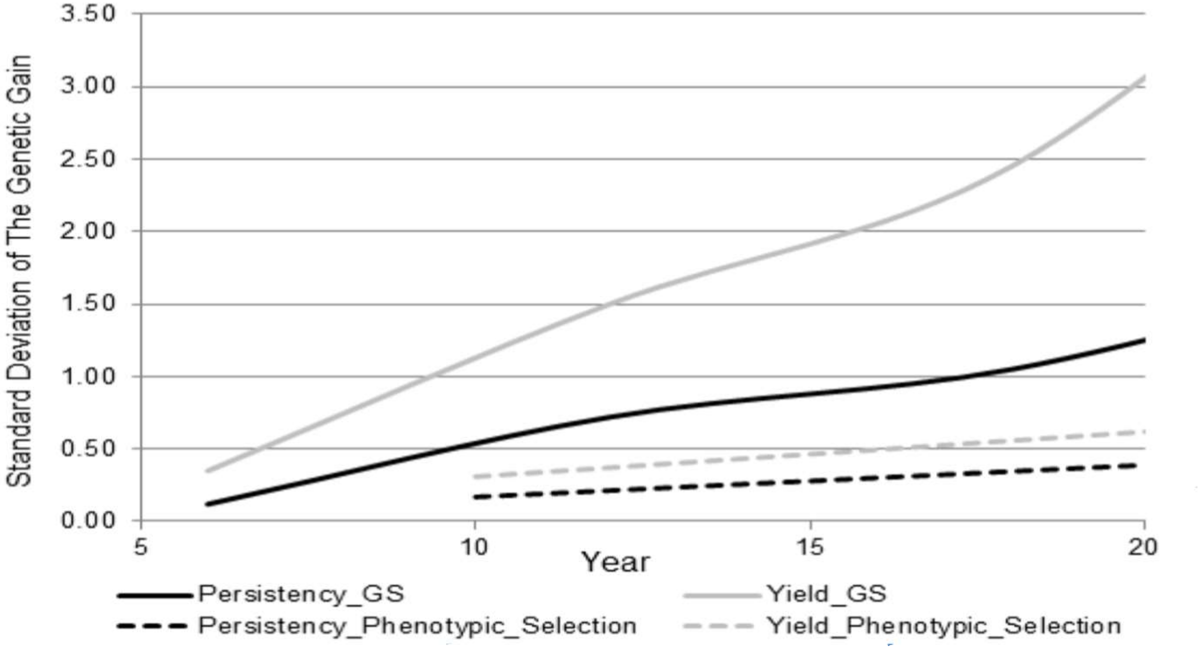


Introduction of genomic selection

Large increases in genetic gain from genomic selection

Garcia-Ruiz et al., 2016. PNAS.
https://www.pnas.org/content/pnas/113/28/E3995.full.pdf

Genetic Gain: Pasture Grasses (Simulations)



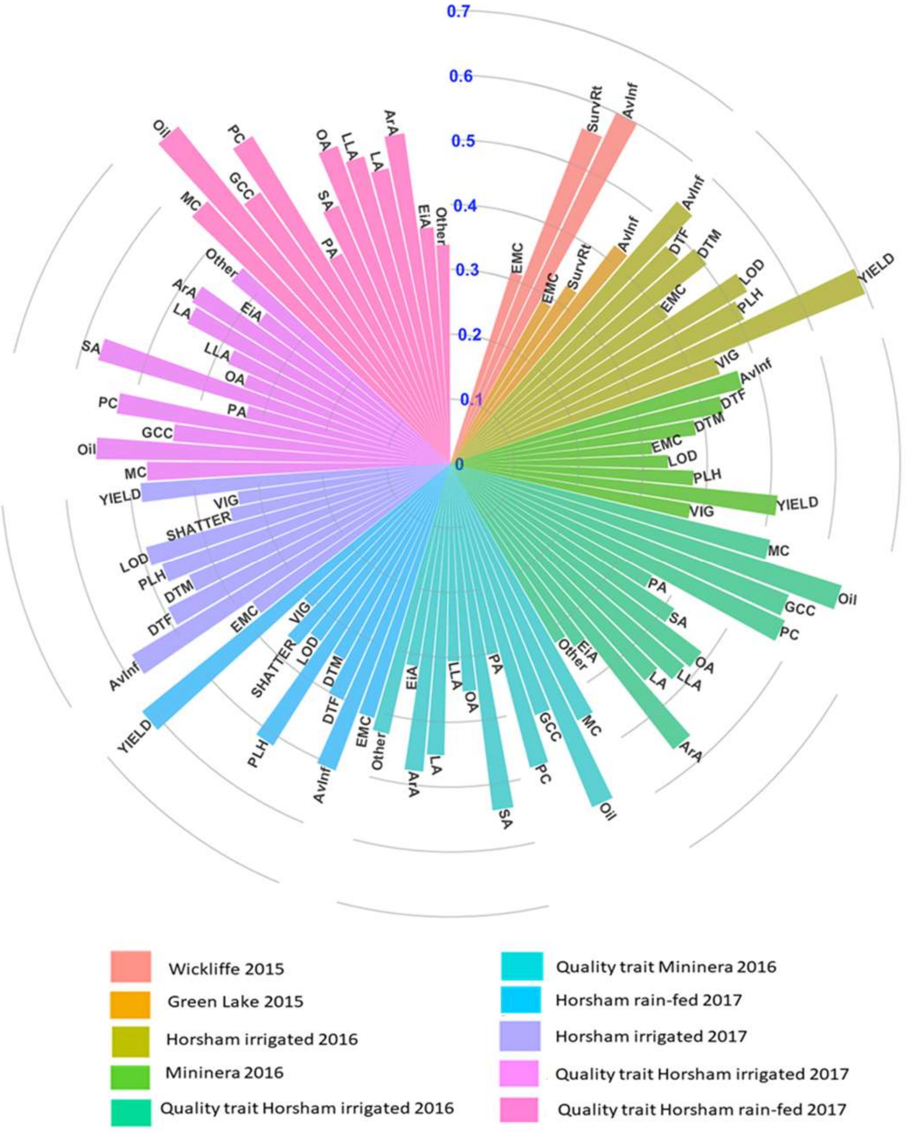
Large increases in genetic gain from genomic selection (GS)

Lin et al., 2016. The Plant Genome. 9:1

Canola Genomic Prediction

- 200 spring canola lines
- 60,000 genotyping-by-sequencing SNP markers
- Within-site GBLUP

Accuracy moderate to high across 22 key canola traits.



Fikere et al., 2020. Plants 9:719