Genome-based genetic evaluation

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-07

Learning objectives

- Understand limitations of estimates from the pedigree-based model → why we would need genome-based model

- Understand how to combine phenotype information from all relatives connected via genomic data

- Practice inference of breeding values with the genome-based model
  - simple cases using R matrix algebra
  - using other packages

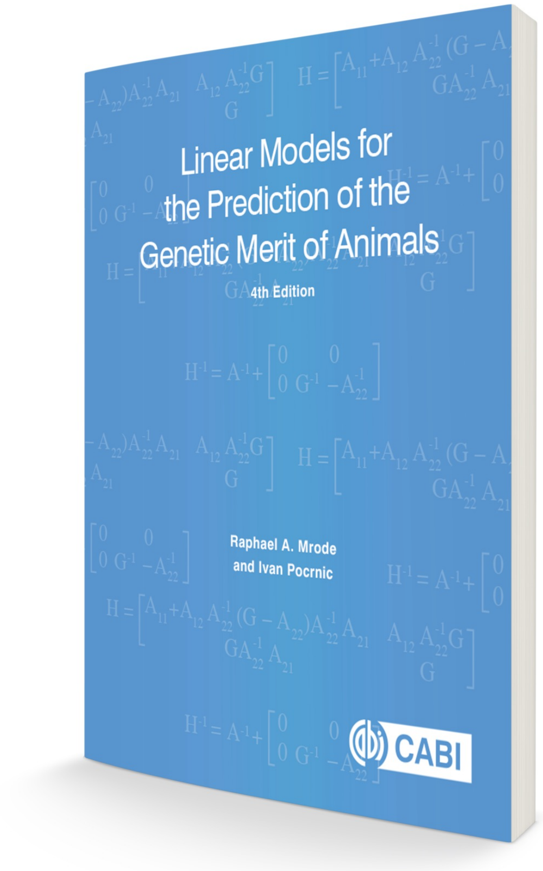
Linear Models for the Prediction of the Genetic Merit of Animals

CABI Biotechnology Series

September 2023 | 412pp

Raphael A Mrode
Ivan Pocrnic

Robin Thompson
Gregor Gorjanc


Linear Models for the Prediction of the Genetic Merit of Animals
4th Edition
Raphael A. Mrode
and Ivan Pocrnic
CABI

See chapters 11-14!

Learning objectives

- Understand limitations of estimates from the pedigree-based model

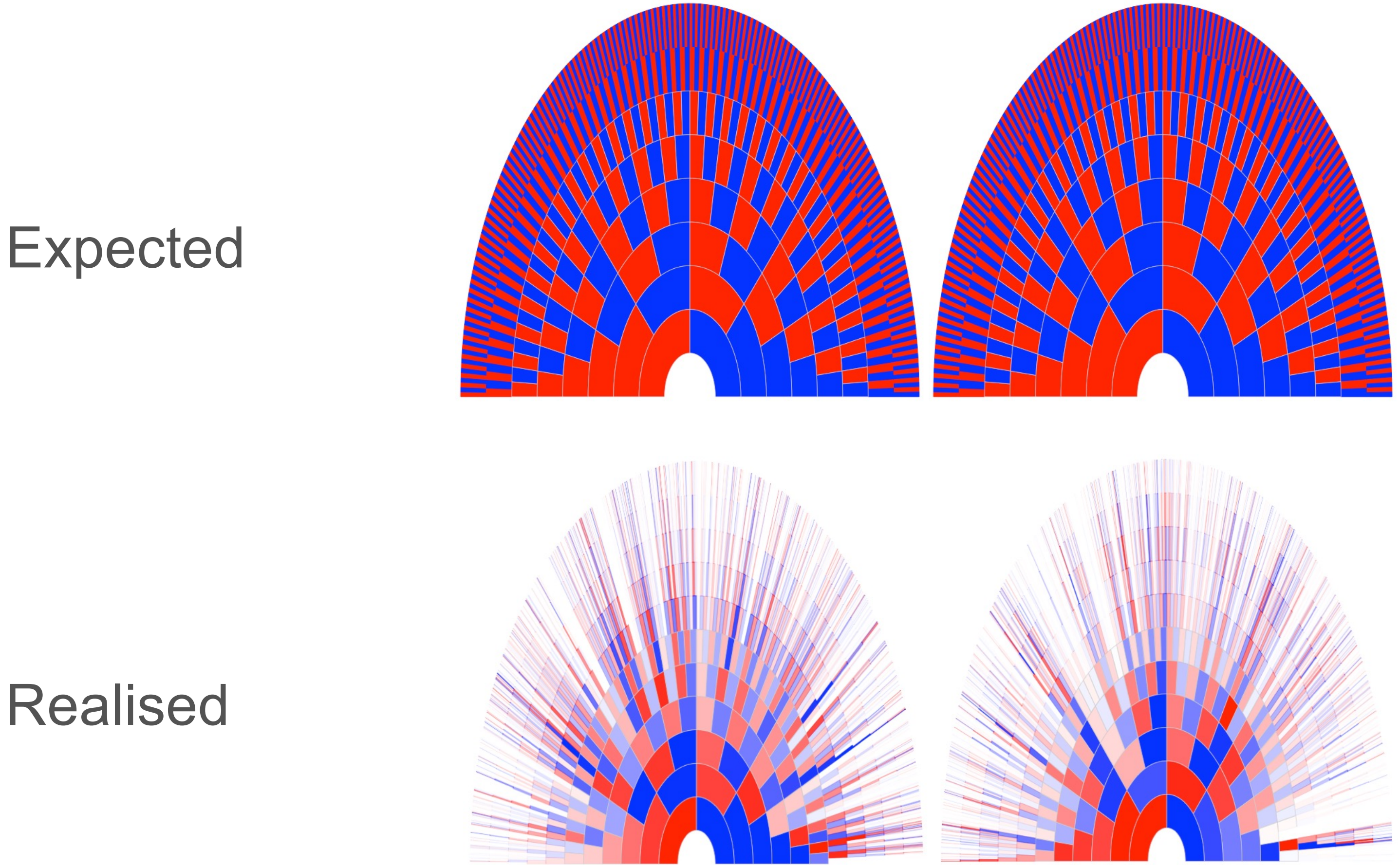
- Understand how to combine phenotype information from all relatives connected via genomic data

- Practice inference of breeding values with the genome-based model
  - simple cases using R matrix algebra
  - using other packages

Limitations with pedigree-based model

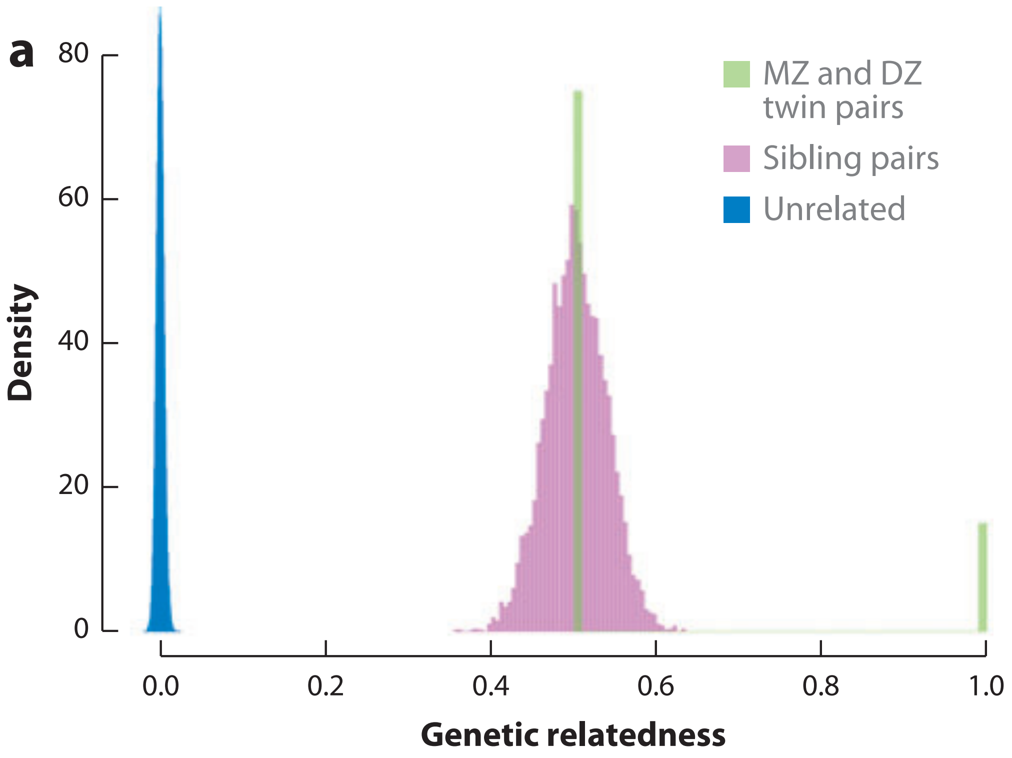
- **With pedigrees we can *apriori* describe expected amount of variation**
  - between pedigree founders (assumed unrelated)
  - between families
    (variation between family means / parent average terms)
  - within families
    (variation between Mendelian sampling terms)

Expected and realised relatedness



Coop (2013)

Expected and realised relatedness



b

$$G = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

Pedigree design

4 individuals from 2 families, including 1 MZ twin pair and 1 DZ twin pair

$$G = \begin{bmatrix} 1 & & & \\ 0.543 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0.476 & 1 \end{bmatrix}$$

Within-family design

4 individuals from 2 families, including 2 full sibling pairs

$$G = \begin{bmatrix} 1.002 & & & \\ -0.016 & 1.018 & & \\ -0.003 & -0.006 & 0.994 & \\ 0.015 & 0.021 & -0.011 & 0.983 \end{bmatrix}$$

Population design

4 'unrelated' individuals from the same population

Vinkhuyzen et al. (2013)

Limitations with pedigree-based model

- **With pedigrees we can *apriori* describe expected amount of variation**
  - between pedigree founders (assumed unrelated)
  - between families
    (variation between family means / parent average terms)
  - within families
    (variation between Mendelian sampling terms)
- **When we fit the model, we *aposteriori* estimate "realised" deviations**
  (phenotype resemblance updates assumed pedigree relationships)
  → the more information per individual, the higher accuracy

Limitations with pedigree-based model

- What does all this mean in practice:
  - Decent accuracy of estimated breeding values for individuals with own phenotypic data or progeny with phenotypic data (genomic data won’t add much more information!)

Limitations with pedigree-based model

- What does all this mean in practice:
  - Decent accuracy of estimated breeding values for individuals with own phenotypic data or progeny with phenotypic data (genomic data won’t add much more information!)
  - Low accuracy of estimated breeding values for individuals without own phenotypic data or progeny with phenotypic data (genomic data can add more information)

Limitations with pedigree-based model

- What does all this mean in practice:
  - Decent accuracy of estimated breeding values for individuals with own phenotypic data or progeny with phenotypic data (genomic data won't add much more information!)
  - Low accuracy of estimated breeding values for individuals without own phenotypic data or progeny with phenotypic data (genomic data can add more information)
  - Zero accuracy of estimated breeding values within a family with progeny prediction!!! → we can not differentiate full-sibs :((progeny prediction does not capture Mendelian sampling terms, so genomic data can add a lot of information)

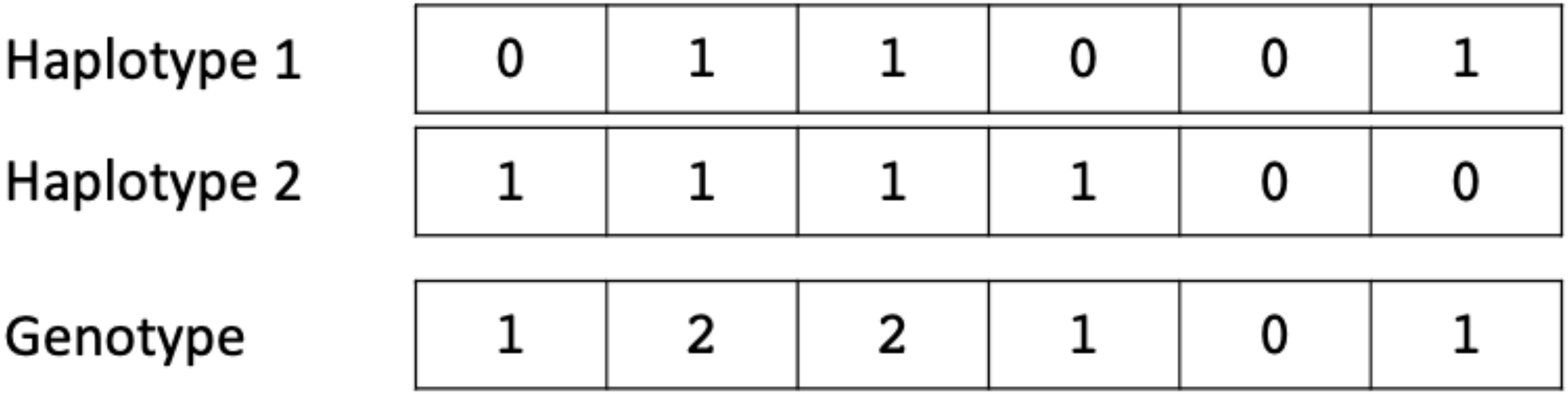
Limitations with pedigree-based model

- Pedigree could be
  - wrong!
  - partially missing
  - missing altogether!

- Genomic data should help with all the mentioned issues!

Data

Recall the 0/1 encoding of haplotypes and 0/1/2 encoding of genotypes



Haplotype 1	0	1	1	0	0	1
Haplotype 2	1	1	1	1	0	0
Genotype	1	2	2	1	0	1

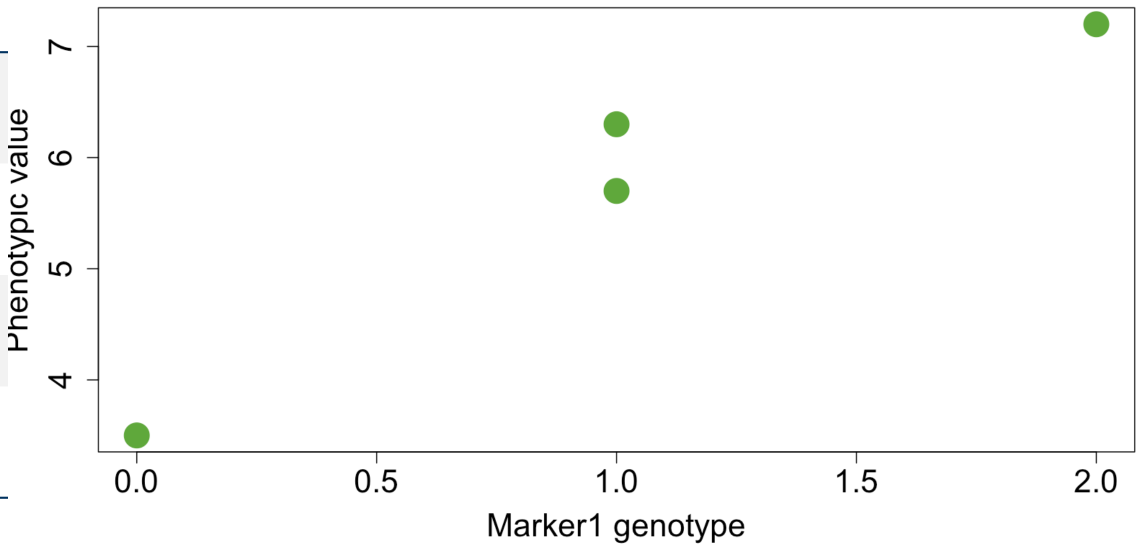
Data - example

ID	Pheno	Marker1	Marker2	Marker3	Marker4	Marker5
1	7.2	2	2	2	0	1
2	3.5	0	2	1	1	0
3	5.7	1	1	1	1	1
4	6.3	2	1	0	1	2

How could we model this data?

- Let's focus on one locus first

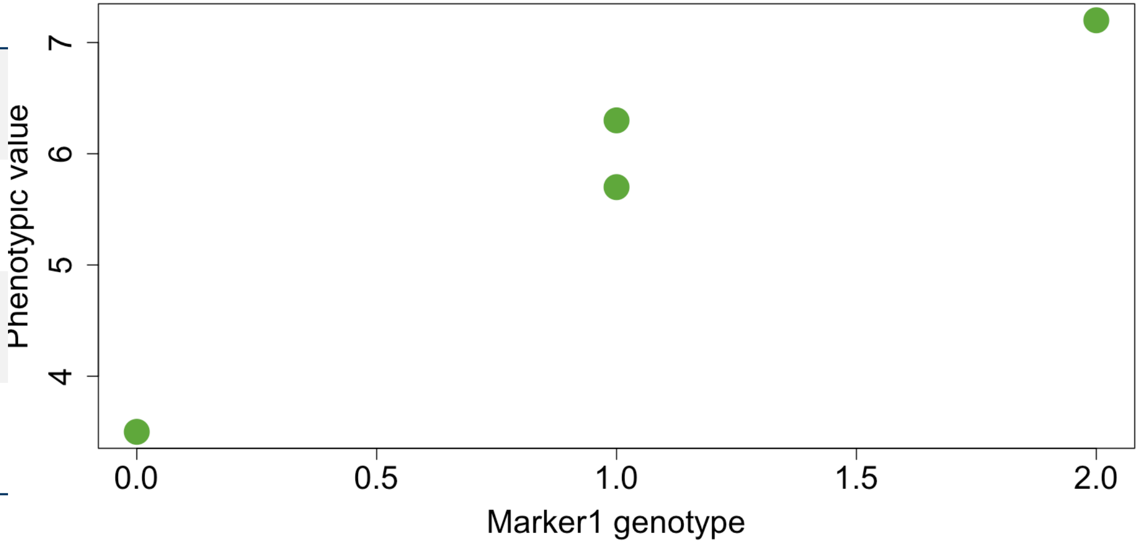
ID	Pheno	Marker1
1	7.2	2
2	3.5	0
3	5.7	1
4	6.3	2



How could we model this data?

- Let’s focus on one locus first

ID	Pheno	Marker1
1	7.2	2
2	3.5	0
3	5.7	1
4	6.3	2



- We have:
  - – continuous variable (Pheno) → response
  - – continuous variable (Marker1) → covariate

Linear regression (single marker)

- Estimating the association between phenotypic value and marker 1 genotypes (as allele dosage)

$$y_1 = 7.2 = \mu + 2\alpha_1 + e_1$$
$$y_2 = 3.5 = \mu + 0\alpha_1 + e_2$$
$$y_3 = 5.7 = \mu + 1\alpha_1 + e_3$$
$$y_4 = 6.3 = \mu + 2\alpha_1 + e_4$$

$$e_i \sim N(0, \sigma_e^2)$$

- Assuming causality, α is allele substitution effect

Linear regression (single marker)

- Estimating the association between phenotypic value and marker 1 genotypes (as allele dosage)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 7.2 \\ 3.5 \\ 5.7 \\ 6.3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} (\alpha_1) + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{W} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{W} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{y} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{y} \end{pmatrix}$$

$$Var(\boldsymbol{\alpha}|\boldsymbol{y}) = diag(\boldsymbol{C}^{-1})_{\alpha}\, \sigma_e^2$$

Breeding values at single marker

- Model: $$\begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \\ a_{4,1} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} (\alpha_1) = \boldsymbol{a}_1 = \boldsymbol{W\alpha}$$

$$E(\boldsymbol{a}_1) = E(\boldsymbol{W\alpha}) = \boldsymbol{W}E(\boldsymbol{\alpha})$$

$$Var(\boldsymbol{a}_1) = Var(\boldsymbol{W\alpha}) = \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T$$

Breeding values at single marker

- Model:

$$\begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \\ a_{4,1} \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} (\alpha_1) = \boldsymbol{a}_1 = \boldsymbol{W\alpha}$$

$$E(\boldsymbol{a}_1) = E(\boldsymbol{W\alpha}) = \boldsymbol{W}E(\boldsymbol{\alpha})$$

$$Var(\boldsymbol{a}_1) = Var(\boldsymbol{W\alpha}) = \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T$$

- Estimator/Predictor:

$$E(\boldsymbol{a}_1|\boldsymbol{y}) = \widehat{\boldsymbol{a}}_1 = \boldsymbol{W}\widehat{\boldsymbol{\alpha}}$$

$$Var(\boldsymbol{a}_1|\boldsymbol{y}) = \boldsymbol{W}Var(\boldsymbol{\alpha}|\boldsymbol{y})\boldsymbol{W}^T$$

Questions?!

Multiple linear regression (multiple markers)

- Estimating the association between phenotypic value and marker 1-5 genotypes (as allele dosage)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 7.2 \\ 3.5 \\ 5.7 \\ 6.3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 2 & 2 & 2 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

Multiple linear regression (multiple markers)

- Estimating the association between phenotypic value and marker 1-5 genotypes (as allele dosage)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 7.2 \\ 3.5 \\ 5.7 \\ 6.3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 2 & 2 & 2 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$

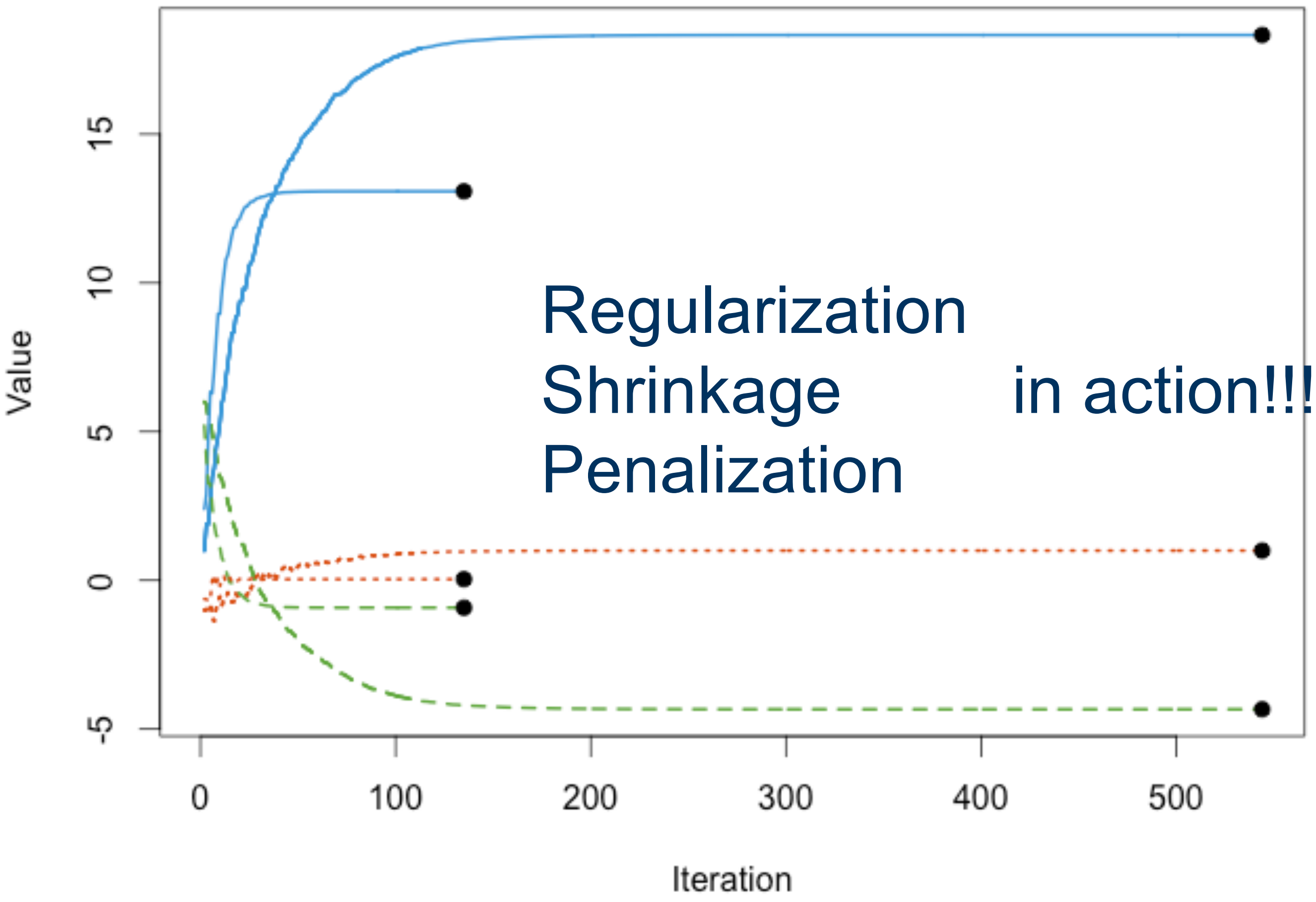
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

$$\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$$

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{W} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{W} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{y} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{y} \end{pmatrix}$$

$$Var(\boldsymbol{\alpha}|\boldsymbol{y}) = diag(\boldsymbol{C}^{-1})_{\boldsymbol{\alpha}}\, \sigma_e^2$$

Role of the prior for marker effects $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{I}\sigma_{\alpha}^{2})$



Regularization
Shrinkage in action!!!
Penalization

Breeding values over all markers

- Model: $$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \boldsymbol{a} = \boldsymbol{W\alpha}$$

$$E(\boldsymbol{a}) = E(\boldsymbol{W\alpha}) = \boldsymbol{W}E(\boldsymbol{\alpha}) = \mathbf{0}$$

$$Var(\boldsymbol{a}) = Var(\boldsymbol{W\alpha}) = \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T = \boldsymbol{WW}^T\sigma_\alpha^2$$

Breeding values over all markers

- Model: $$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{pmatrix} = \boldsymbol{a} = \boldsymbol{W\alpha}$$

$$E(\boldsymbol{a}) = E(\boldsymbol{W\alpha}) = \boldsymbol{W}E(\boldsymbol{\alpha}) = \mathbf{0}$$

$$Var(\boldsymbol{a}) = Var(\boldsymbol{W\alpha}) = \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{W}^T\sigma_\alpha^2$$

- Estimator/Predictor:

$$E(\boldsymbol{a}_1|\boldsymbol{y}) = \widehat{\boldsymbol{a}}_1 = \boldsymbol{W}\widehat{\boldsymbol{\alpha}}$$

$$Var(\boldsymbol{a}_1|\boldsymbol{y}) = \boldsymbol{W}Var(\boldsymbol{\alpha}|\boldsymbol{y})\boldsymbol{W}^T$$

Questions?!

Prediction of genomic prediction accuracy ("global")

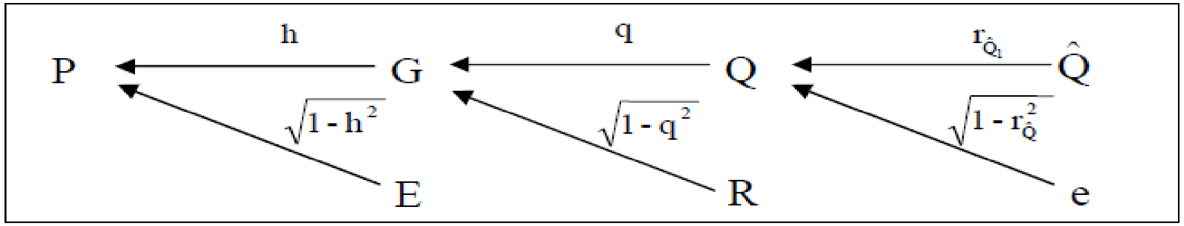
- Effective no. of chr. segments

$$M_e = 2N_eLC/ln(N_eL)$$

- Prop. of genetic variance captured by markers

$$q^2 = M/(M + M_e)$$

- Reliability of GEBV $R^2 = T/(1 + T), T = n\,q^2h^2/M_e$
- Reliability of EBV $R^2 = (T/(1 + T))q^2$



Goddard (2011), Dekkers (2007)

Inputs

- M no. of genome-wide markers
- N_e effective population size
- L average size of chromosomes in Morgans
- C no. of chromosomes
- h^2 heritability of training phenotypes
- n no. of training individuals

Maize example (train and predict in family)

- M no. of genome-wide markers = 200
- N_e effective population size = 1
- L average size of chromosomes = 2
- C no. of chromosomes = 10
- h^2 heritability of phenotype included into training = 0.25
- n no. of training individuals = 100
- Effective no. of chr. segments

  $M_e=2N_eLC/\ln(N_eL)=2\times1\times2\times10/\ln(1\times2)=58$
- Prop. of genetic variance captured by markers

  $q^2=M/(M+M_e)=200/(200+58)=0.76$
- Reliability of GEBV

  $R^2\approx T/(1+T)$, $T=nq^2h^2/M_e$

  $T=100\times0.76\times0.25/58=0.34$, $R^2\approx0.25$, $r\approx0.5$
- Reliability of EBV

  $R^2\approx(T/(1+T))q^2=0.19$, $r\approx0.44$

Maize example (predict from other families)

- M no. of genome-wide markers = 10,000
- N_e effective population size = 50
- L average size of chromosomes = 2
- C no. of chromosomes = 10
- h^2 heritability of phenotype included into training = 0.25
- n no. of training individuals = 2000
- Effective no. of chr. segments
  $M_e=2N_eLC/\ln(N_eL)=2\times50\times2\times10/\ln(50\times2)=434$
- Prop. of genetic variance captured by markers
  $q^2=M/(M+M_e)=10000/(10000+434)=0.96$
- Reliability of GEBV
  $R^2\approx T/(1+T)$, $T=nq^2h^2/M_e$
  $T=2000\times0.96\times0.25/434=1.1$, $R^2\approx0.53$, $r\approx0.72$
- Reliability of   EBV
  $R^2\approx(T/(1+T))q^2=0.50$, $r\approx0.71$

Dairy bulls example

- M no. of genome-wide markers = 50,000
- N_e effective population size = 50
- L average size of chromosomes = 1
- C no. of chromosomes = 30
- h^2 heritability of phenotype included into training = 0.80
- n no. of training individuals = 1000
- Effective no. of chr. segments

  $M_e=2N_eLC/\ln(N_eL)=2\times50\times1\times30/\ln(50\times1)=767$
- Prop. of genetic variance captured by markers

  $q^2=M/(M+M_e)=50{,}000/(50{,}000+767)=0.98$
- Reliability of GEBV

  $R^2\approx T/(1+T)$, $T=nq^2h^2/M_e$

  $T=1000\times0.98\times0.80/767=1.02$, $R^2\approx0.50$, $r\approx0.71$
- Reliability of   EBV

  $R^2\approx(T/(1+T))q^2=0.50$, $r\approx0.70$

Dairy cows example

- M no. of genome-wide markers = 50,000
- N_e effective population size = 50
- L average size of chromosomes = 1
- C no. of chromosomes = 30
- h^2 heritability of phenotype included into training = 0.30
- n no. of training individuals = ??? How many to get R^2 EBV of 0.50???
- Effective no. of chr. segments

  $M_e=2N_eLC/\ln(N_eL)=2\times50\times1\times30/\ln(50\times1)=767$
- Prop. of genetic variance captured by markers

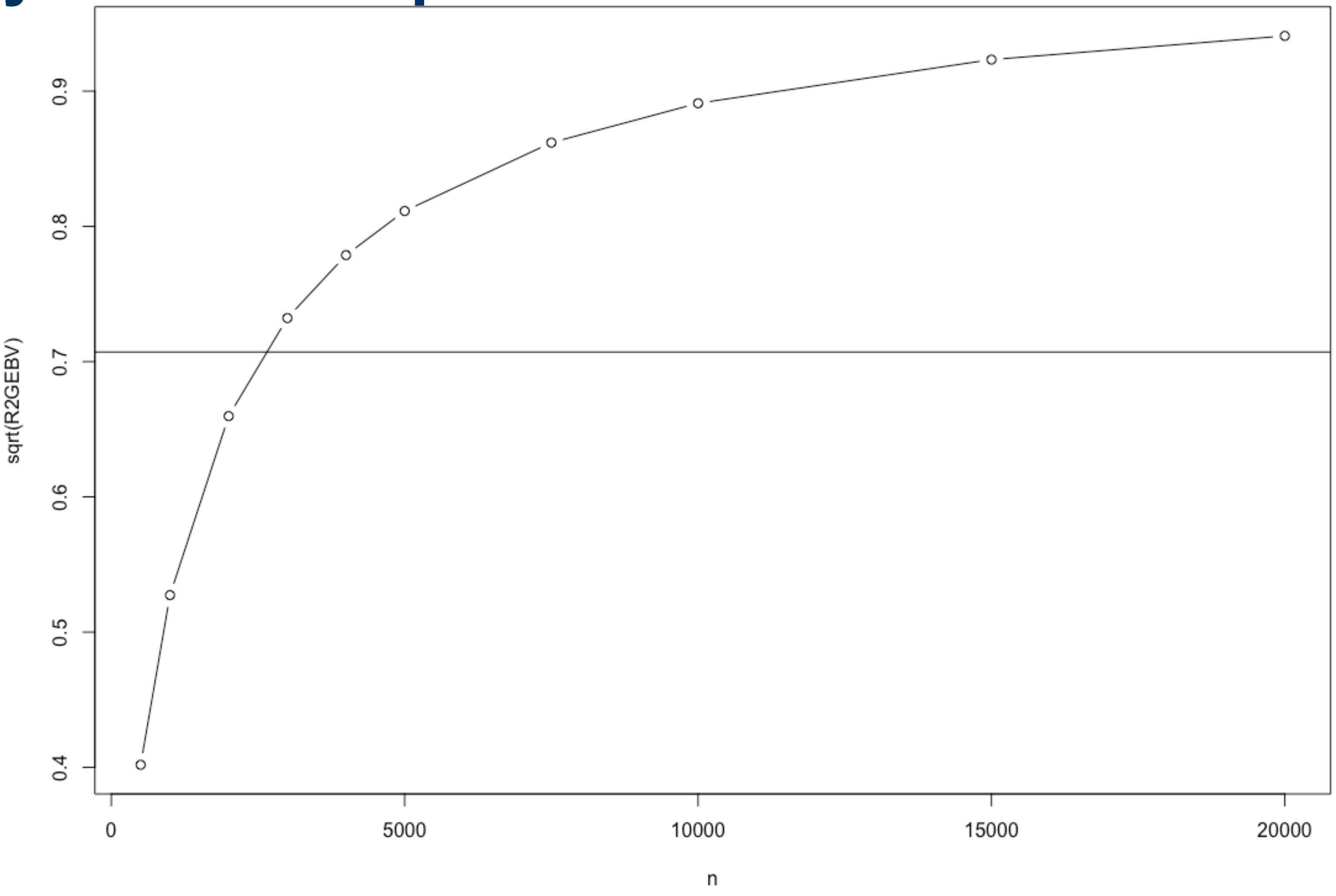
  $q^2=M/(M+M_e)=50000/(50000+767)=0.98$
- Reliability of GEBV

  $R^2\approx T/(1+T)$, $T=nq^2h^2/M_e$

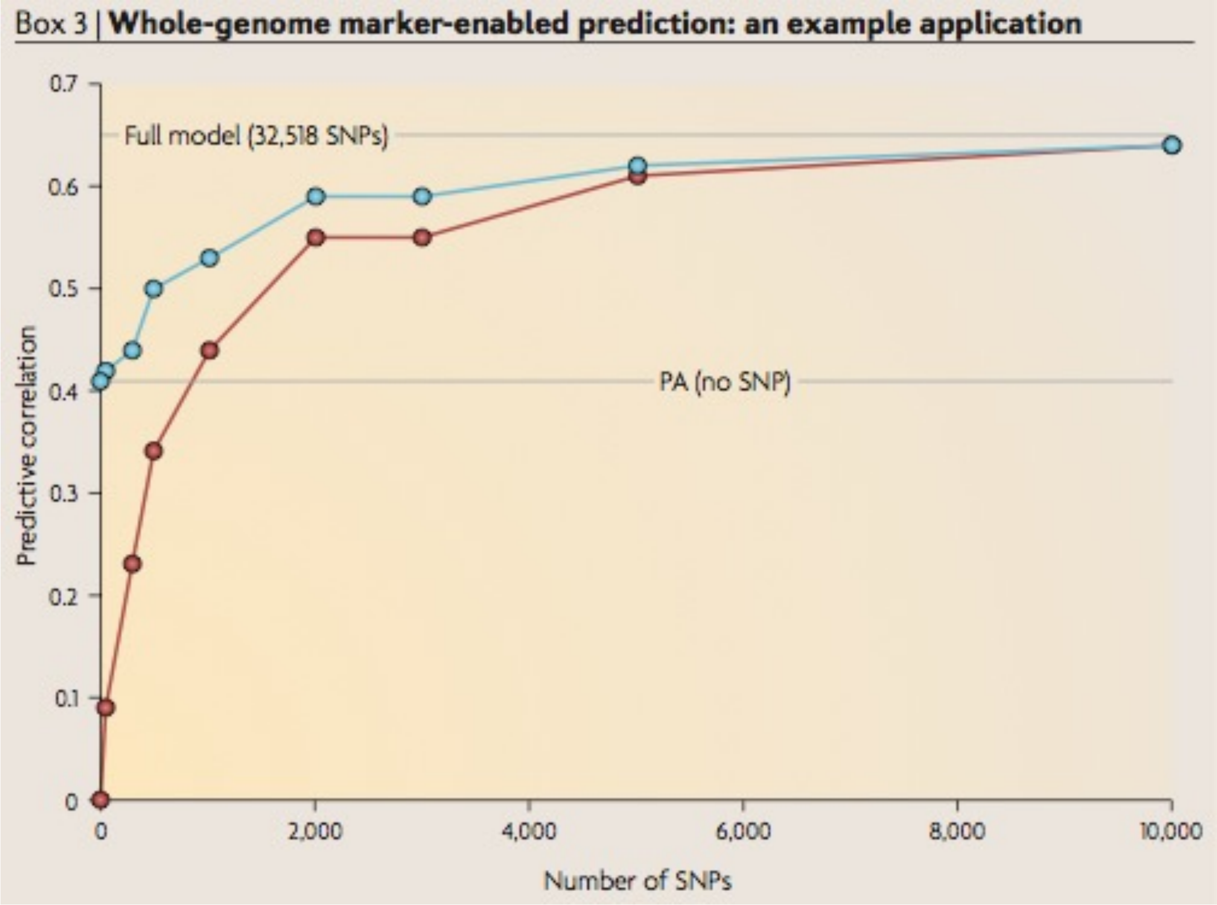
  $T=???\times0.98\times0.30/767=???$, $R^2\approx???$, $r\approx???$
- Reliability of EBV

  $R^2\approx(T/(1+T))q^2=???$, $r\approx???$

Dairy cows example

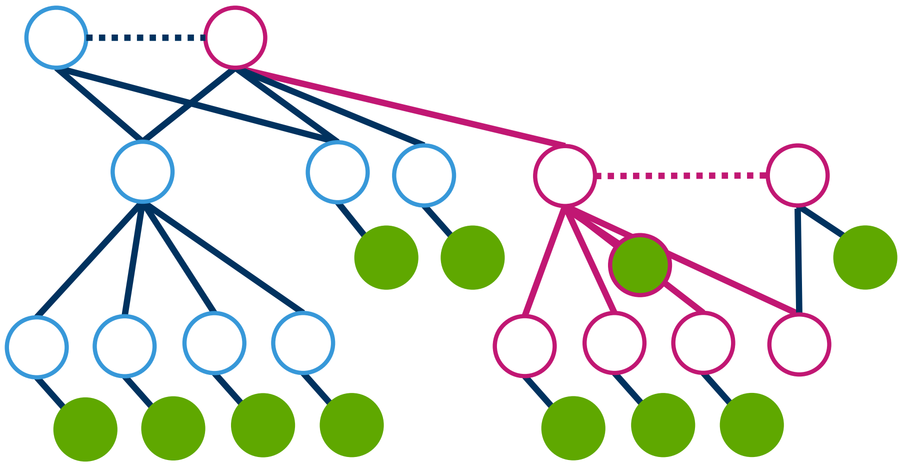


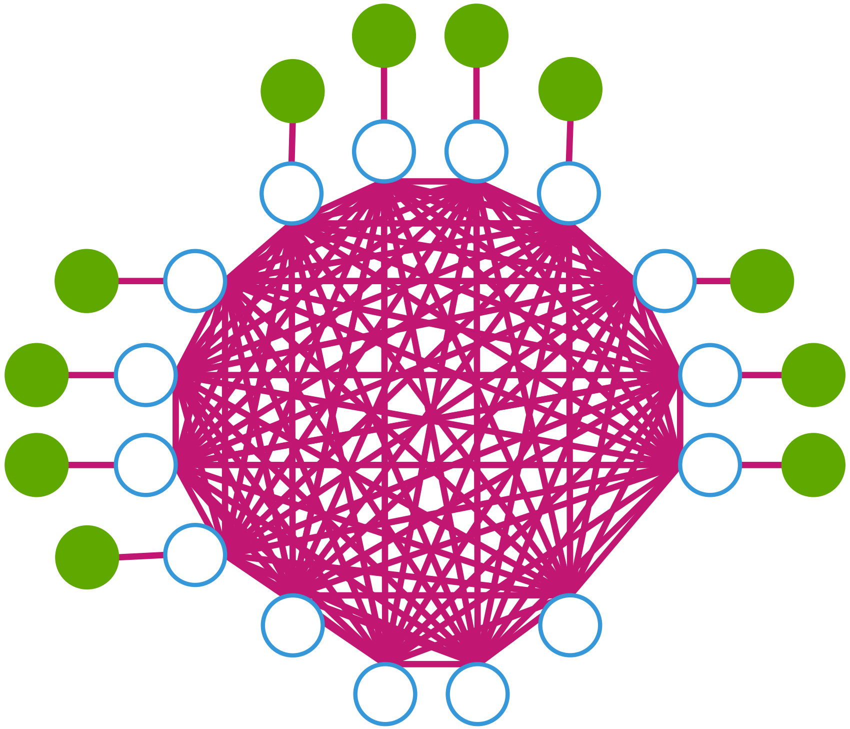
~10,000 ***good*** markers works quite well



de Los Campos et al. (2010)

Information for an individual – pedigree vs. genomics





Questions?!

Marker & individual genome-based models

- Marker genome-based model (SNP-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

$$\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$$

- Individual genome-based model (G-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{ZW\alpha} + \boldsymbol{e}$$

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

$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$

$$\boldsymbol{a} \sim N(\boldsymbol{0}, ?\,\sigma_\alpha^2)$$

Z so we can include non-phenotyped individuals

Marker & individual genome-based models

- Marker genome-based model (SNP-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$$

- Individual genome-based model (G-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{ZW\alpha} + \boldsymbol{e}$$
$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, ?\,\sigma_\alpha^2)$$

Z so we can include non-phenotyped individuals

$$\begin{aligned} Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\ &= \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T \\ &= \boldsymbol{WW}^T\sigma_\alpha^2 \end{aligned}$$

Marker & individual genome-based models

- Marker genome-based model (SNP-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$$

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{W} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{W} + \boldsymbol{I}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{y} \\ \boldsymbol{W}^T\boldsymbol{E}^{-1}\boldsymbol{y} \end{pmatrix}$$

$$Var(\boldsymbol{\alpha}|\boldsymbol{y}) = diag(\boldsymbol{C}^{-1})_{\boldsymbol{\alpha}}\ \sigma_e^2$$

- Individual genome-based model (G-BLUP)

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{ZW\alpha} + \boldsymbol{e}$$
$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{WW}^T\sigma_\alpha^2)$$

Z so we can include non-phenotyped individuals

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{E}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{E}^{-1}\boldsymbol{Z} + \boldsymbol{WW}^{T^{-1}}\frac{\sigma_e^2}{\sigma_\alpha^2} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{E}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{E}^{-1}\boldsymbol{y} \end{pmatrix}$$

$$Var(\boldsymbol{a}|\boldsymbol{y}) = diag(\boldsymbol{C}^{-1})_{\boldsymbol{a}}\sigma_e^2$$

Genomic covariance-like coefficient matrices

- Genotype matrix **W** is nInd x nLoc
- Between individuals

$$
\begin{aligned}
Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\
&= \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T \\
&= \boldsymbol{W}\boldsymbol{W}^T\sigma_\alpha^2
\end{aligned}
\longrightarrow \boldsymbol{W}\boldsymbol{W}^T
$$

Covariance-like coefficients between individuals (nInd x nInd) **similar to NRM matrix**

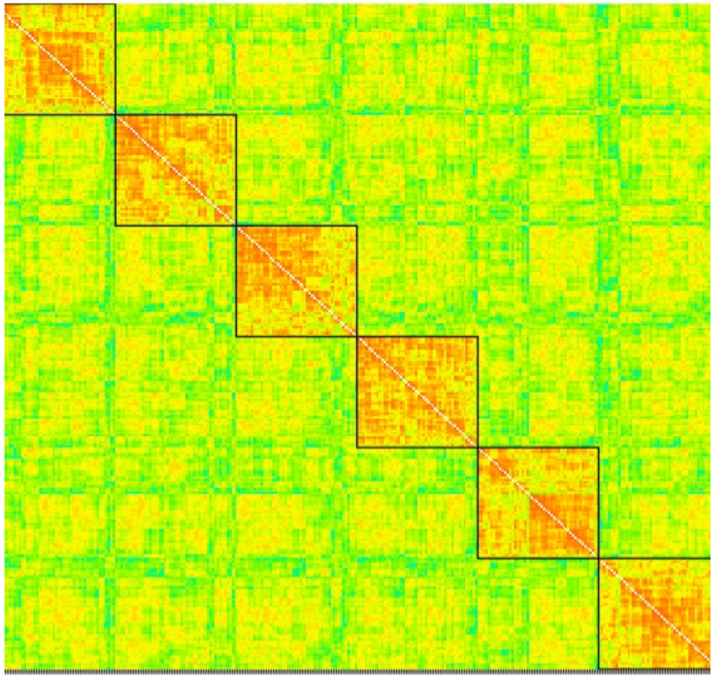
- Between loci
  - sum-of-squares $\boldsymbol{W}^T\boldsymbol{W}$

Covariance-like coefficients between loci (nLoc x nLoc) **similar to LD matrix**

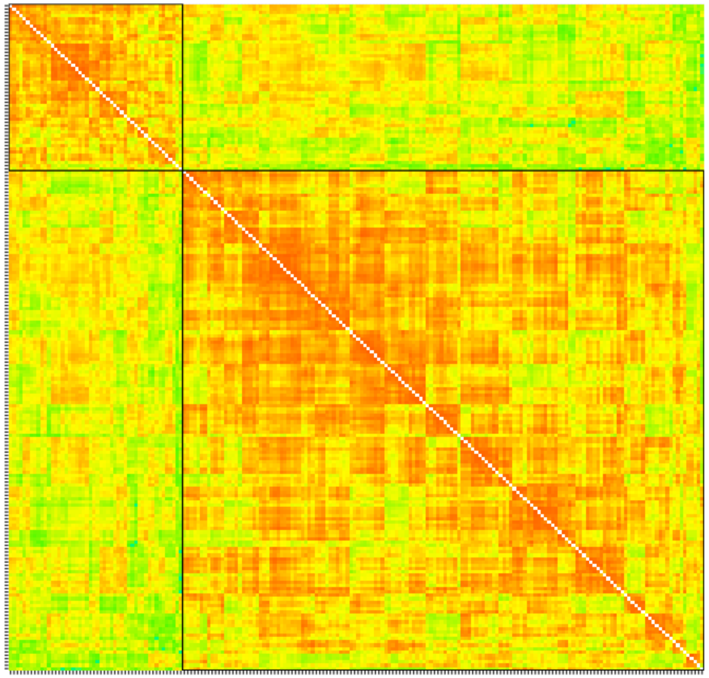
  - covariance $Cov(\boldsymbol{W}) = \boldsymbol{C} = \left(\boldsymbol{W} - E(\boldsymbol{W})\right)^T\left(\boldsymbol{W} - E(\boldsymbol{W})\right)/(n-1)$
  - correlation $Cor(\boldsymbol{W}) = diag(\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{C}diag(\boldsymbol{C})^{-\frac{1}{2}}$

Genomic covariance-like coefficient matrices

Between loci



Between individuals



Genomic covariance-like coefficient matrices

- Genotype matrix **W** is nInd x nLoc
- Between individuals
  - sum-of-squares $\boldsymbol{W}\boldsymbol{W}^T$
  - covariance $Cov(\boldsymbol{W}^T) = \boldsymbol{C} = \left(\boldsymbol{W} - E(\boldsymbol{W})\right)\left(\boldsymbol{W} - E(\boldsymbol{W})\right)^T/(n-1)$
  - correlation $Cor(\boldsymbol{W}^T) = diag(\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{C}diag(\boldsymbol{C})^{-\frac{1}{2}}$

Covariance-like coefficients
between individuals
(nInd x nInd)
similar to NRM matrix

I want the genome-based NRM
(following the pedigree-based NRM)!?

Genome-based NRM

- Maybe we don't need it!

$$\begin{aligned} Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\ &= \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T \\ &= \boldsymbol{WW}^T\sigma_\alpha^2 \end{aligned}$$

Genome-based NRM

- Maybe we don't need it!

$$\begin{aligned} Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\ &= \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T \\ &= \boldsymbol{W}\boldsymbol{W}^T\sigma_\alpha^2 \end{aligned}$$

- Many proposed versions:
  - [-1, 0, 1] centering $(\boldsymbol{W}-1)(\boldsymbol{W}-1)^T$
    - diagonals = the number of homozygous loci for individuals
    - off-diagonals = the number of alleles shared between individuals

Genome-based NRM

- Maybe we don't need it!

$$\begin{aligned} Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\ &= \boldsymbol{W}Var(\boldsymbol{\alpha})\boldsymbol{W}^T \\ &= \boldsymbol{WW}^T\sigma_\alpha^2 \end{aligned}$$

- Many proposed versions:
  - [-1, 0, 1] centering $(\boldsymbol{W}-1)(\boldsymbol{W}-1)^T$
    - diagonals = the number of homozygous loci for individuals
    - off-diagonals = the number of alleles shared between individuals
  - VanRaden 1 (to match pedigree NRM)

$$\boldsymbol{G} = \left(\boldsymbol{W} - E(\boldsymbol{W})\right)\left(\boldsymbol{W} - E(\boldsymbol{W})\right)^T / \sum diag\left(Cov(\boldsymbol{W})\right)$$

$$E(\boldsymbol{W}_i) = 2p_i$$

$$Var(\boldsymbol{W}_i) = 2p_iq_i(1+F_i)$$

  - Many other versions!!!

Genome-based NRM

- Whatever the choice, there is useful information in **G**!
- Take a trio of diploid individuals and use [-1, 0, 1] coding in **w**

$$\boldsymbol{w}_{f(i)} = \boldsymbol{w}_{f(i),1} + \boldsymbol{w}_{f(i),2}$$
$$\boldsymbol{w}_{m(i)} = \boldsymbol{w}_{m(i),1} + \boldsymbol{w}_{m(i),2}$$
$$\boldsymbol{w}_{i} = \boldsymbol{w}_{i,1} + \boldsymbol{w}_{i,2}$$

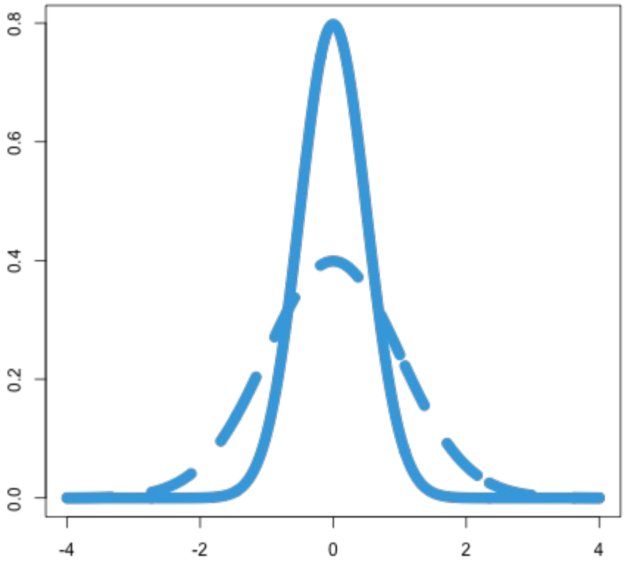
- Realised shared number of alleles between individuals

$$\begin{pmatrix} \boldsymbol{w}_{f(i)}\boldsymbol{w}_{f(i)}^T & & sym. \\ \boldsymbol{w}_{m(i)}\boldsymbol{w}_{f(i)}^T & \boldsymbol{w}_{m(i)}\boldsymbol{w}_{m(i)}^T & \\ \boldsymbol{w}_{i}\boldsymbol{w}_{f(i)}^T & \boldsymbol{w}_{i}\boldsymbol{w}_{m(i)}^T & \boldsymbol{w}_{i}\boldsymbol{w}_{i}^T \end{pmatrix}$$

Genome-based NRM - What do these terms mean?

$$\begin{pmatrix} \boldsymbol{w}_{f(i)}\boldsymbol{w}_{f(i)}^{T} & & sym. \\ \boldsymbol{w}_{m(i)}\boldsymbol{w}_{f(i)}^{T} & \boldsymbol{w}_{m(i)}\boldsymbol{w}_{m(i)}^{T} & \\ \boldsymbol{w}_{i}\boldsymbol{w}_{f(i)}^{T} & \boldsymbol{w}_{i}\boldsymbol{w}_{m(i)}^{T} & \boldsymbol{w}_{i}\boldsymbol{w}_{i}^{T} \end{pmatrix}$$

- Diagonal: prior variances indicating how much individual breeding values ***could*** deviate from population mean



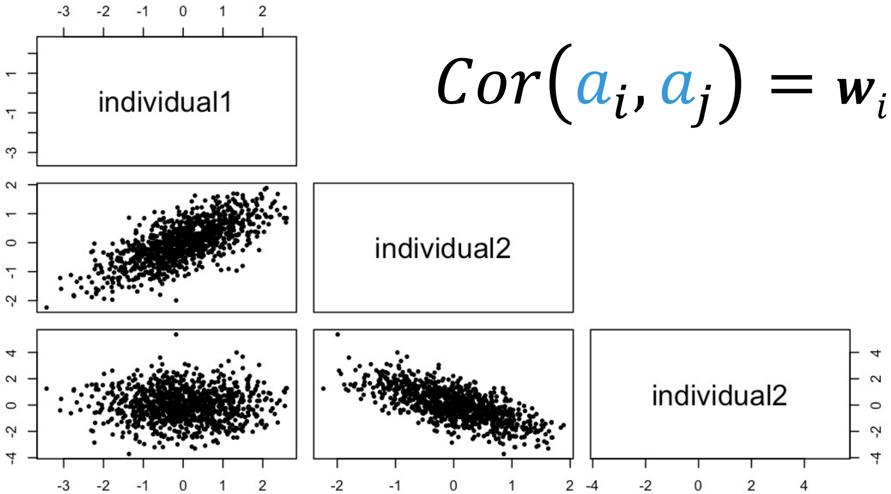
$$a_i \sim N(0, \boldsymbol{w}_i\boldsymbol{w}_i^{T}\sigma_{\alpha}^{2})$$

Genome-based NRM - What do these terms mean?

$$\begin{pmatrix} \boldsymbol{w}_{f(i)}\boldsymbol{w}_{f(i)}^T & & sym. \\ \boldsymbol{w}_{m(i)}\boldsymbol{w}_{f(i)}^T & \boldsymbol{w}_{m(i)}\boldsymbol{w}_{m(i)}^T & \\ \boldsymbol{w}_i\boldsymbol{w}_{f(i)}^T & \boldsymbol{w}_i\boldsymbol{w}_{m(i)}^T & \boldsymbol{w}_i\boldsymbol{w}_i^T \end{pmatrix}$$

- Off-diagonal: prior co-variances indicating how much individual breeding values ***could*** correlate compared to the “average pair”

$$Cor(a_i, a_j) = \boldsymbol{w}_i\boldsymbol{w}_j^T / \sqrt{\boldsymbol{w}_i\boldsymbol{w}_i^T\boldsymbol{w}_j\boldsymbol{w}_j^T}$$



Genome-based NRM - gametic relationships

- If genotypes are phased we can build gametic relationships

$$\boldsymbol{w}_{f(i)} = \boldsymbol{w}_{f(i),1} + \boldsymbol{w}_{f(i),2}$$

$$\boldsymbol{w}_{m(i)} = \boldsymbol{w}_{m(i),1} + \boldsymbol{w}_{m(i),2}$$

$$\boldsymbol{w}_{i} = \boldsymbol{w}_{i,1} + \boldsymbol{w}_{i,2}$$

$$\begin{pmatrix} \boldsymbol{w}_{f(i),1}\boldsymbol{w}_{f(i),1}^T & & & & & sym. \\ \boldsymbol{w}_{f(i),2}\boldsymbol{w}_{f(i),1}^T & \boldsymbol{w}_{f(i),2}\boldsymbol{w}_{f(i),2}^T & & & & \\ \boldsymbol{w}_{m(i),1}\boldsymbol{w}_{f(i),1}^T & \boldsymbol{w}_{m(i),1}\boldsymbol{w}_{f(i),2}^T & \boldsymbol{w}_{m(i),1}\boldsymbol{w}_{m(i),1}^T & & & \\ \boldsymbol{w}_{m(i),2}\boldsymbol{w}_{f(i),1}^T & \boldsymbol{w}_{m(i),2}\boldsymbol{w}_{f(i),2}^T & \boldsymbol{w}_{m(i),2}\boldsymbol{w}_{m(i),1}^T & \boldsymbol{w}_{m(i),2}\boldsymbol{w}_{m(i),2}^T & & \\ \boldsymbol{w}_{i,1}\boldsymbol{w}_{f(i),1}^T & \boldsymbol{w}_{i,1}\boldsymbol{w}_{f(i),2}^T & \boldsymbol{w}_{i,1}\boldsymbol{w}_{m(i),1}^T & \boldsymbol{w}_{i,1}\boldsymbol{w}_{m(i),2}^T & \boldsymbol{w}_{i,1}\boldsymbol{w}_{i,1}^T & \\ \boldsymbol{w}_{i,2}\boldsymbol{w}_{f(i),1}^T & \boldsymbol{w}_{i,2}\boldsymbol{w}_{f(i),2}^T & \boldsymbol{w}_{i,2}\boldsymbol{w}_{m(i),1}^T & \boldsymbol{w}_{i,2}\boldsymbol{w}_{m(i),2}^T & \boldsymbol{w}_{i,2}\boldsymbol{w}_{i,1}^T & \boldsymbol{w}_{i,2}\boldsymbol{w}_{i,2}^T \end{pmatrix}$$

→ How much gametes/genomes could deviate or correlate

Genome-based NRM – between & within family

$$\boldsymbol{w}_{f(i)} = \boldsymbol{w}_{f(i),1} + \boldsymbol{w}_{f(i),2}$$
$$\boldsymbol{w}_{m(i)} = \boldsymbol{w}_{m(i),1} + \boldsymbol{w}_{m(i),2}$$
$$\boldsymbol{w}_{i} = \boldsymbol{w}_{i,1} + \boldsymbol{w}_{i,2}$$

- Expected genotype (=parent average) & deviation (=Mendelian sampling)

$$E(\boldsymbol{w}_i) = E\left(\tfrac{1}{2}\boldsymbol{w}_{f(i)}+\tfrac{1}{2}\boldsymbol{w}_{m(i)}+\boldsymbol{w}_i^r\right) = \tfrac{1}{2}\boldsymbol{w}_{f(i)} + \tfrac{1}{2}\boldsymbol{w}_{m(i)}$$

→ How many alt. alleles do we expect from parents (vs. mean)

$$\boldsymbol{w}_i^r = \boldsymbol{w}_i - \left(\tfrac{1}{2}\boldsymbol{w}_{f(i)} + \tfrac{1}{2}\boldsymbol{w}_{m(i)}\right)$$

→ How many more or less alt. alleles did individual get

Genome-based NRM – between & within family

- Expected genotype (=parent average) & deviation (=Mendelian sampling) per genome

$$E\left(\boldsymbol{w}_{i,1}\right) = E\left(\tfrac{1}{2}\boldsymbol{w}_{f(i),1} + \tfrac{1}{2}\boldsymbol{w}_{f(i),2} + \boldsymbol{w}_{i,1}^{r}\right) = \tfrac{1}{2}\boldsymbol{w}_{f(i),1} + \tfrac{1}{2}\boldsymbol{w}_{f(i),2}$$

$$\boldsymbol{w}_{i,1}^{r} = \boldsymbol{w}_{i,1} - \left(\tfrac{1}{2}\boldsymbol{w}_{f(i),1} + \tfrac{1}{2}\boldsymbol{w}_{f(i),2}\right)$$

→ from father

$$E\left(\boldsymbol{w}_{i,2}\right) = E\left(\tfrac{1}{2}\boldsymbol{w}_{m(i),1} + \tfrac{1}{2}\boldsymbol{w}_{m(i),2} + \boldsymbol{w}_{i,1}^{r}\right) = \tfrac{1}{2}\boldsymbol{w}_{m(i),1} + \tfrac{1}{2}\boldsymbol{w}_{m(i),2}$$

$$\boldsymbol{w}_{i,2}^{r} = \boldsymbol{w}_{i,2} - \left(\tfrac{1}{2}\boldsymbol{w}_{m(i),1} + \tfrac{1}{2}\boldsymbol{w}_{m(i),2}\right)$$

→ from mother

Genome-based NRM variants & interpretation

- Centering shifts reference population

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{Xb} + \boldsymbol{W\alpha} + \boldsymbol{e} \\
&= \boldsymbol{Xb} + \boldsymbol{W\alpha} - E(\boldsymbol{W})\boldsymbol{\alpha} + E(\boldsymbol{W})\boldsymbol{\alpha} + \boldsymbol{e} \\
&= \boldsymbol{Xb} + (\boldsymbol{W} - E(\boldsymbol{W}))\boldsymbol{\alpha} + E(\boldsymbol{W})\boldsymbol{\alpha} + \boldsymbol{e} \\
&= \boldsymbol{Xb} + (\boldsymbol{W} - E(\boldsymbol{W}))\boldsymbol{\alpha} + c + \boldsymbol{e} \\
&= (\boldsymbol{Xb} + c) + (\boldsymbol{W} - E(\boldsymbol{W}))\boldsymbol{\alpha} + \boldsymbol{e} \\
&= \boldsymbol{Xb}^c + \boldsymbol{W}^c\boldsymbol{\alpha} + \boldsymbol{e}
\end{aligned}
$$

Genome-based NRM variants & interpretation

- Scaling changes variance meaning

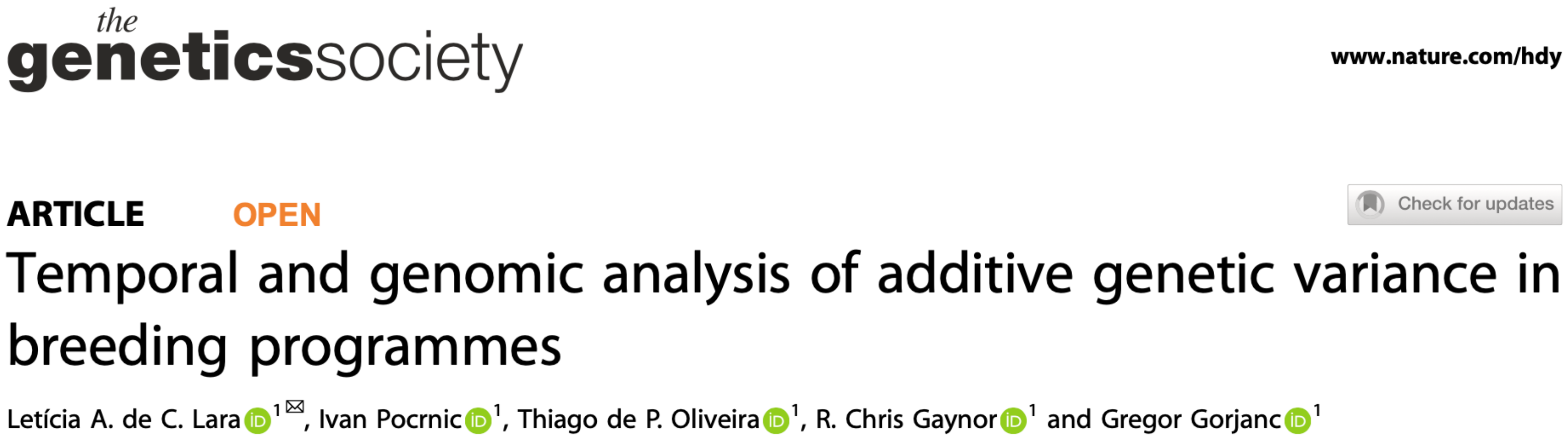
$$
\begin{aligned}
Var(\boldsymbol{a}) &= Var(\boldsymbol{W\alpha}) \\
&= \boldsymbol{WW}^T\sigma_\alpha^2 \\
&= \boldsymbol{WW}^T\sigma_\alpha^2 k\frac{1}{k} \\
&= \frac{\boldsymbol{WW}^T}{\boldsymbol{k}}\sigma_\alpha^2 k \\
&= \boldsymbol{G}\sigma_a^{2*}
\end{aligned}
$$

$$k = \sum 2p_i q_i$$

$$\sigma_a^{2*} = \sigma_\alpha^2 \sum 2p_i q_i$$

- Depending on k we can get very different estimates of σ_a^{2*} (genomic variance)
- Many pedigree and genomic variance comparisons may be dubious?

Flexible (temporal and genomic) analysis of genetic variation



the geneticssociety

www.nature.com/hdy

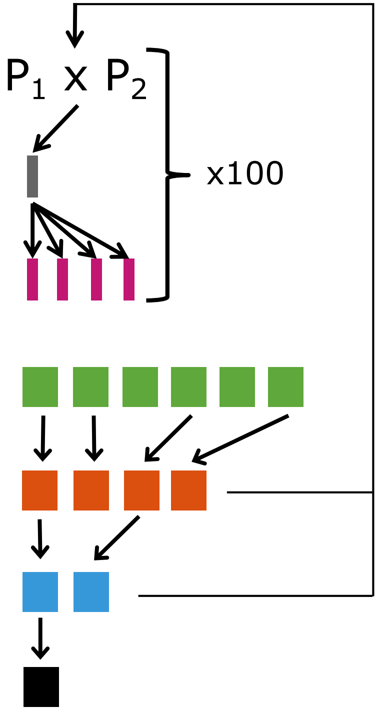
ARTICLE OPEN

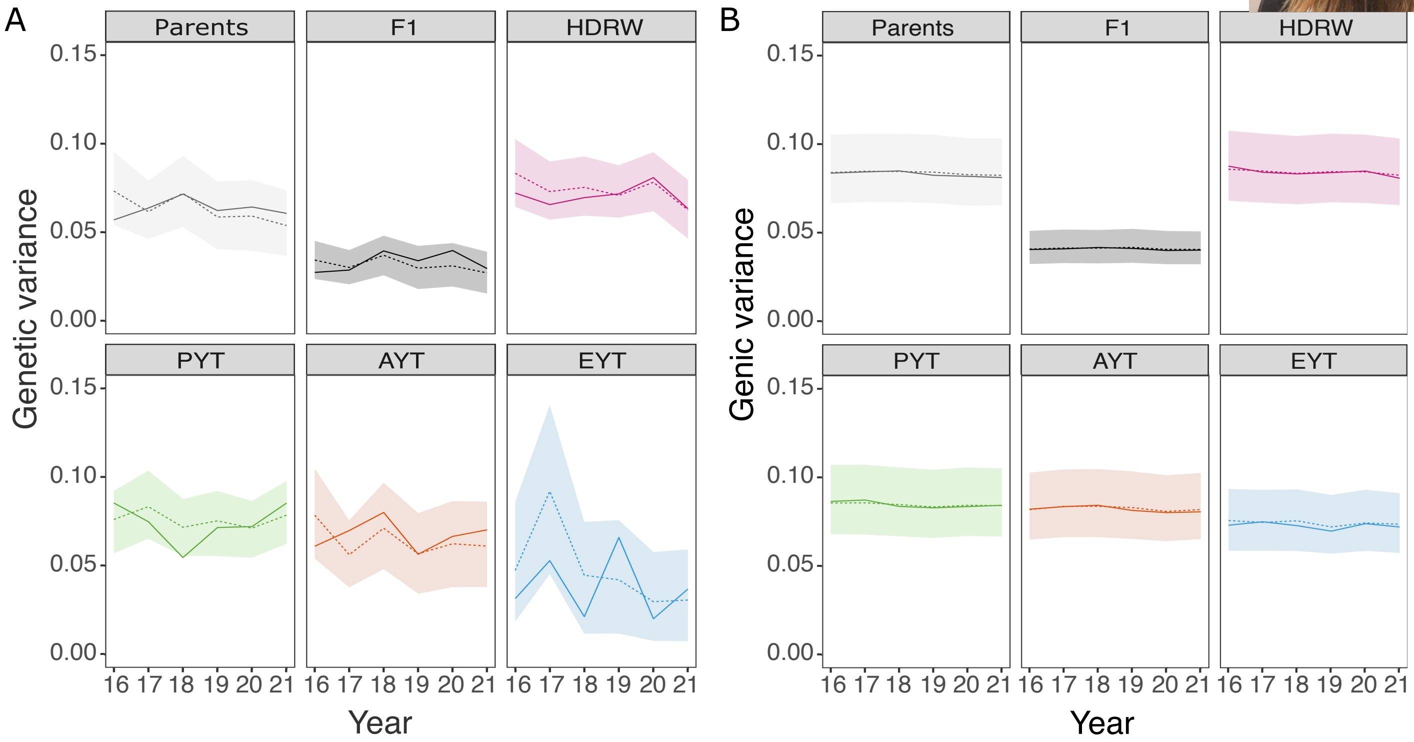
Check for updates

Temporal and genomic analysis of additive genetic variance in breeding programmes

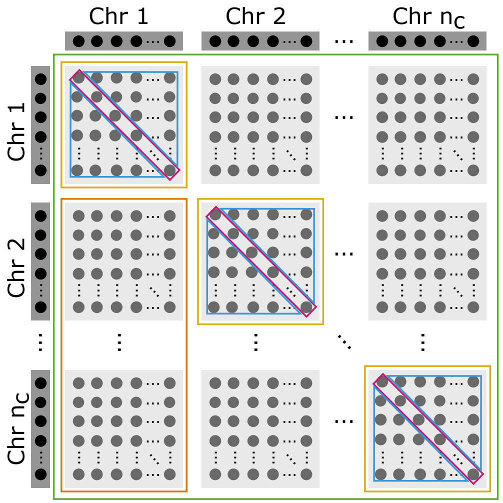
Letícia A. de C. Lara[1] ✉, Ivan Pocrnic[1], Thiago de P. Oliveira[1], R. Chris Gaynor[1] and Gregor Gorjanc[1]

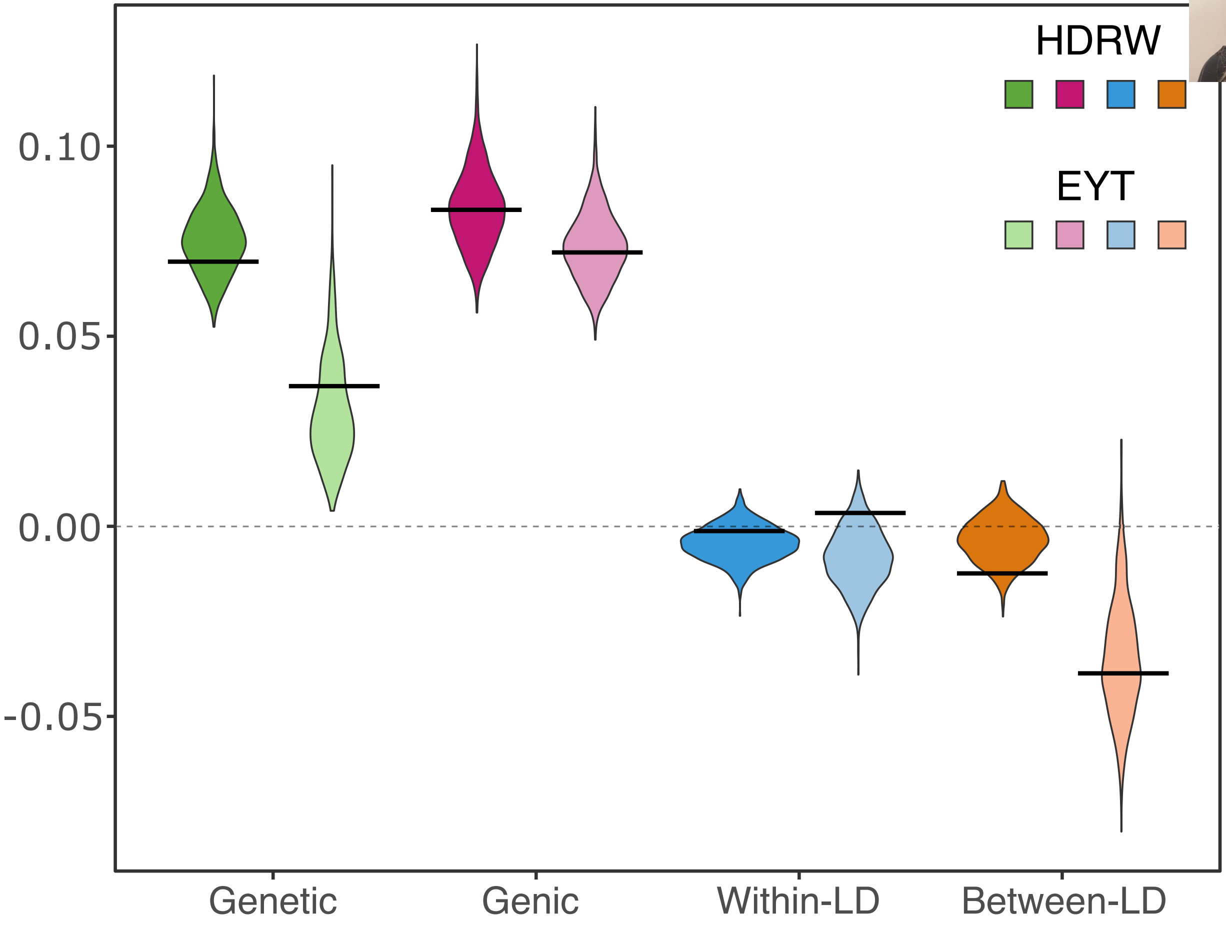
Temporal analysis of genetic variation



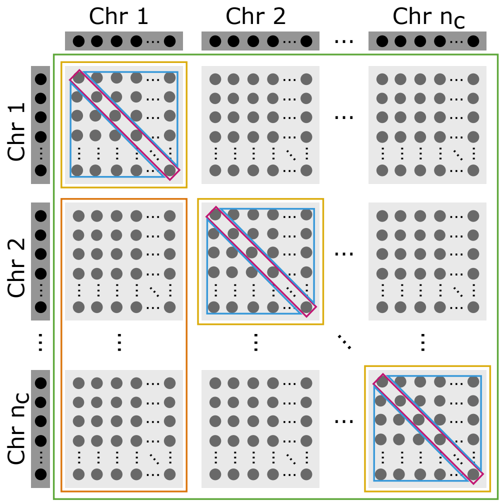


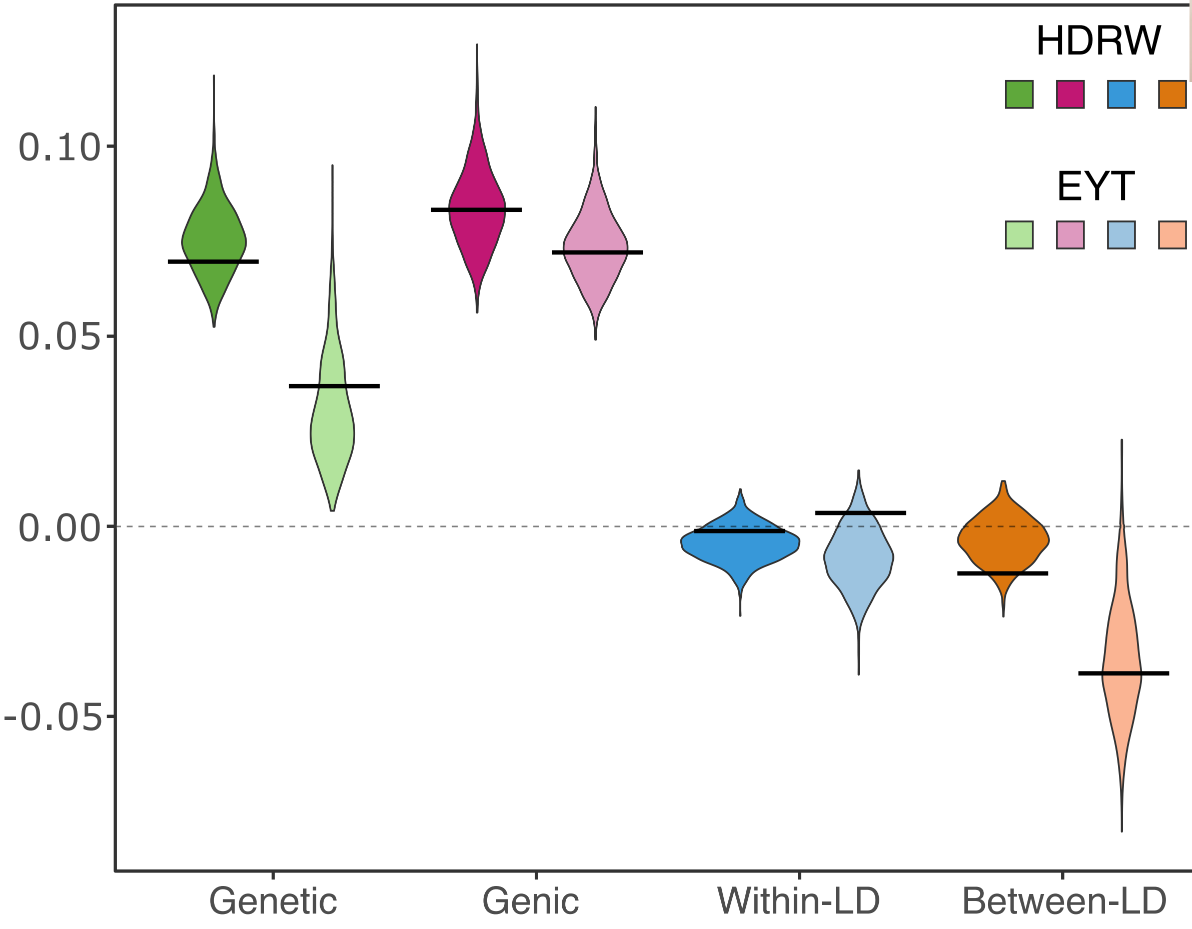
Genomic analysis of genetic variation





Genomic analysis of genetic variation





Topics not covered

- “Bayesian models” – different assumptions about marker effects & commonly approached with methods used in Bayesian statistics (MCMC/VB)
- Single-step GBLUP (ssGBLUP and variants) – combining all phenotype, pedigree, and genomic data
- ”APY”/SVD/… – approximations for large-scale
- Non-additive genetic or other effects
(note that α captures a bit of dominance, epistasis, GxE, …)

Limitations with current genome-based models?

- Markers vs. QTL
- Admixed populations, multiple populations, …
- Whole-genome sequence data
- …

Learning objectives

- Understand limitations of estimates from the pedigree-based model → why we would need genome-based model

- Understand how to combine phenotype information from all relatives connected via genomic data

- Practice inference of breeding values with the genome-based model
  - simple cases using R matrix algebra
  - using other packages

Questions?!

THE UNIVERSITY *of* EDINBURGH

ROSLIN

UKRI Biotechnology and Biological Sciences Research Council

THE ROYAL SOCIETY

Genome-based genetic evaluation

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-07

HLab