Introduction to ssGBLUP

Jorge Hidalgo

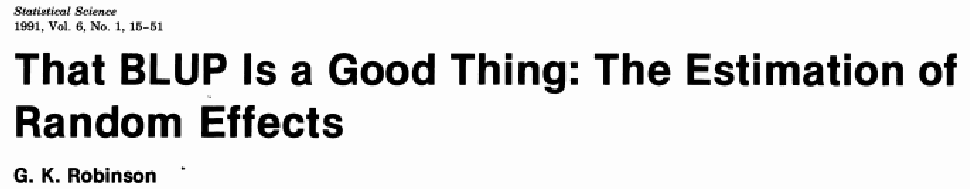
Animal Breeding and Genetics Group

College of Agricultural & Environmental Sciences

UNIVERSITY OF GEORGIA

Pedigree BLUP

BLUP-based methods



Statistical Science
1991, Vol. 6, No. 1, 15–51

That BLUP Is a Good Thing: The Estimation of Random Effects

G. K. Robinson

Our genetic model is: $$\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \mathbf{e}$$

$\mathbf{y}$ = vector of phenotypes

$\mathbf{Xb}$ = matrix relating $\mathbf{y}$ with fixed effects in $\mathbf{b}$

$\mathbf{Wu}$ = matrix relating $\mathbf{y}$ with random effects in $\mathbf{u}$, $Var(\mathbf{u}) = \mathbf{A}\sigma_u^2$

$\mathbf{e}$ = vector of random errors, $Var(\mathbf{e}) = \mathbf{I}\sigma_e^2$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

Best: minimizes MSE

Linear: linear function of the data

Unbiased: $E(u) = E(\hat{u})$

Prediction: for random effects

$$u_i = u_{s\_i} + u_{d\_i}$$

$$p(\mathbf{y}, \mathbf{u}) = p(\mathbf{u}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{u})p(\mathbf{u})$$

Henderson, 1949

Pedigree BLUP

BLUP-based methods

Statistical Science
1991, Vol. 6, No. 1, 15–51

That BLUP Is a Good Thing: The Estimation of Random Effects

G. K. Robinson

- Unbalanced data and information from relatives

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

$$\lambda = \frac{1 - h^2}{h^2}$$

h^2 is high

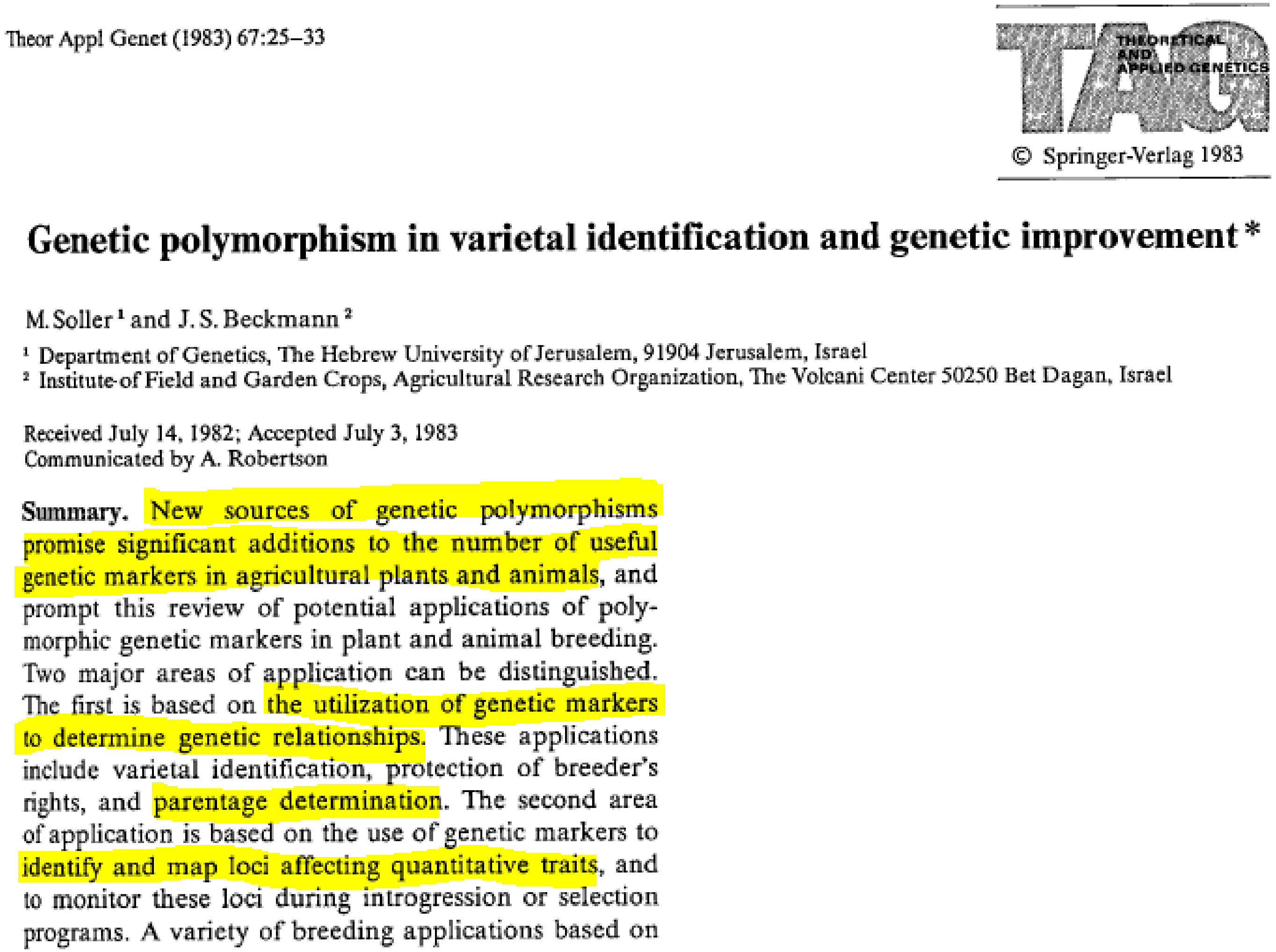
λ goes to zero
$\mathbf{A}^{-1}\lambda$ goes to zero
“Relationships don’t matter”

h^2 is low

λ goes to infinity
$\mathbf{A}^{-1}\lambda$ goes to infinity
“Relationships matter a lot”

Henderson, 1949

BLUP-based methods



Theor Appl Genet (1983) 67:25–33

THEORETICAL AND APPLIED GENETICS

Genetic polymorphism in varietal identification and genetic improvement *

M. Soller [1] and J. S. Beckmann [2]

[1] Department of Genetics, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel
[2] Institute of Field and Garden Crops, Agricultural Research Organization, The Volcani Center 50250 Bet Dagan, Israel

Received July 14, 1982; Accepted July 3, 1983
Communicated by A. Robertson

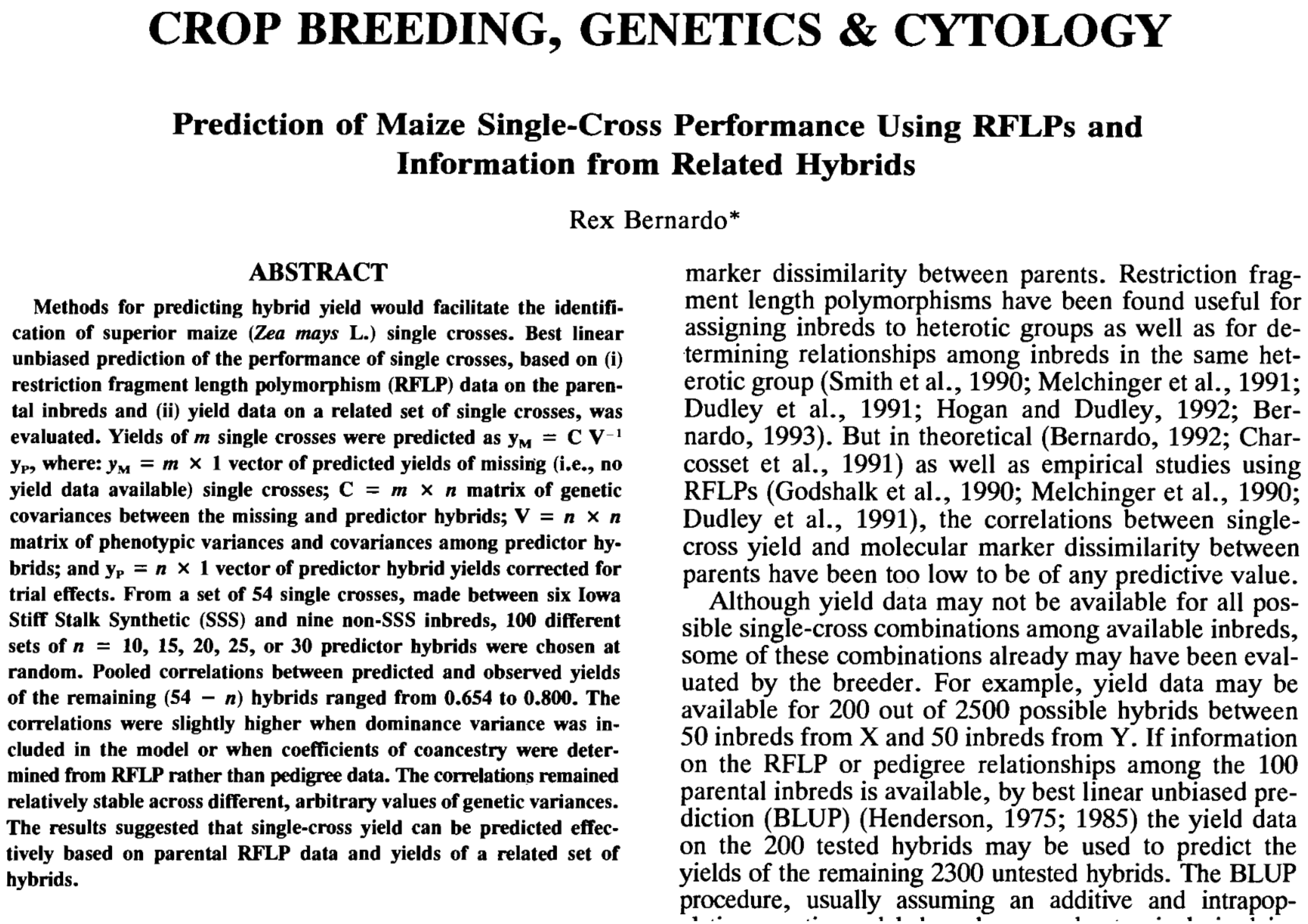
Summary. New sources of genetic polymorphisms promise significant additions to the number of useful genetic markers in agricultural plants and animals, and prompt this review of potential applications of polymorphic genetic markers in plant and animal breeding. Two major areas of application can be distinguished. The first is based on the utilization of genetic markers to determine genetic relationships. These applications include varietal identification, protection of breeder's rights, and parentage determination. The second area of application is based on the use of genetic markers to identify and map loci affecting quantitative traits, and to monitor these loci during introgression or selection programs. A variety of breeding applications based on

- Use of DNA polymorphisms as genetic markers
- Construct genetic relationships
- Parentage determination
- Identification of QTL
- RFLP (expensive)

Soller and Beckman, 1982

Genomic BLUP

BLUP-based methods



CROP BREEDING, GENETICS & CYTOLOGY

Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids

Rex Bernardo*

ABSTRACT

Methods for predicting hybrid yield would facilitate the identification of superior maize (*Zea mays* L.) single crosses. Best linear unbiased prediction of the performance of single crosses, based on (i) restriction fragment length polymorphism (RFLP) data on the parental inbreds and (ii) yield data on a related set of single crosses, was evaluated. Yields of *m* single crosses were predicted as $y_M = C V^{-1} y_P$, where: $y_M = m \times 1$ vector of predicted yields of missing (i.e., no yield data available) single crosses; $C = m \times n$ matrix of genetic covariances between the missing and predictor hybrids; $V = n \times n$ matrix of phenotypic variances and covariances among predictor hybrids; and $y_P = n \times 1$ vector of predictor hybrid yields corrected for trial effects. From a set of 54 single crosses, made between six Iowa Stiff Stalk Synthetic (SSS) and nine non-SSS inbreds, 100 different sets of $n = 10, 15, 20, 25$, or 30 predictor hybrids were chosen at random. Pooled correlations between predicted and observed yields of the remaining $(54 - n)$ hybrids ranged from 0.654 to 0.800. The correlations were slightly higher when dominance variance was included in the model or when coefficients of coancestry were determined from RFLP rather than pedigree data. The correlations remained relatively stable across different, arbitrary values of genetic variances. The results suggested that single-cross yield can be predicted effectively based on parental RFLP data and yields of a related set of hybrids.

marker dissimilarity between parents. Restriction fragment length polymorphisms have been found useful for assigning inbreds to heterotic groups as well as for determining relationships among inbreds in the same heterotic group (Smith et al., 1990; Melchinger et al., 1991; Dudley et al., 1991; Hogan and Dudley, 1992; Bernardo, 1993). But in theoretical (Bernardo, 1992; Charcosset et al., 1991) as well as empirical studies using RFLPs (Godshalk et al., 1990; Melchinger et al., 1990; Dudley et al., 1991), the correlations between single-cross yield and molecular marker dissimilarity between parents have been too low to be of any predictive value.

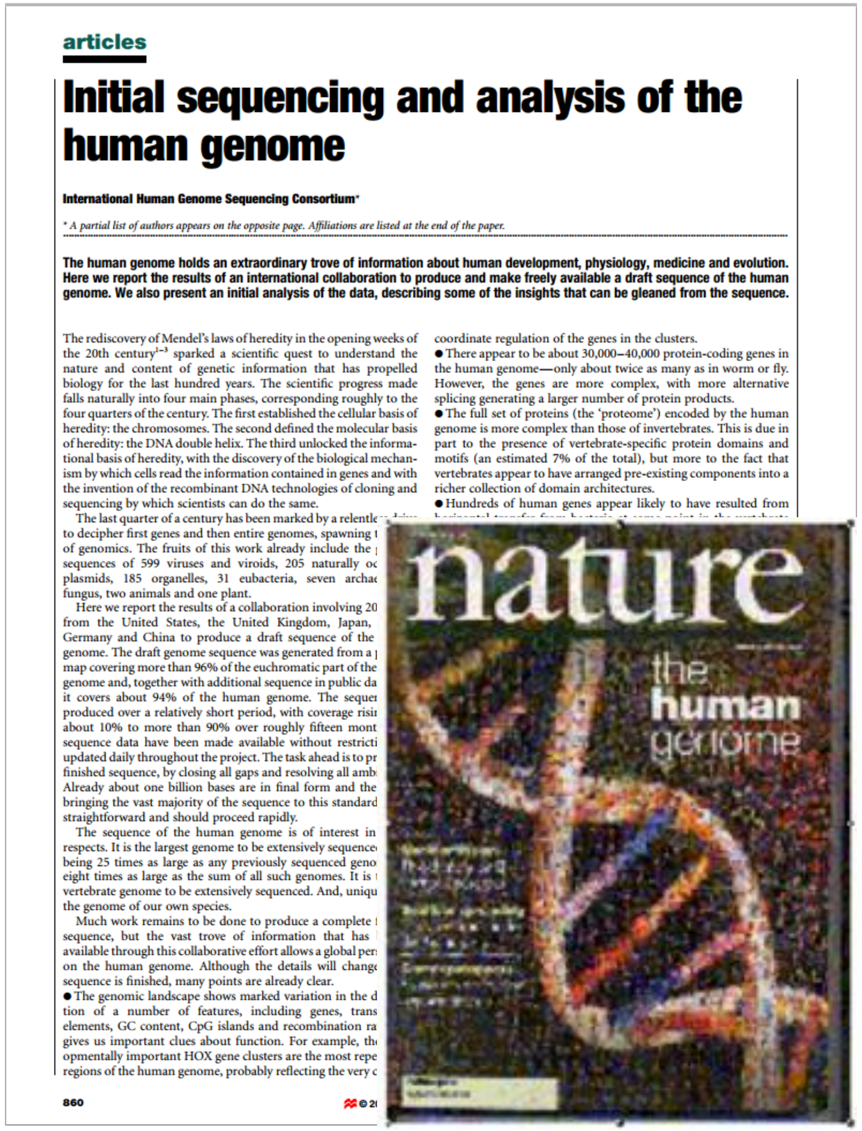
Although yield data may not be available for all possible single-cross combinations among available inbreds, some of these combinations already may have been evaluated by the breeder. For example, yield data may be available for 200 out of 2500 possible hybrids between 50 inbreds from X and 50 inbreds from Y. If information on the RFLP or pedigree relationships among the 100 parental inbreds is available, by best linear unbiased prediction (BLUP) (Henderson, 1975; 1985) the yield data on the 200 tested hybrids may be used to predict the yields of the remaining 2300 untested hybrids. The BLUP procedure, usually assuming an additive and intrapop-

- Use of DNA polymorphisms as genetic markers
- Construct genetic relationships
- Parentage determination
- Identification of QTL
- RFLP (expensive)

Bernardo, 1994

Genomic BLUP

BLUP-based methods



articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentle
to decipher first genes and then entire genomes, spawning t
of genomics. The fruits of this work already include the
sequences of 599 viruses and viroids, 205 naturally oc
plasmids, 185 organelles, 31 eubacteria, seven archae
fungus, two animals and one plant.

Here we report the results of a collaboration involving 20
from the United States, the United Kingdom, Japan,
Germany and China to produce a draft sequence of the
genome. The draft genome sequence was generated from a
map covering more than 96% of the euchromatic part of the
genome and, together with additional sequence in public da
it covers about 94% of the human genome. The sequen
produced over a relatively short period, with coverage risin
about 10% to more than 90% over roughly fifteen mont
sequence data have been made available without restricti
updated daily throughout the project. The task ahead is to pr
finished sequence, by closing all gaps and resolving all ambi
Already about one billion bases are in final form and the
bringing the vast majority of the sequence to this standard
straightforward and should proceed rapidly.

The sequence of the human genome is of interest in
respects. It is the largest genome to be extensively sequence
being 25 times as large as any previously sequenced geno
eight times as large as the sum of all such genomes. It is
vertebrate genome to be extensively sequenced. And, uniqu
the genome of our own species.

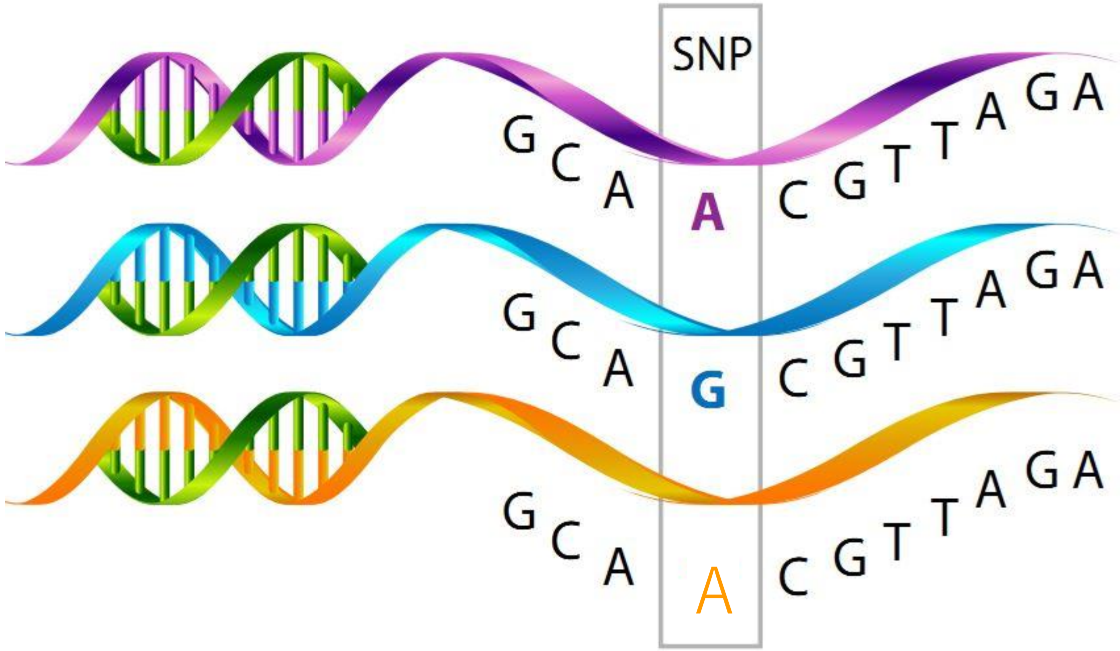
Much work remains to be done to produce a complete
sequence, but the vast trove of information that has
available through this collaborative effort allows a global pers
on the human genome. Although the details will change
sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the d
tion of a number of features, including genes, trans
elements, GC content, CpG islands and recombination ra
gives us important clues about function. For example, th
opmentally important HOX gene clusters are the most repe
regions of the human genome, probably reflecting the very c

coordinate regulation of the genes in the clusters.

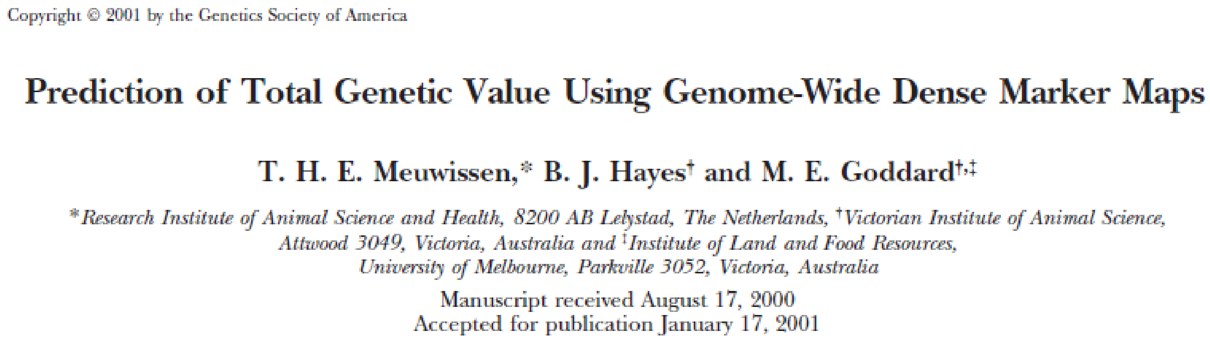
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from

860 © 20



http://neuroendoimmune.files.wordpress.com/2014/03/snp.png

Mutation < 1% < SNP



Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps

T. H. E. Meuwissen,* B. J. Hayes† and M. E. Goddard†,‡

*Research Institute of Animal Science and Health, 8200 AB Lelystad, The Netherlands, †Victorian Institute of Animal Science, Attwood 3049, Victoria, Australia and ‡Institute of Land and Food Resources, University of Melbourne, Parkville 3052, Victoria, Australia

Manuscript received August 17, 2000
Accepted for publication January 17, 2001

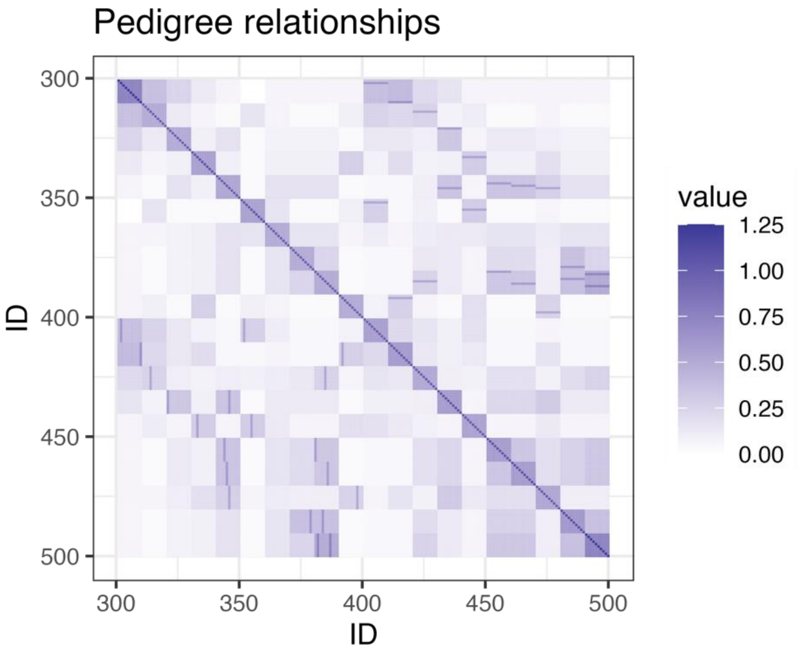
Genomic BLUP

BLUP-based methods

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

Only for genotyped animals

- Better Mendelian sampling tracking



(G)

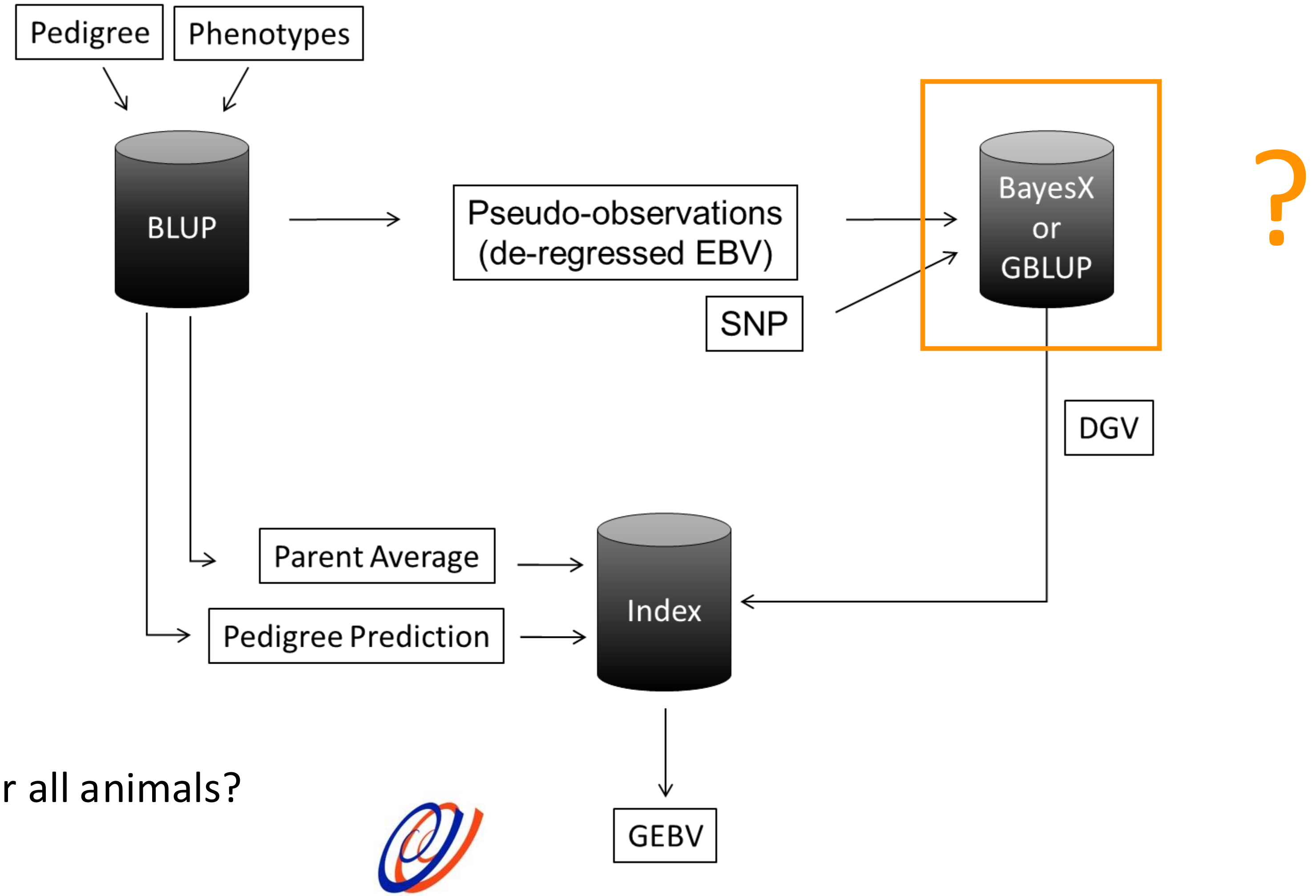
Bernardo, 1994
Nejati-Javaremi et al. (1997)
VanRaden, 2008

Lourenco et al. (2015)

Genomic BLUP

BLUP-based methods

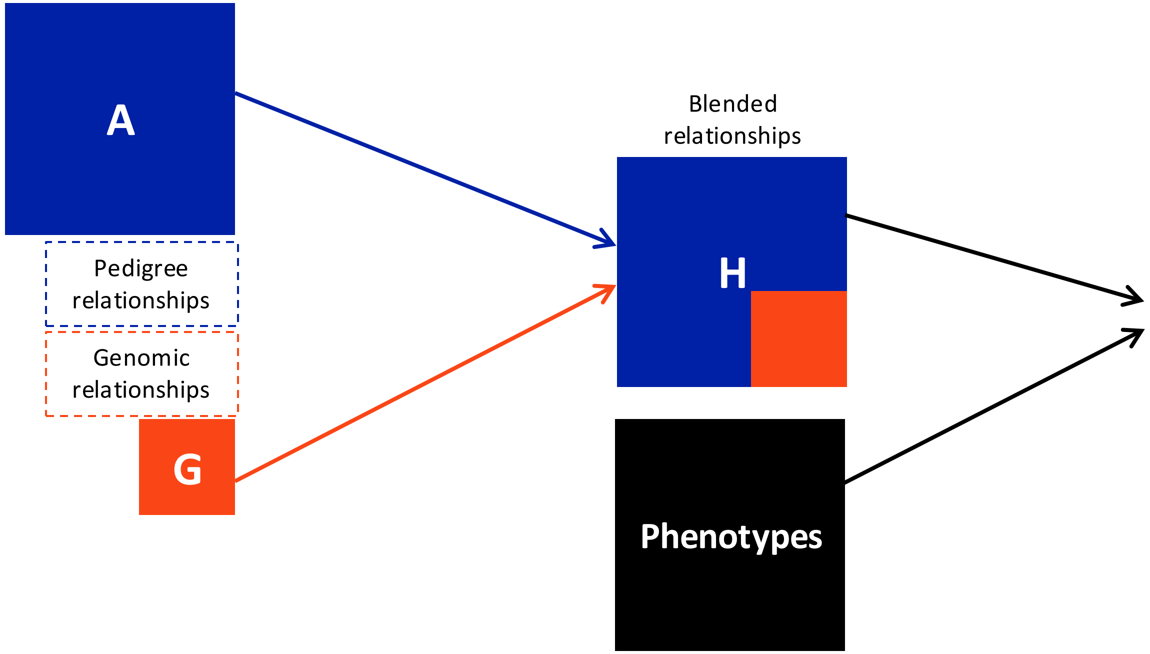
- In practice, not all individuals are genotyped



- How to obtain covariances for all animals?

BLUP-based methods

- In practice, not all individuals are genotyped



$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

- How to obtain covariances for all animals?

Misztal et al. (2009)

Single-step Genomic BLUP

BLUP-based methods

- Genomic evaluation would be simpler if all individuals were genotyped
- What should be done when there are genotyped and non-genotyped individuals?
  - SNPs are capturing relationships
  - Pedigrees give information about relationships
  - Genomic and pedigree relationships can be combined in

Non-genotyped

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

Genotyped

$$\mathbf{H} = \mathbf{A} + \xcancel{\begin{bmatrix} 0 & 0 \\ 0 & -\mathbf{A}_{22} \end{bmatrix}}$$

Misztal et al. (2009)

Single-step Genomic BLUP

BLUP-based methods

- $\mathbf{A}$ is the expectation of realized or observed relationships
- Consider $\mathbf{A}$ as *prior* and $\mathbf{G}$ as *observed* relationships, then construct *posterior* relationships

$$\mathrm{p}(\mathbf{u}_2) = \mathrm{N}(\mathbf{0},\quad \mathbf{G}\sigma_u^2)$$

$$\mathrm{p}(\mathbf{u}_1|\mathbf{u}_2) = \mathrm{N}(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2,\quad \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$$

$$\mathrm{p}(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{p}(\mathbf{u}_1|\mathbf{u}_2)\mathrm{p}(\mathbf{u}_2)$$

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} = \mathbf{H} = \begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

Legarra et al. (2009); Aguilar et al. (2010)

Single-step Genomic BLUP

BLUP-based methods

Error in the prediction

Prediction variance of genotypes for ungenotyped animals

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} = \mathbf{H} = \begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}$$

Covariance

Relationships from genotypes

But … we need $\mathbf{H}^{-1}$

Legarra et al. (2009); Aguilar et al. (2010)

BLUP-based methods

Surprisingly…

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

Legarra et al. (2009); Aguilar et al. (2010)

Single-step Genomic BLUP

BLUP-based methods

Animal	Sire	Dam
1	0	0
2	0	0
3	1	2
4	1	2

A

$$\begin{bmatrix} 1.0 & 0.0 & 0.5 & 0.5 \\ . & 1.0 & 0.5 & 0.5 \\ . & . & 1.0 & 0.5 \\ . & . & . & 1.0 \end{bmatrix}$$

G

$$\begin{bmatrix} 1.0 & 0.52 \\ . & 1.0 \end{bmatrix}$$

H

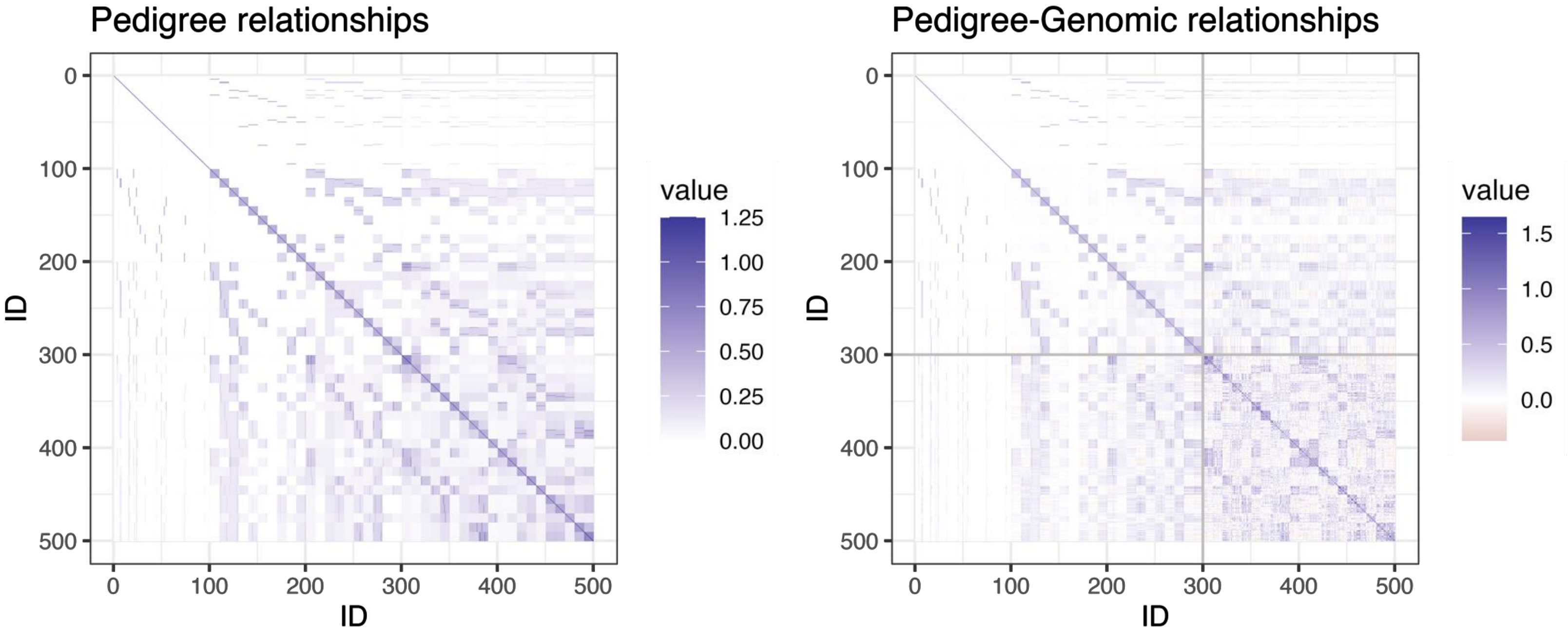
$$\begin{bmatrix} 1.004 & 0.0 & 0.507 & 0.507 \\ . & 1.004 & 0.507 & 0.507 \\ . & . & 1.0 & 0.52 \\ . & . & . & 1.0 \end{bmatrix}$$

Legarra et al. (2009); Aguilar et al. (2010)

BLUP-based methods

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

- Projection of genomic relationships on the rest of individuals
- Bayesian update of $\mathbf{A}$ based on new information from $\mathbf{G}$



BLUP-based methods

- $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$

- **A**
  - Contains expected relationships
  - It is limited by the pedigree depth and completeness
  - Depends on the accuracy of recording pedigrees

- $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{G}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$

- **G**
  - Contains the number of shared alleles between animals weighted by heterozygosity
  - There are no limitations regarding the number of past generations
  - It depends on allele frequency and quality of genomic data

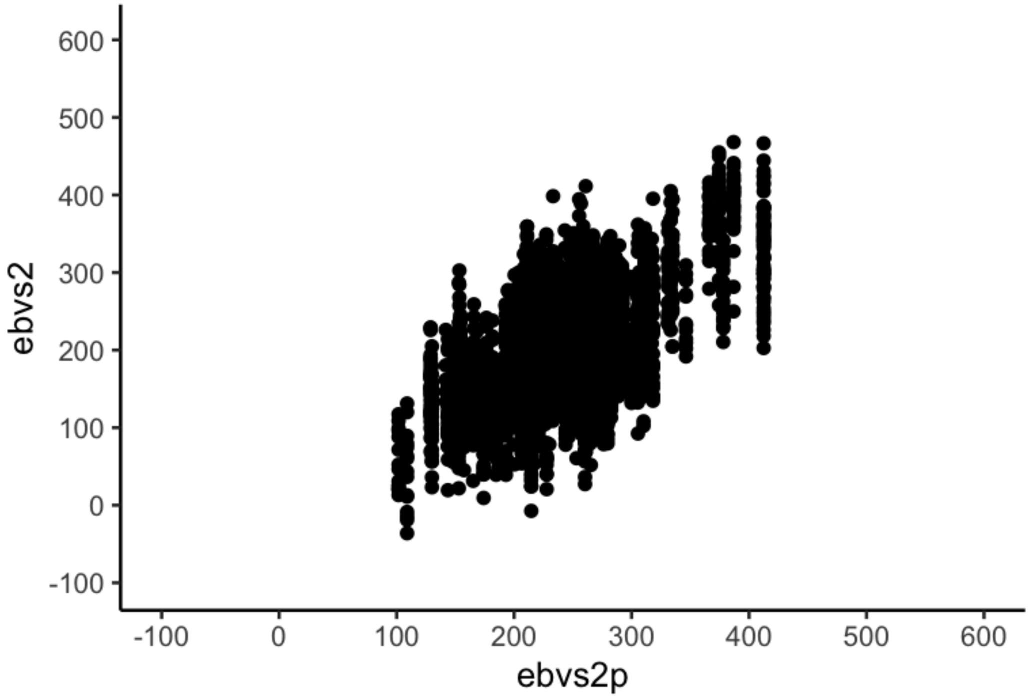
- $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \mathbf{H}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$

- **H**
  - Projection of genomic relationships on the ungenotyped individuals
  - Bayesian updating of $\mathbf{A}$ based on new information from $\mathbf{G}$

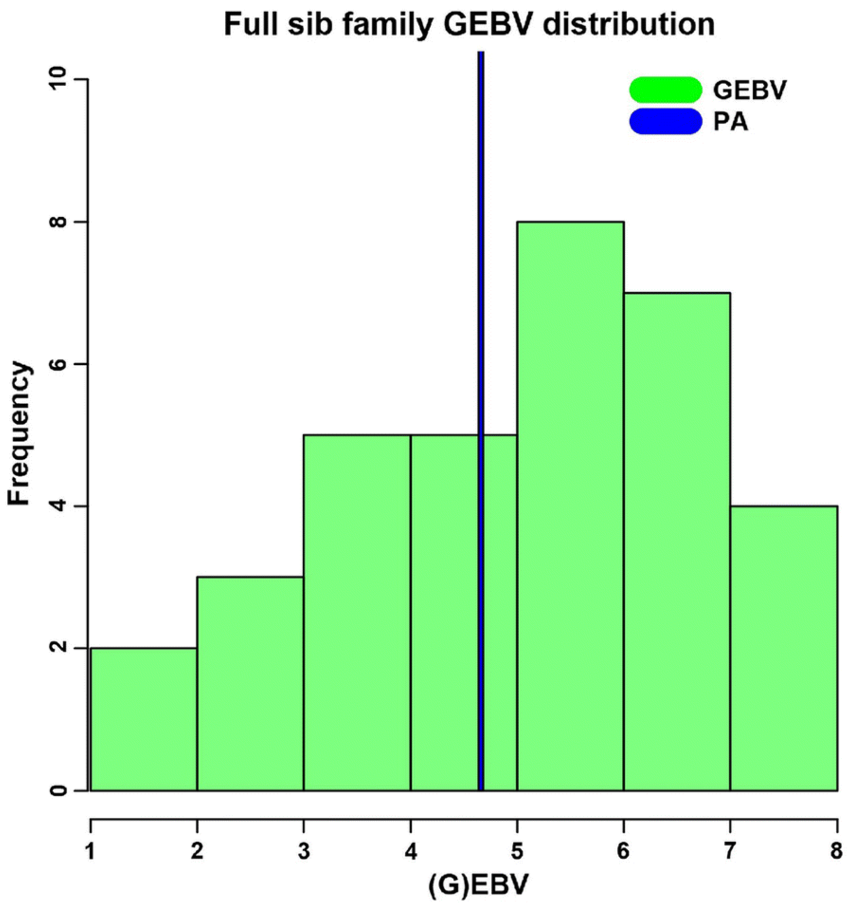
Single-step Genomic BLUP

BLUP-based methods

- Pedigree BLUP



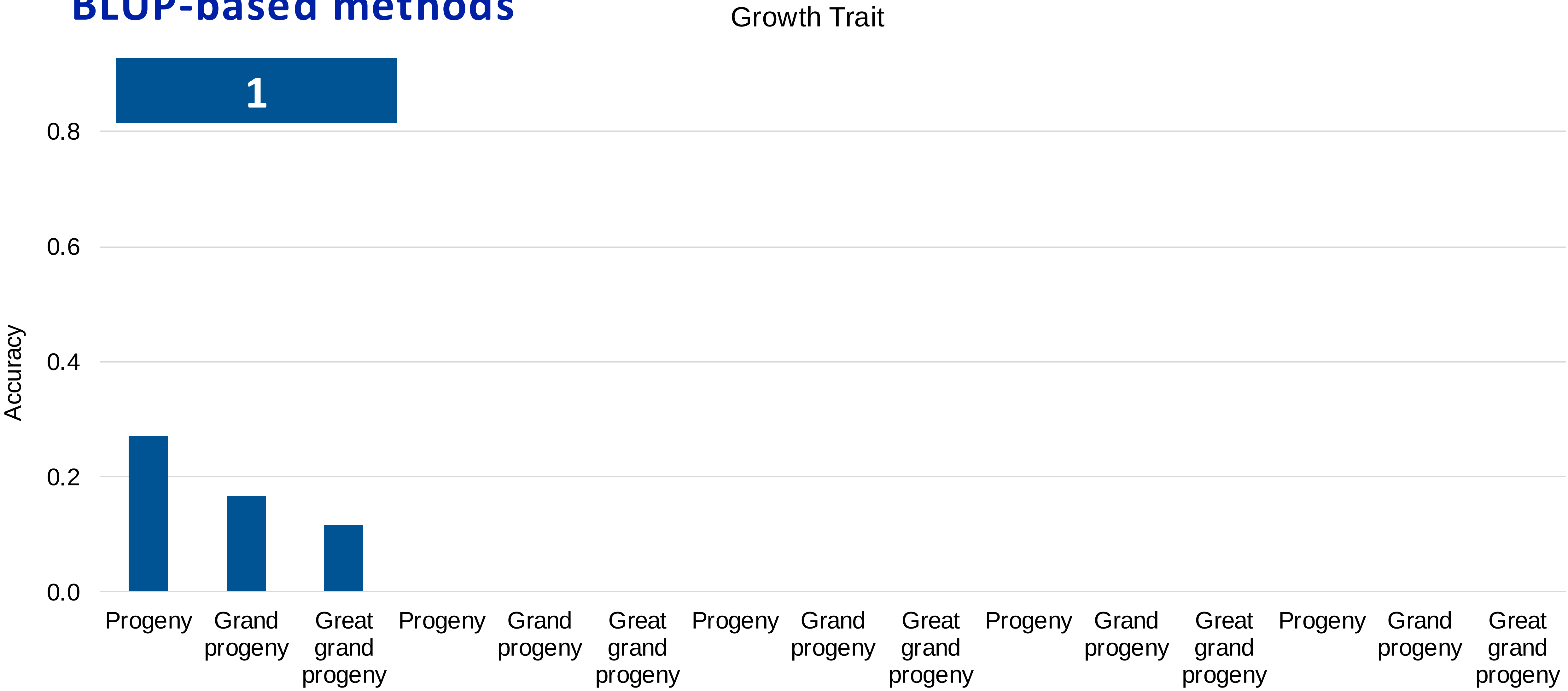
BLUP-based methods



Garcia et al. (2018)

Single-step Genomic BLUP

BLUP-based methods



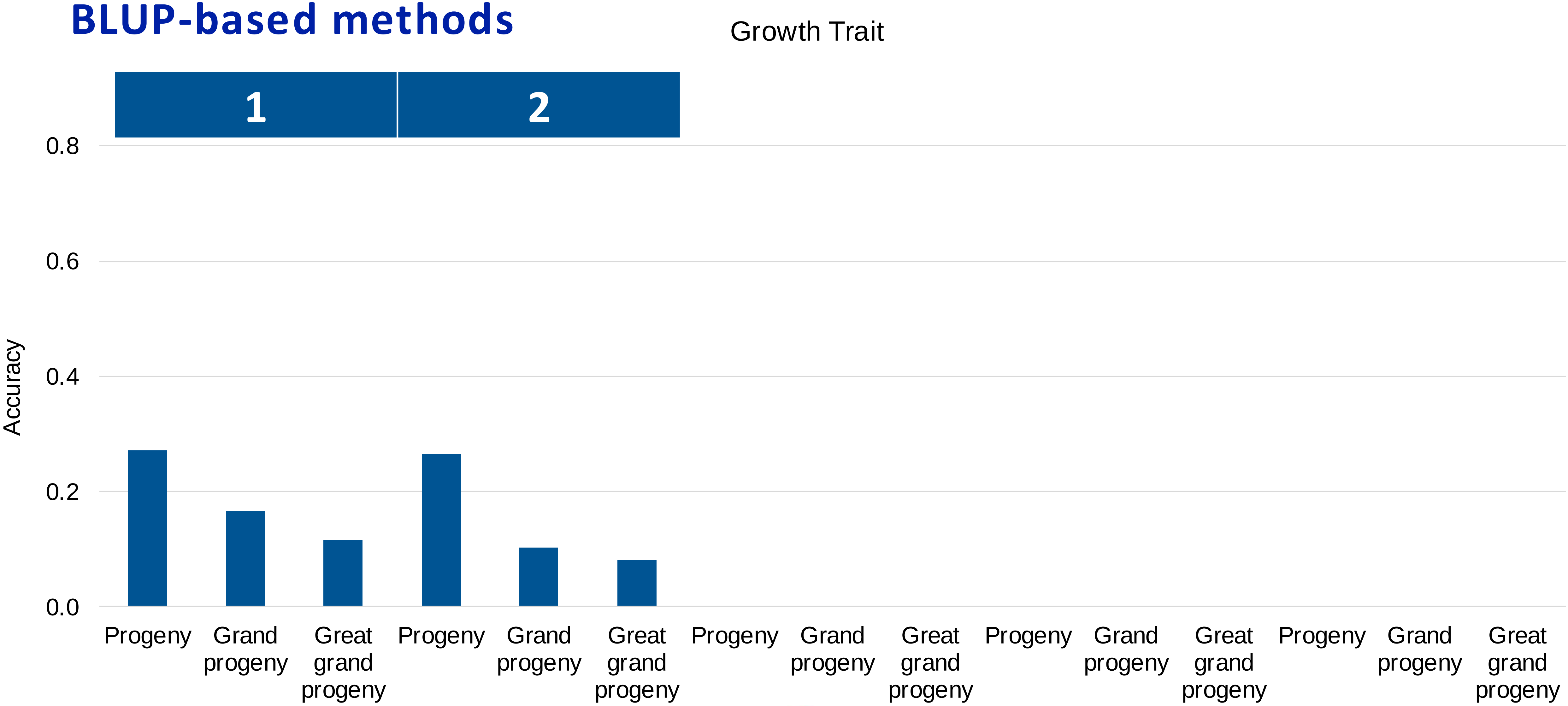
Hidalgo et al. (2022)

Pedigree + Phenotypes

Pedigree + Phenotypes + Genotypes

Single-step Genomic BLUP

BLUP-based methods



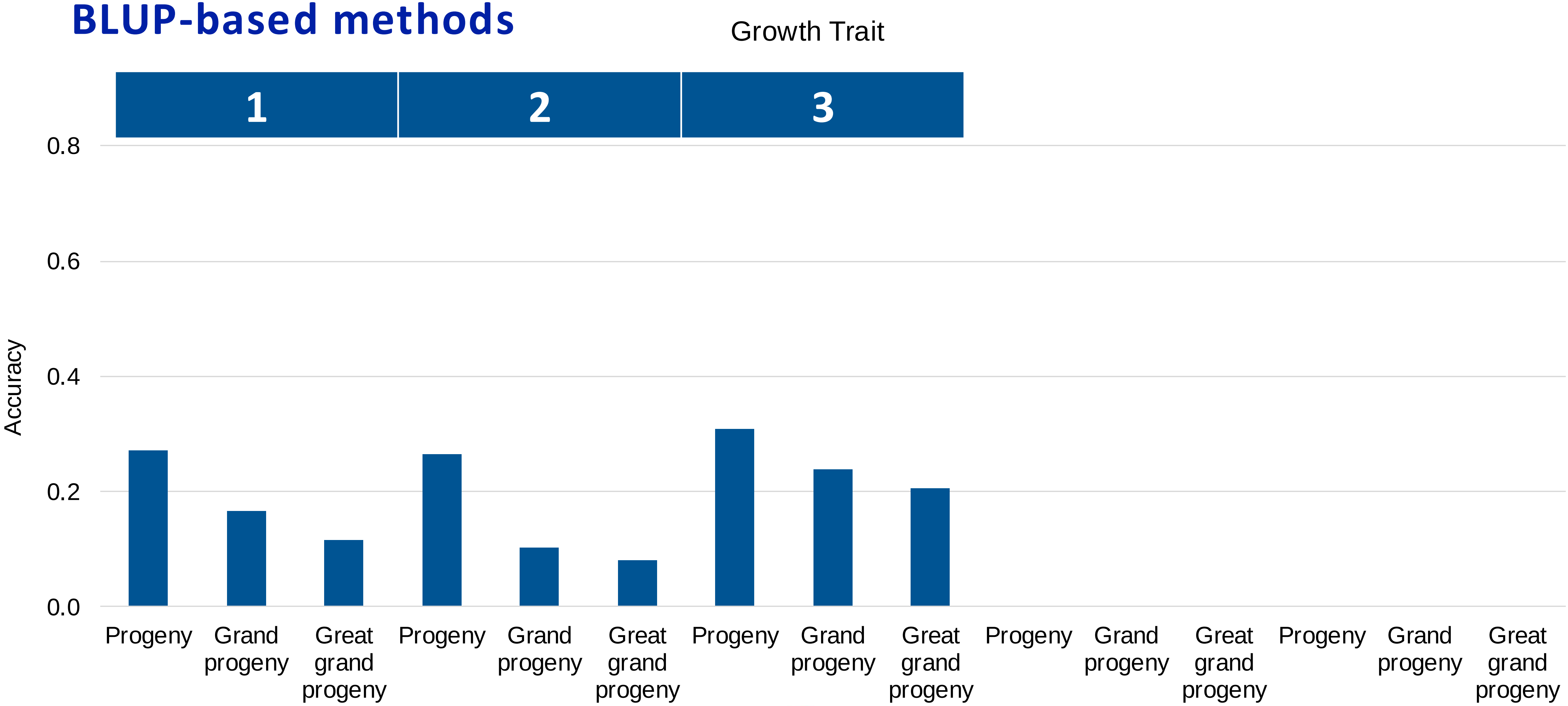
Hidalgo et al. (2018)

Pedigree + Phenotypes

Pedigree + Phenotypes + Genotypes

Single-step Genomic BLUP

BLUP-based methods



Hidalgo et al. (2018)

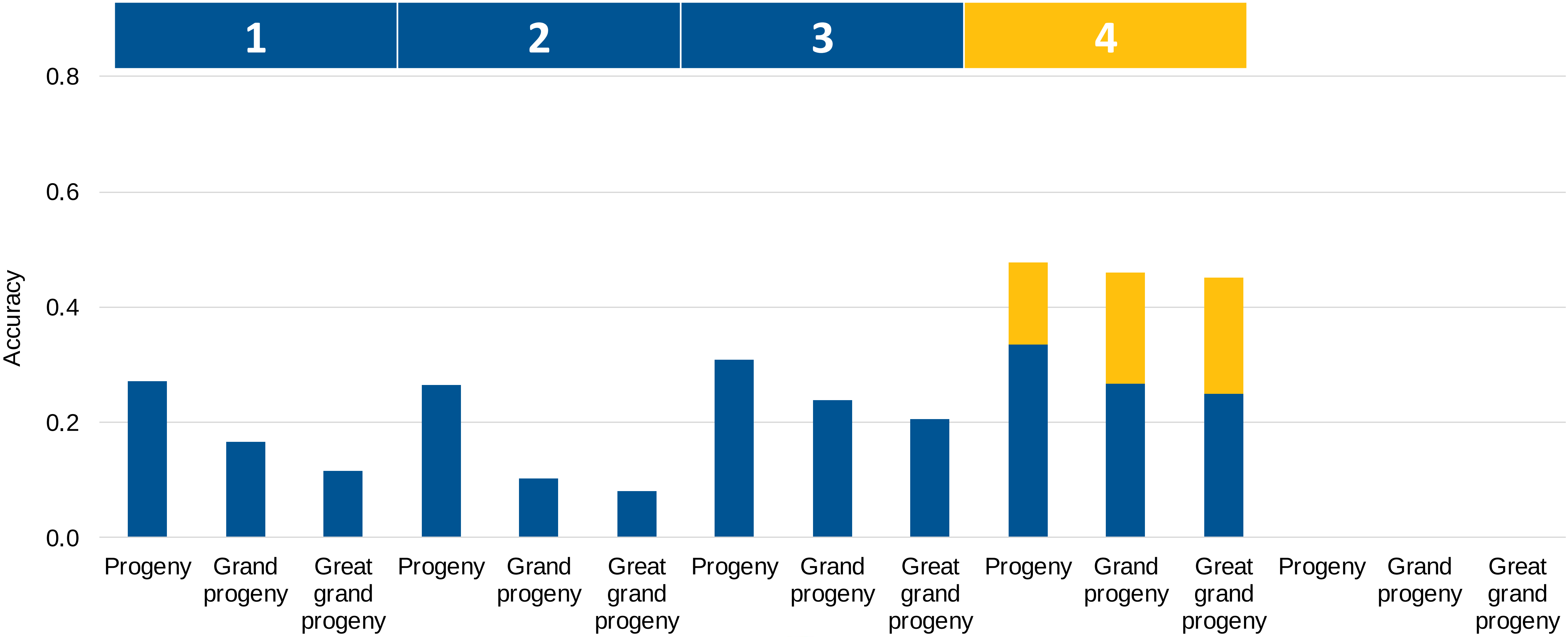
Pedigree + Phenotypes

Pedigree + Phenotypes + Genotypes

Single-step Genomic BLUP

BLUP-based methods

Growth Trait



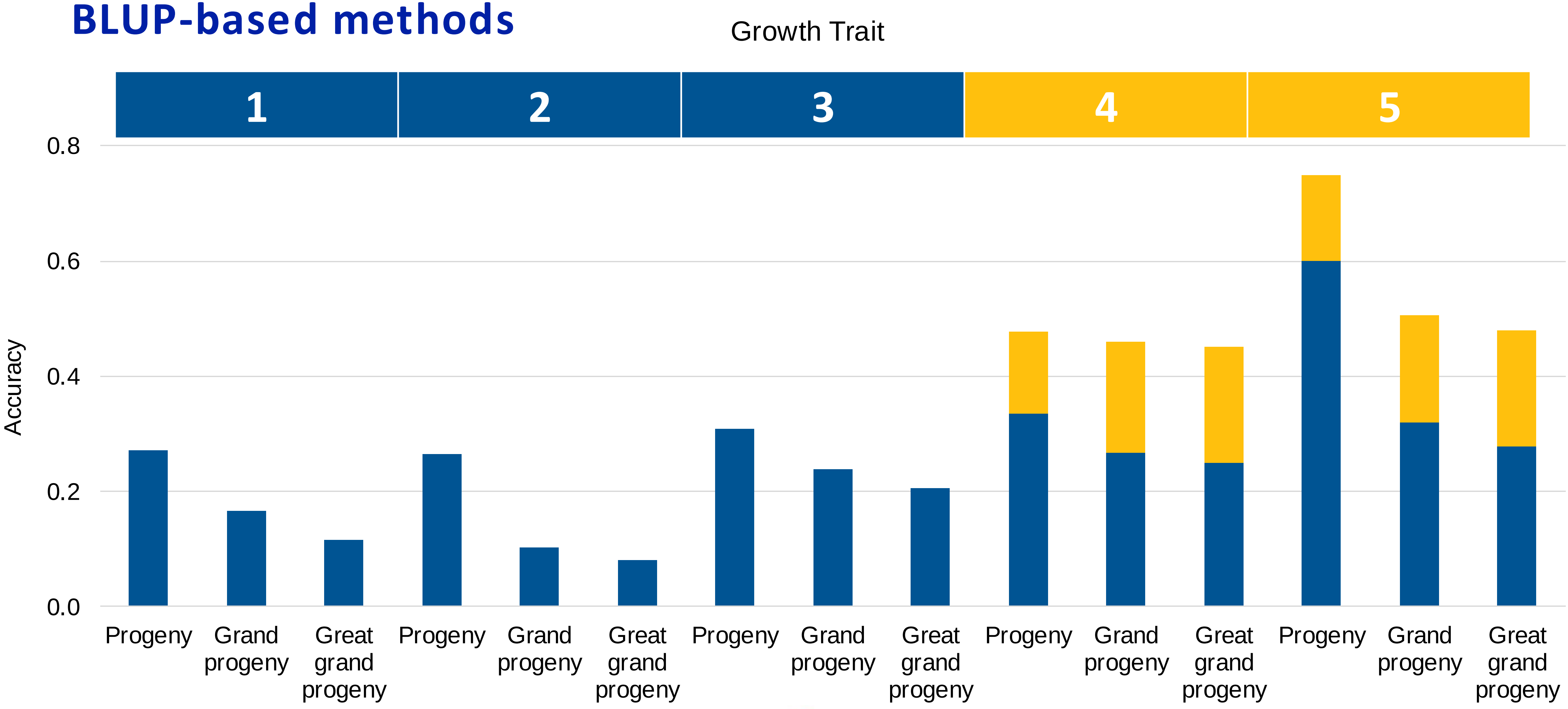
Hidalgo et al. (2022)

Pedigree + Phenotypes

Pedigree + Phenotypes + Genotypes

Single-step Genomic BLUP

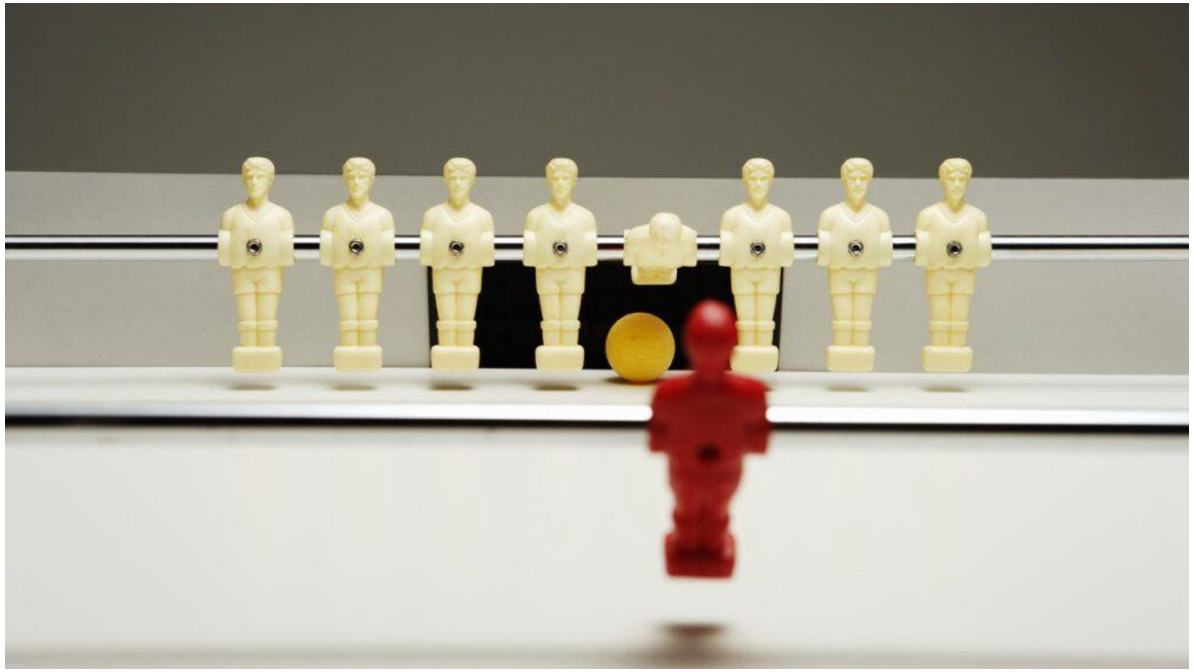
BLUP-based methods



Hidalgo et al. (2022)

Pedigree + Phenotypes

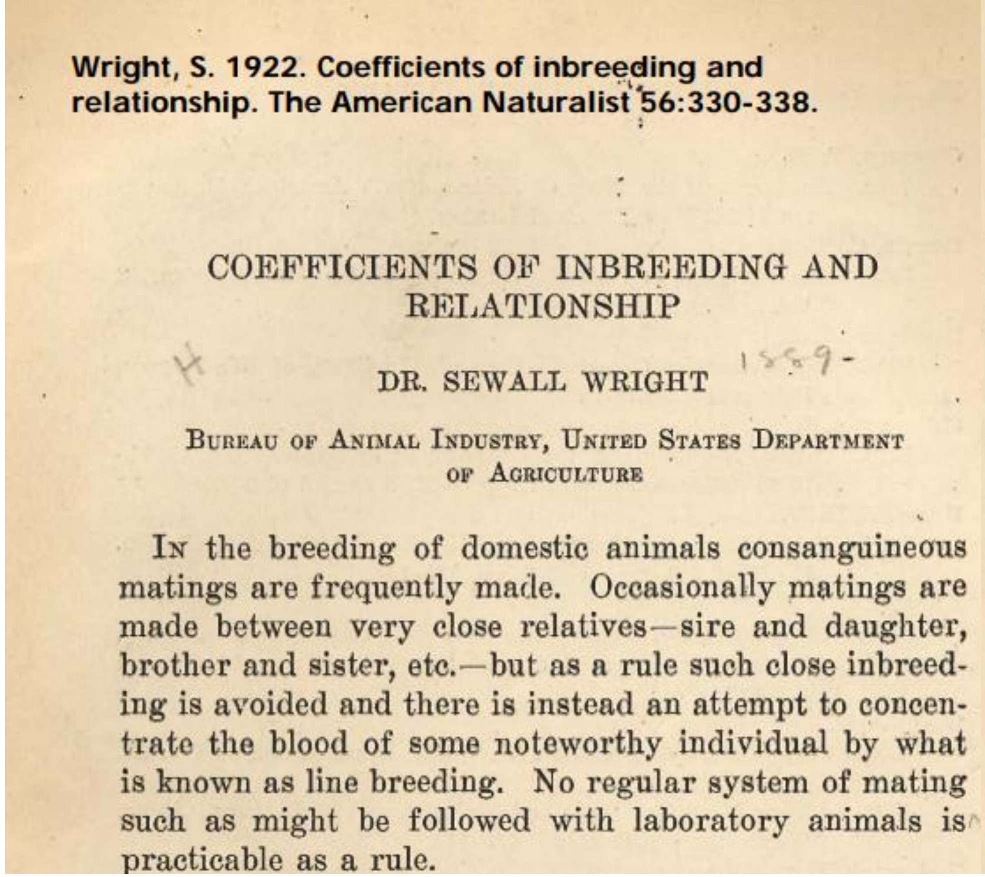
Pedigree + Phenotypes + Genotypes



How to construct G

Realized relationship matrix

- Back to 1922, Wright's relationships matrix ($\mathbf{A}$)
- Relationships were conceived as standardized covariances



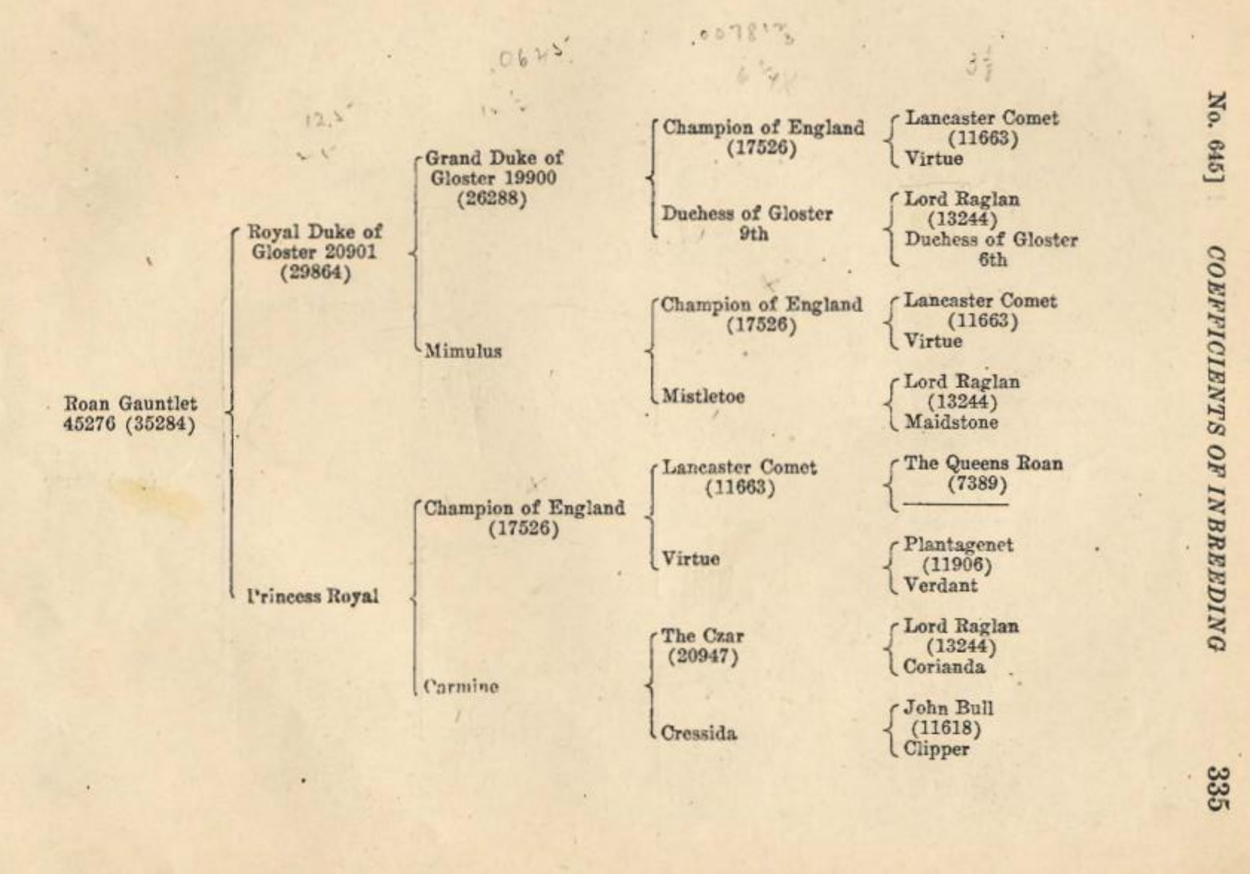
Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56:330-338.

COEFFICIENTS OF INBREEDING AND RELATIONSHIP

DR. SEWALL WRIGHT 1559-

Bureau of Animal Industry, United States Department of Agriculture

In the breeding of domestic animals consanguineous matings are frequently made. Occasionally matings are made between very close relatives—sire and daughter, brother and sister, etc.—but as a rule such close inbreeding is avoided and there is instead an attempt to concentrate the blood of some noteworthy individual by what is known as line breeding. No regular system of mating such as might be followed with laboratory animals is practicable as a rule.



How to construct G

Realized relationship matrix

- How much DNA do two individuals share looking to DNA?
  - Let gene content be coded as 0, 1, and 2 copies of a reference allele
  - Define z_{ij} for locus i, individual j as the gene content
  - The mean of gene content is twice the allele frequency; $\bar{z} = 2p$
  - The variance of the gene content; $\sigma_z^2 = 2p(1-p)$

  - Center z_{ij} subtracting the mean; $z_{ij} - 2p_i$
  - Scale dividing by the sum of variances = $\sum 2p_i(1-p_i)$

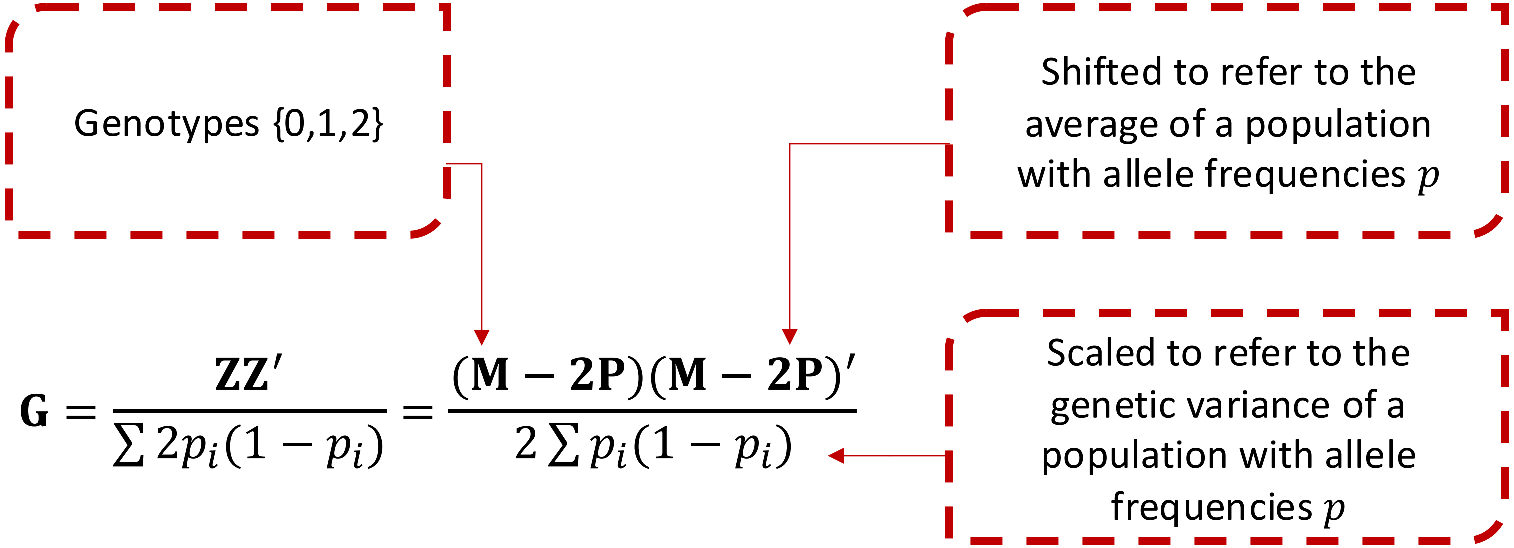
ID1	0221200101202211002222121
ID2	2211212121101211212012121
ID3	1212120020202120120122111

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{\sum 2p_i(1-p_i)}$$

VanRaden, 2008

How to construct $\mathbf{G}$

Realized relationship matrix



Genotypes {0,1,2}

Shifted to refer to the average of a population with allele frequencies p

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{\sum 2p_i(1-p_i)} = \frac{(\mathbf{M}-\mathbf{2P})(\mathbf{M}-\mathbf{2P})'}{2\sum p_i(1-p_i)}$$

Scaled to refer to the genetic variance of a population with allele frequencies p

Tuning and Blending $\mathbf{G}$

Realized relationship matrix

- Tuning scales $\mathbf{G}$ to $\mathbf{A}_{22}$ to refer to the same genetic base
- $\mathrm{p}(\mathbf{u}_2) = \mathrm{N}(\mathbf{0},\quad \mathbf{G}\sigma_u^2)$
- If the population is undergoing selection, the mean is not 0
- Different genetic variance in genotyped and ungenotyped animals
- Accounts for the selection, improves accuracy, and reduces bias
    - $(\overline{diag(\mathbf{G})})b + a = \ (\overline{diag(\mathbf{A}_{22})})$
    - $a + b\overline{\mathbf{G}} = \overline{\mathbf{A}_{22}}$
    - $\mathbf{G}_{tun} = a + b\mathbf{G}_{\mathrm{o}}$
- Blending avoids singularity; the procedure consists of a weighted sum of $\mathbf{G}_{\mathrm{o}}$ and a positive-definitive matrix
- Improves convergence
    - $\mathbf{G} = \alpha\mathbf{G}_{tun} + \beta\mathbf{A}_{22}$
        - This also assigns part of the genetic variance to pedigrees

Christensen et al. (2012)

Importance of inbreeding

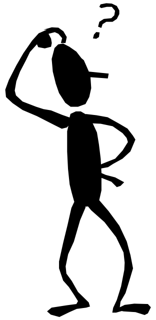
Realized relationship matrix

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

The $\mathbf{G}$ matrix computed using VanRaden's method considers inbreeding, so $\mathbf{G}^{-1}$ does. Therefore, $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ should be constructed considering inbreeding to avoid inflation in the estimated breeding values

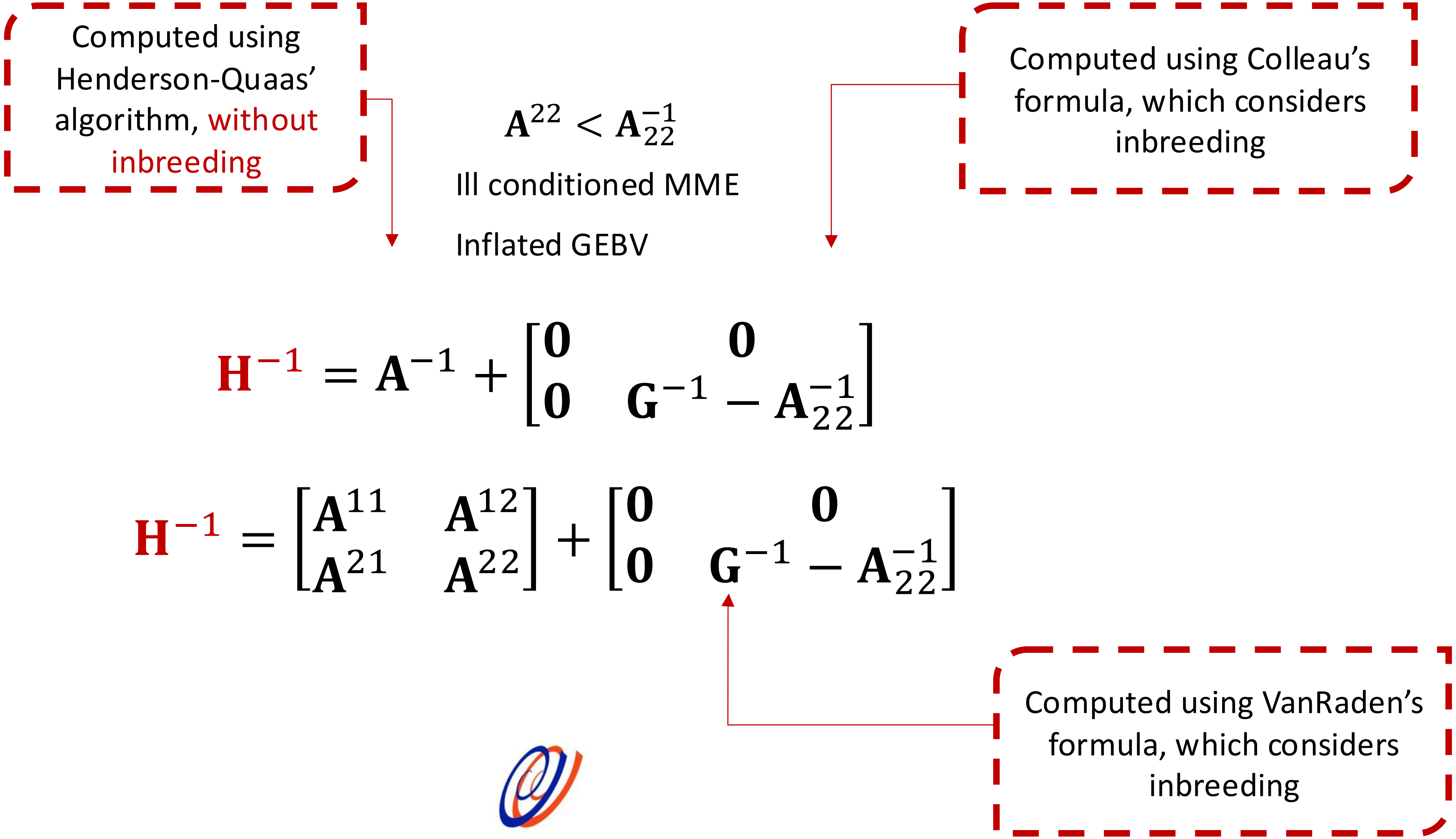
- Pocrnic et al. (2016)
- 10 generations: 5 males mated 12.5k females
- 138k pedigree | 75k genotyped animals
- Average inbreeding in generation 10 = 0.21
- **No convergence after 5000 iterations**

- Ideal simulated population
- No missing pedigree
- All recent generations were in the pedigree file



Importance of inbreeding

Realized relationship matrix



Computed using Henderson-Quaas' algorithm, without inbreeding

$$\mathbf{A}^{22} < \mathbf{A}_{22}^{-1}$$

Ill conditioned MME

Inflated GEBV

Computed using Colleau's formula, which considers inbreeding

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Computed using VanRaden's formula, which considers inbreeding

Importance of inbreeding

Realized relationship matrix

Computed using Henderson-Quaas' algorithm, with inbreeding

Computed using Colleau's formula, which considers inbreeding

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

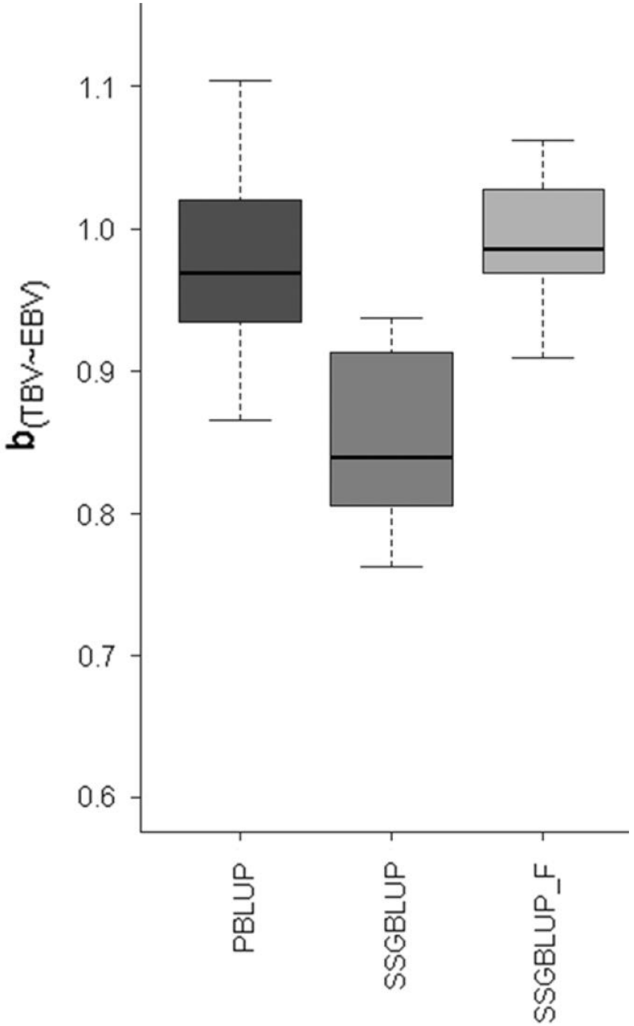
✓

Computed using VanRaden's formula, which considers inbreeding

Importance of inbreeding

Realized relationship matrix

- Garcia-Baccino et al. (2017)
- 29k pedigree | 5.3k genotyped animals
- PBLUP vs. ssGBLUP vs. ssGBLUP_inbreeding (F)
- Inflated GEBV with ssGBLUP
- No inflation with inbreeding

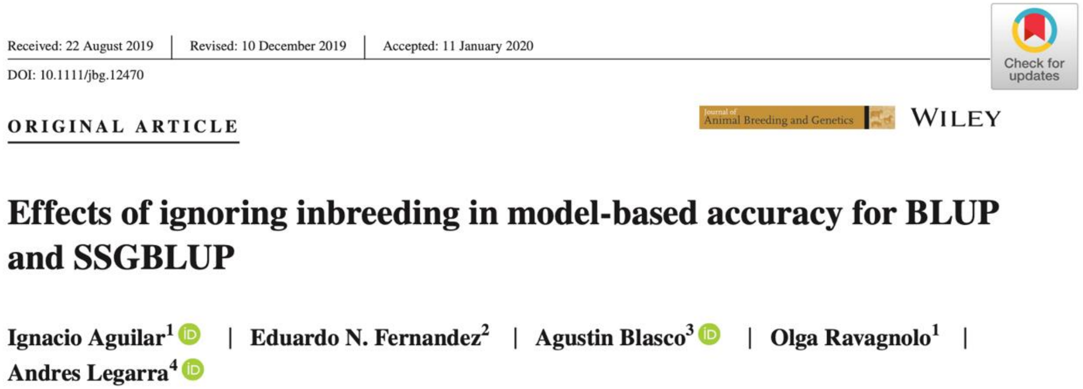


Importance of inbreeding

Realized relationship matrix

Inbreeding is also important in the estimation of accuracies

$$Accuracy_i = \sqrt{1 - \left[\frac{PEV_i}{\sigma_u^2(1 + F_i)}\right]}$$



Received: 22 August 2019 | Revised: 10 December 2019 | Accepted: 11 January 2020

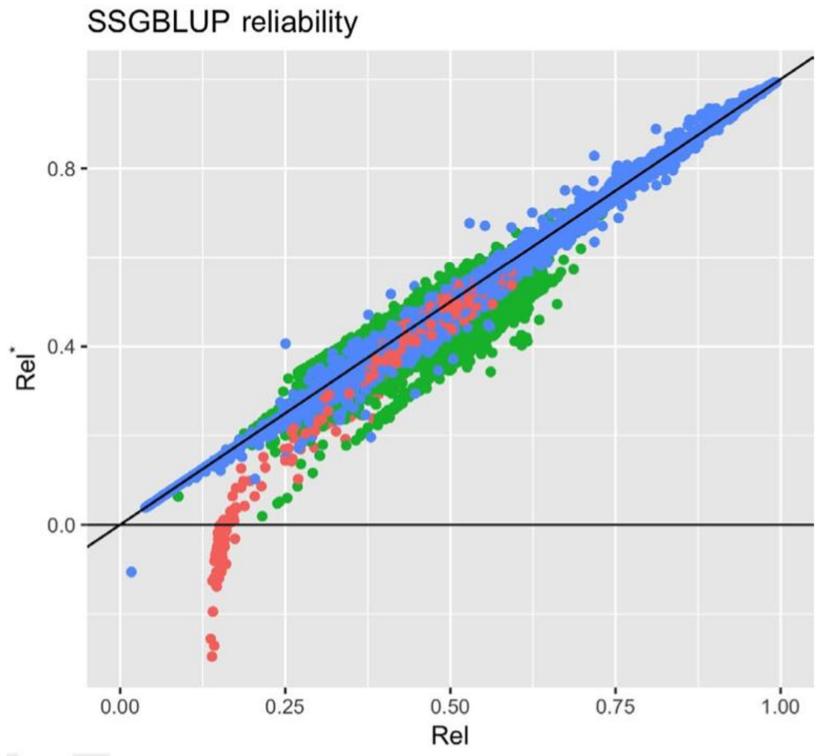
DOI: 10.1111/jbg.12470

Check for updates
Journal of Animal Breeding and Genetics
WILEY

ORIGINAL ARTICLE

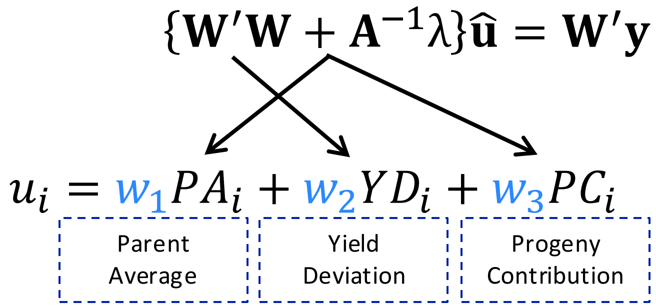
Effects of ignoring inbreeding in model-based accuracy for BLUP and SSGBLUP

Ignacio Aguilar[1] | Eduardo N. Fernandez[2] | Agustin Blasco[3] | Olga Ravagnolo[1] | Andres Legarra[4]



Decomposition of EBV and GEBV

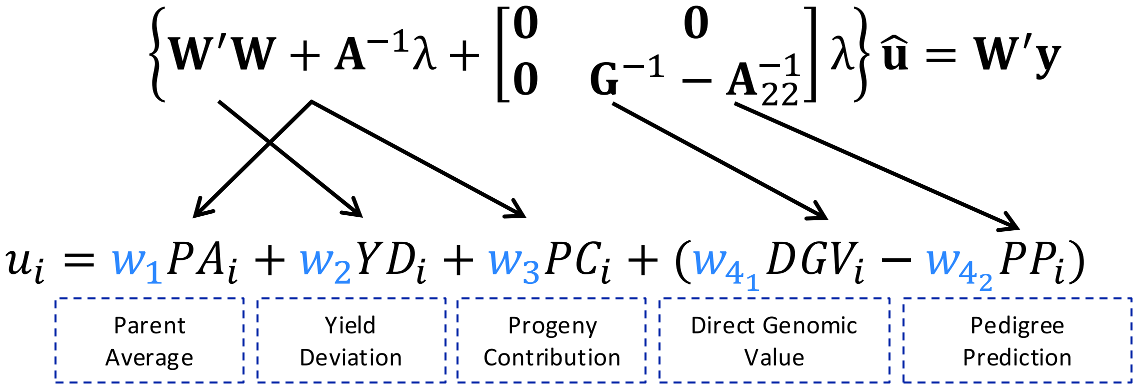
EBV



$$\{\mathbf{W'W} + \mathbf{A}^{-1}\lambda\}\hat{\mathbf{u}} = \mathbf{W'y}$$

$$u_i = w_1 PA_i + w_2 YD_i + w_3 PC_i$$

GEBV



$$\left\{\mathbf{W'W} + \mathbf{A}^{-1}\lambda + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}\lambda\right\}\hat{\mathbf{u}} = \mathbf{W'y}$$

$$u_i = w_1 PA_i + w_2 YD_i + w_3 PC_i + (w_{4_1} DGV_i - w_{4_2} PP_i)$$

For young animals $u_i = w_1 PA_i + (w_{4_1} DGV_i - w_{4_2} PP_i)$

With many genotypes $u_i \approx w_{4_1} DGV_i$

Importance of inbreeding

Decomposition of EBV and GEBV

- For young animals $u_i = w_1 PA_i + \left(w_{4_1} DGV_i - w_{4_2} PP_i\right)$

$$u_i = \frac{\frac{2}{1-F_i}}{\frac{2}{1-F_i}+g^{ii}-a_{22}^{ii}} PA_i + \left(\frac{g^{ii}}{\frac{2}{1-F_i}+g^{ii}-a_{22}^{ii}} DGV_i - \frac{a^{ii}}{\frac{2}{1-F_i}+g^{ii}-a_{22}^{ii}} PP_i\right)$$

- Ignoring F

$$u_i = \frac{2}{2+g^{ii}-a_{22}^{ii}} PA_i + \left(\frac{g^{ii}}{2+g^{ii}-a_{22}^{ii}} DGV_i - \frac{a^{ii}}{2+g^{ii}-a_{22}^{ii}} PP_i\right)$$

- Inbreeding increases the denominator
- GEBV is smaller
- Inflation is reduced

Estimating Variance Components

We require VC or at least some function of them

EM-REML

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

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

$$\lambda = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_u^2}$$

1. Set initial variance components
2. Compute $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ solving the MME
3. Update variance components

$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\mathbf{A}^{-1}\hat{\mathbf{u}} + tr(\mathbf{A}^{-1}\mathbf{C}^{\mathbf{uu}})\hat{\sigma}_e^2}{N}$$

Inverse of LHS for individual effect

Number of individuals, rank of **A**

$$\hat{\sigma}_e^2 = \frac{\mathbf{y}'\,(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}})}{N - rank(\mathbf{X})}$$

4. Go to 1 or stop if variance components do not change anymore

Patterson and Thompson (1971)

Dempster et al. (1977)

REML, GREML and SSGREML

Estimating Variance Components

AI-REML

Vector of VC

AI matrix

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_{n} - \mathbf{s}_{n}[\mathbf{AI}_{n}]^{-1}$$

Gradient (score vector)

AI- algorithm uses this matrix as Hessian

$\mathbf{P}$ = Projection or hat matrix

$$\mathbf{AI} = \begin{bmatrix} \frac{1}{2}\left\{\frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)'\mathbf{P}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)}{(\hat{\sigma}_e^2)^2}\right\} & \frac{1}{2}\left\{\frac{\mathbf{u}'\mathbf{Z}'\mathbf{P}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)}{\hat{\sigma}_e^2(\hat{\sigma}_u^2)^2}\right\} \\ \frac{1}{2}\left\{\frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)'\mathbf{P}\mathbf{Z}\mathbf{u}}{\hat{\sigma}_e^2(\hat{\sigma}_u^2)^2}\right\} & \frac{1}{2}\left\{\frac{\mathbf{u}'\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{u}}{(\hat{\sigma}_u^2)^2}\right\} \end{bmatrix}$$

Gradient

$$\mathbf{s} = \begin{bmatrix} \frac{1}{2}\left\{\frac{\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{u}}\right)}{(\hat{\sigma}_e^2)^2} - \frac{(\mathrm{n}-\mathrm{p}-\mathrm{q})}{(\hat{\sigma}_e^2)^2} - \frac{tr(\mathbf{A}^{-1}\mathbf{C}^{\mathrm{uu}})}{\hat{\sigma}_u^2}\right\} \\ \frac{1}{2}\left\{\frac{\hat{\mathbf{u}}'\mathbf{A}^{-1}\hat{\mathbf{u}}}{(\hat{\sigma}_u^2)^2} - \frac{q}{\hat{\sigma}_u^2} + tr(\mathbf{A}^{-1}\mathbf{C}^{\mathrm{uu}})\frac{\hat{\sigma}_e^2}{(\hat{\sigma}_u^2)^2}\right\} \end{bmatrix}$$

Estimating Variance Components

EM-REML

- Simple equations
  - More complex in multiple-trait models
- Very slow convergence
- Computationally demanding ($\mathbf{C}^{\mathrm{uu}}$)

AI-REML

- Faster than EM-REML
  - Fewer iterations
- Provides estimation of standard errors
- For complex models and poor starting values
  - Slow convergence
  - Estimates out of the parameter space
- Initial rounds with EM-REML may help
- Computationally demanding ($\mathbf{C}^{\mathrm{uu}}$)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

gibbsf90+

- gibbs1f90: stores single trait matrices once – fast for multi-trait models
- gibbs2f90: gibbs1f90 with joint sampling of correlated effects – Maternal effects and RRM
- gibbs3f90: gibbs2f90 with heterogeneous residual variance
- thrgibbs1f90: for linear-threshold models
- thrgibbs3f90: thrgibbs1f90 with heterogeneous residual variance

Variance Components Estimation

Mixed Model Equations Solver

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}W} \\ \mathbf{W'R^{-1}X} & \mathbf{W'R^{-1}W+A^{-1}\otimes G_0^{-1}} \end{bmatrix}\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix}=\begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{W'R^{-1}y} \end{bmatrix}$$

gibbsf90+

Linear

Default

Threshold (-Linear)

```
OPTION cat 0 2 5
```

- Categories renumbered from **1**
- Missing records is only **0**

gibbsf90+

Bayes Theorem

$$p(\theta|y) = p(y|\theta)\, p(\theta)$$

Likelihood function

indicates how likely the observations are from a distribution (with particular parameters)

prior probability of unknown θ

posterior probability of unknown θ with known y

- Basic idea of Gibbs Sampling:
- Numerical method to draw samples from a posterior distribution (not always explicitly available)
- Draw samples = generate random numbers following a distribution
- The results are random numbers (not theoretical formulas)
- The posterior distribution will be drawn based on the numerical values (like a histogram)

gibbsf90+

Ingredients for Gibbs sampling

1) Theoretical derivation: conditional posterior distribution for each unknown parameter

2) Software: a random number generator for a particular distribution

```r
# Basic Gibbs sampling for mu (normal) and sigma2 (inverted chi-square)
y <- c(14,16,18)
N <- length(y)
n.samples <- 100
mu <- rep(0,n.samples)
sigma2 <- rep(0,n.samples)

# initial value
mu[1] <- 0
sigma2[1] <- 10

# sampling
for(i in 2:n.samples){
    mu[i] <- rnorm(1, mean=mean(y), sd=sqrt(sigma2[i-1]/N)) # using the most recent sigma2
    df <- N-2
    S <- sum((y-mu[i])^2)
    sigma2[i] <- rinvchisq(1, df=df, scale=S) # using the most recent mu
}
```

gibbsf90+

- Name of parameter file?

  gibbs1.par

- Number of samples and length of burn-in?

  samples=10,000 to 100,000; burn-in=0

- Give n to store every n-th sample?

  10

- `gibbsf90+ parfile.par --samples i --burnin j --interval k`

gibbsf90+

- Procedure
  - Run `gibbsf90+` to estimate variance components
  - Run `postgibbsf90` to process the samples and check convergence
  - Run `gibbsf90+` with new variance components to compute EBV (2k to 10k samples)

```
OPTION fixed_var mean X
```

Number of the animal effect

postgibbsf90

- Basic idea of post-Gibbs analysis:

- Summarize and visualize the samples drawn by gibbsf90+
- Confirm if the chain converged
- Find the most probable value = posterior mode as a “point estimate”
- Find the reliability of the estimates = the highest posterior density as a “confidence interval”

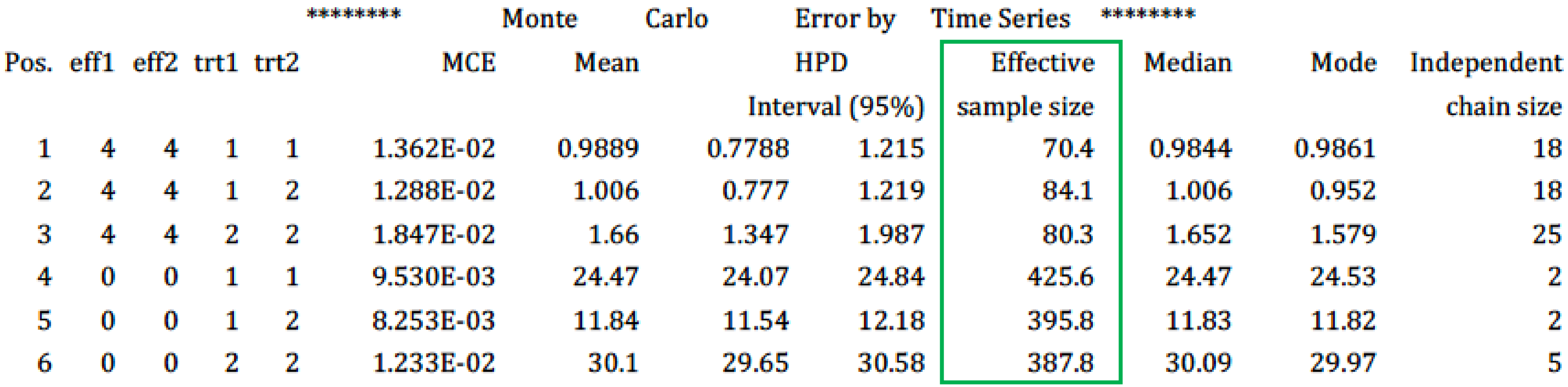
postgibbsf90

- Name of parameter file?
  - gibbs1.par
- Burn-in?
  - 0
- Give n to store every n-th sample? (1 means read all samples)
  - 10

- input files
  - gibbs_samples, fort.99
- output files
  - "postgibbs_samples"
    - all Gibbs samples for additional post analyses
  - "postmean"
    - posterior means
  - "postsd"
    - posterior standard deviations
  - "postout"

postgibbsf90

at least > 10 is recommended
> 30 may be better



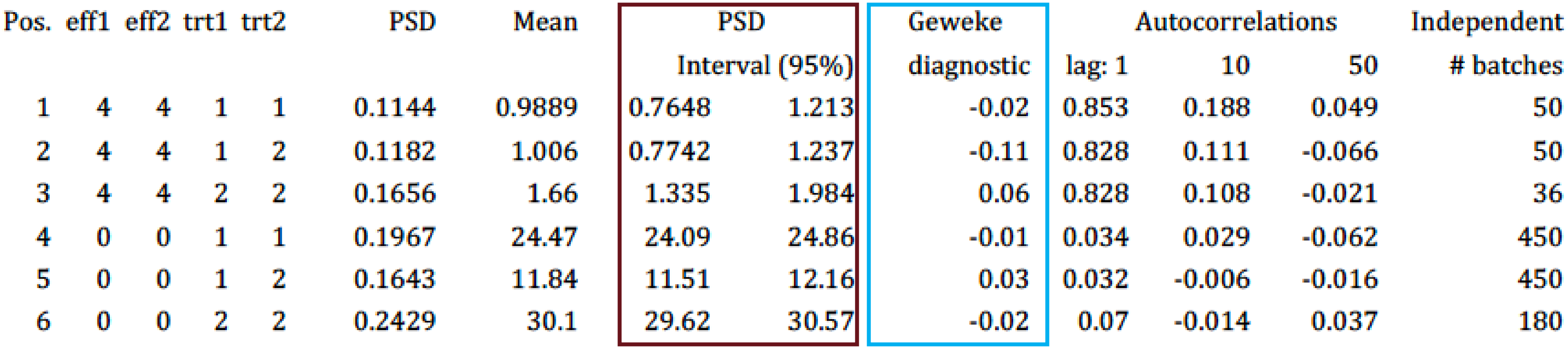
******** Monte Carlo Error by Time Series ********

Pos.	eff1	eff2	trt1	trt2	MCE	Mean	HPD		Effective	Median	Mode	Independent
							Interval (95%)		sample size			chain size
1	4	4	1	1	1.362E-02	0.9889	0.7788	1.215	70.4	0.9844	0.9861	18
2	4	4	1	2	1.288E-02	1.006	0.777	1.219	84.1	1.006	0.952	18
3	4	4	2	2	1.847E-02	1.66	1.347	1.987	80.3	1.652	1.579	25
4	0	0	1	1	9.530E-03	24.47	24.07	24.84	425.6	24.47	24.53	2
5	0	0	1	2	8.253E-03	11.84	11.54	12.18	395.8	11.83	11.82	2
6	0	0	2	2	1.233E-02	30.1	29.65	30.58	387.8	30.09	29.97	5

******** P tio

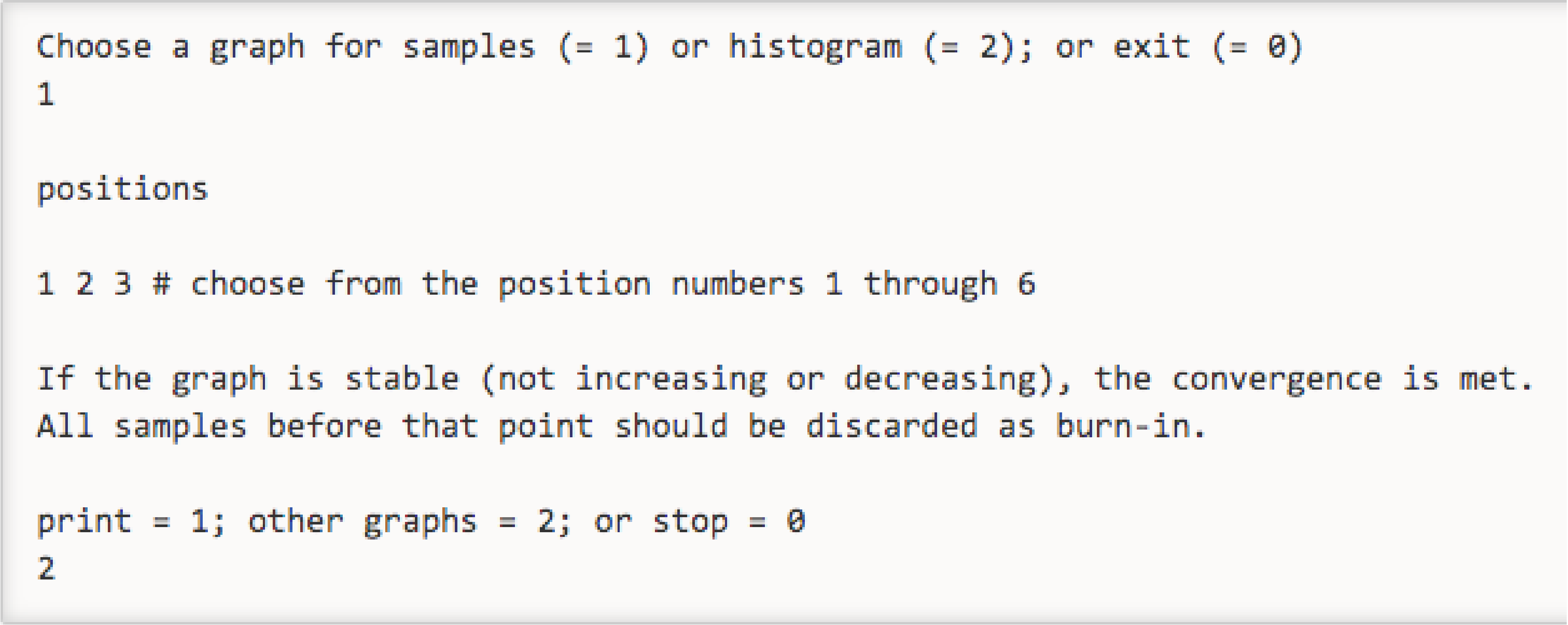
Lower and upper bounds
of Mean ± 1.96PSD

ratio between first half and second
half of the samples ; should be < 1.0



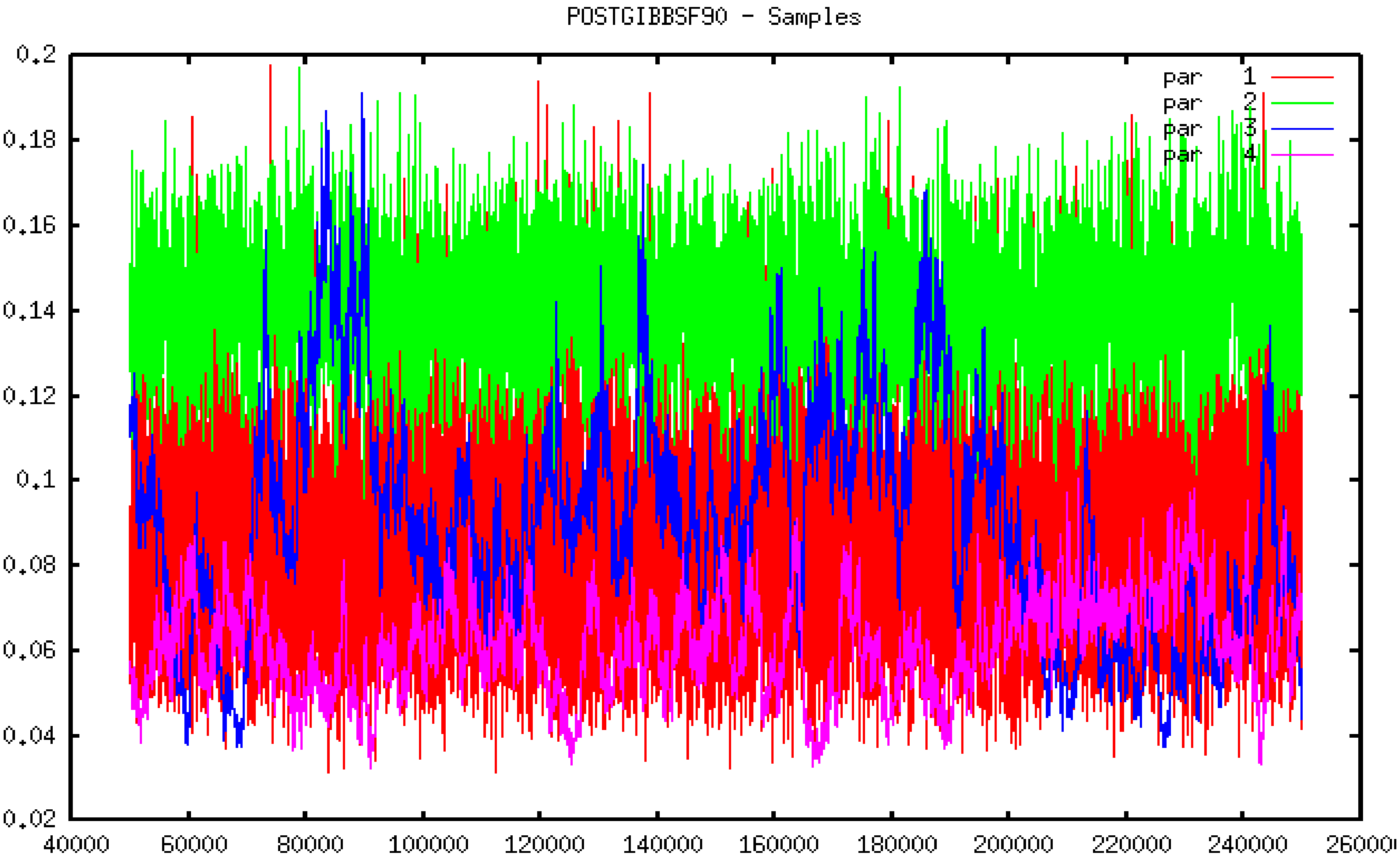
Pos.	eff1	eff2	trt1	trt2	PSD	Mean	PSD		Geweke	Autocorrelations			Independent
							Interval (95%)		diagnostic	lag: 1	10	50	# batches
1	4	4	1	1	0.1144	0.9889	0.7648	1.213	-0.02	0.853	0.188	0.049	50
2	4	4	1	2	0.1182	1.006	0.7742	1.237	-0.11	0.828	0.111	-0.066	50
3	4	4	2	2	0.1656	1.66	1.335	1.984	0.06	0.828	0.108	-0.021	36
4	0	0	1	1	0.1967	24.47	24.09	24.86	-0.01	0.034	0.029	-0.062	450
5	0	0	1	2	0.1643	11.84	11.51	12.16	0.03	0.032	-0.006	-0.016	450
6	0	0	2	2	0.2429	30.1	29.62	30.57	-0.02	0.07	-0.014	0.037	180

postgibbsf90

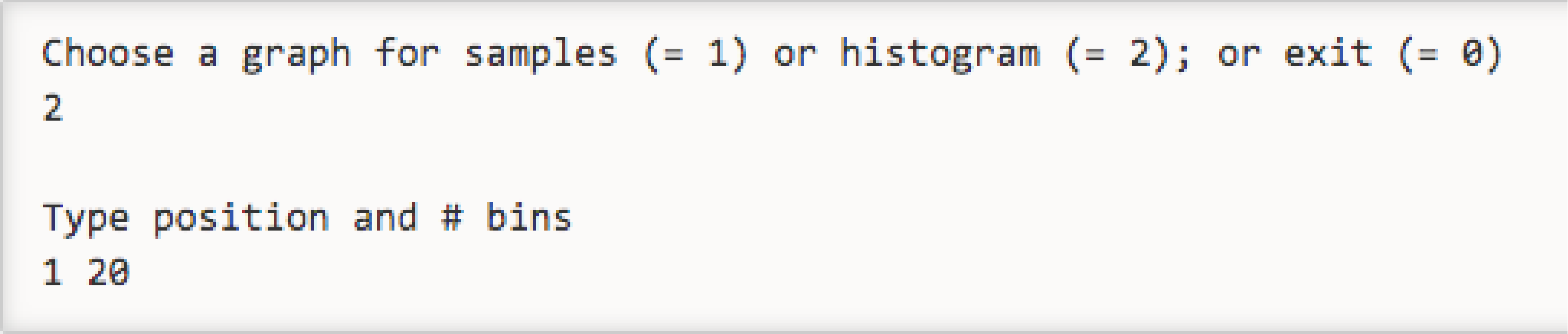


```
Choose a graph for samples (= 1) or histogram (= 2); or exit (= 0)
1

positions

1 2 3 # choose from the position numbers 1 through 6

If the graph is stable (not increasing or decreasing), the convergence is met.
All samples before that point should be discarded as burn-in.

print = 1; other graphs = 2; or stop = 0
2
```

postgibbsf90

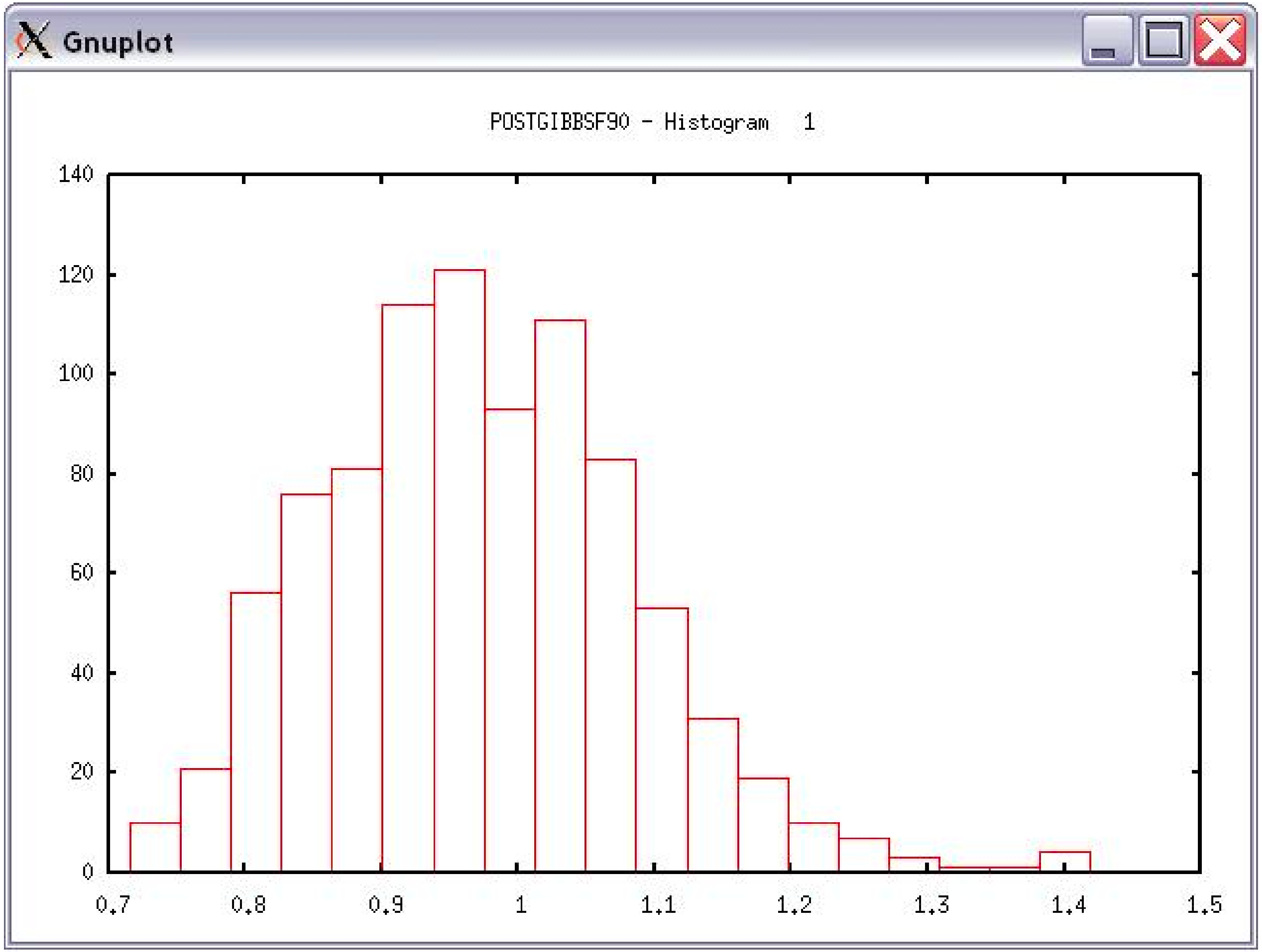


postgibbsf90



```
Choose a graph for samples (= 1) or histogram (= 2); or exit (= 0)
2

Type position and # bins
1 20
```

postgibbsf90



Common problems for BLUPF90 family

- Wrong position or formats for observation and effects

- Misspelling of Keywords
  - Program may stop

- (Co)variance matrices not symmetric, not positive definite
  - Program may not stop

- Large numbers (e.g., 305-day milk yield 10,000 kg)
  - Scale down i.e., 10,000 /1,000 = 10

General output from BLUPF90 family

- Output printed on the screen is not saved to any file!
- Should use redirection or pipes to store output

renumf90

```
renumf90 renum.par | tee renum.log
```

blupf90+

```
blupf90+ renf90.par | tee blup.log
```

gibbsf90+

```
gibbsf90+ exmr99s1 --samples 1000 --burnin 0 --interval 1 | tee gibbs.log
```

Run in background + Save output

```
$vi gibbs.sh
#type the following commands inside gibbs.sh
      gibbsf90+ <<AA > gibbs.log
      renf90.par
      1000 0
      10
      AA
#save and exit
$bash gibbs.sh &  #can replace bash with sh
```

```
$vi bp.sh
#type the following commands inside bp.sh
      blupf90+ <<AA > blup.log
      renf90.par
      AA
#save and exit
$bash bp.sh &   #can replace bash by sh
```

REML, GREML and SSGREML

Estimating Variance Components

blupf90+

MME Solver

Default

- Preconditioner Conjugate Gradient (PCG)
  - Default Iterative method (fast)

- Successive over-relaxation (SOR)
  - An iterative method based on Gauss-Seidel

- Direct solution using sparse Cholesky factorization
  - FSPAK or YAMS (greater memory requirements)

VC Estimation

- AI-REML:

```
OPTION method VCE
```

- EM-REML:

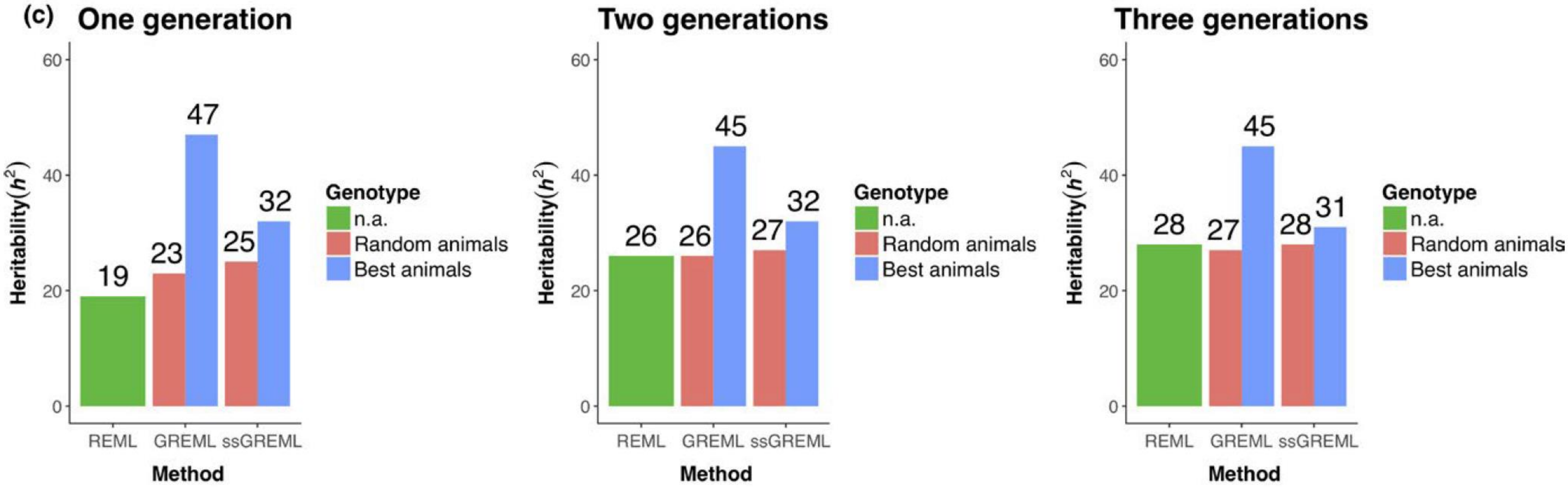
```
OPTION method VCE

OPTION EM-REML xx
```

_ (empty for pure EM)
# of EM rounds
ai (until convergence)

Difference in estimates depending on population structure

Estimating Variance Components

- In practice, it is hard to have base allele frequencies
- SSGREML was less affected by selective or limited genotyping

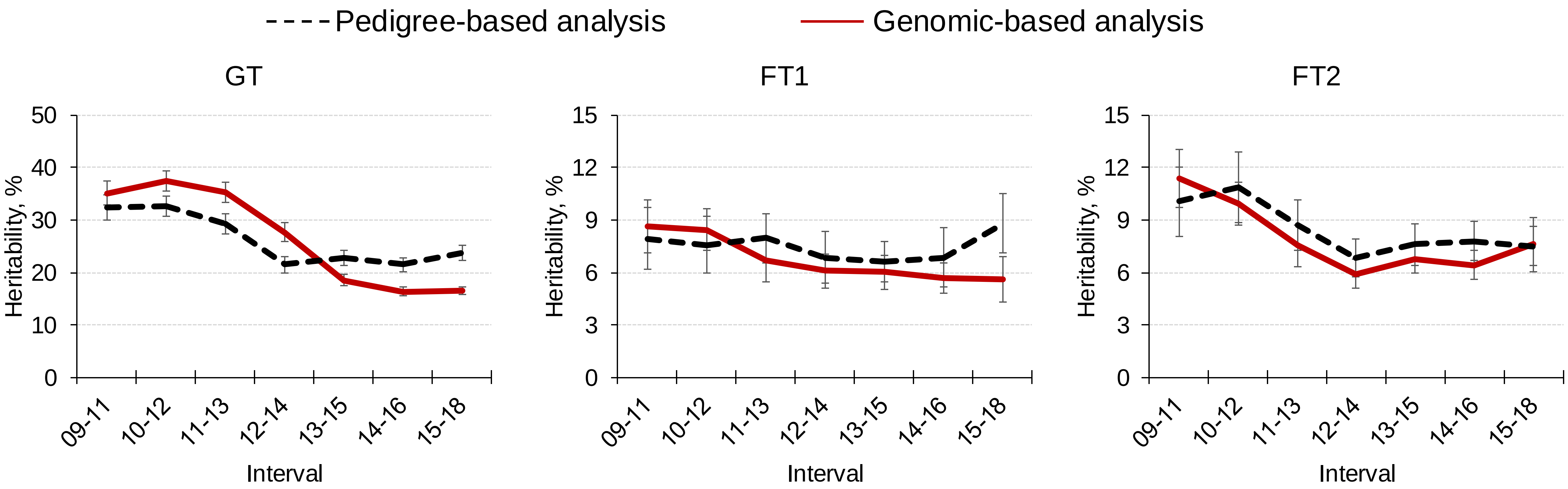


- Estimated heritability = 30%

Cesarani et al. (2018)

Difference in estimates depending on population structure

Estimating Variance Components



35.1 to 16.5%

8.6 to 5.6%

11.4 to 7.6%

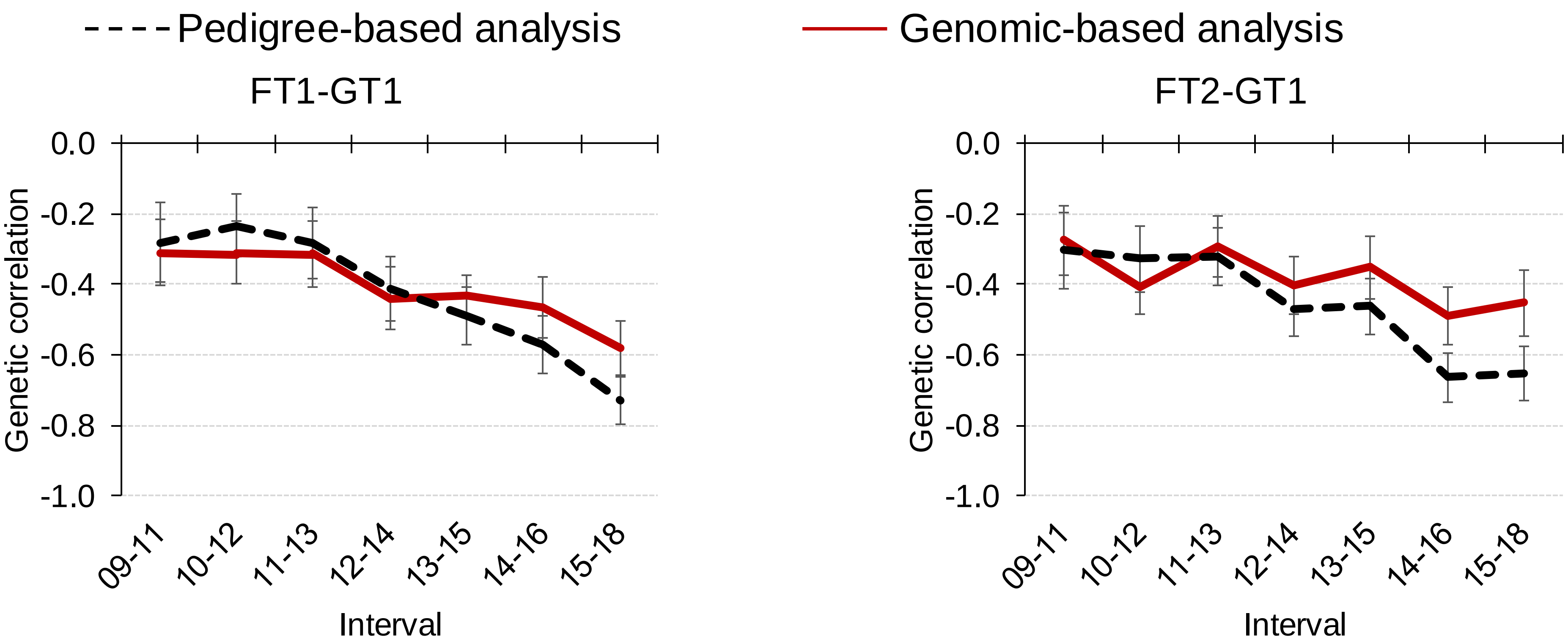
A reduction of ~ 50%

~ 20%

Hidalgo et al. (2020)

Difference in estimates depending on population structure

Estimating Variance Components



-0.31 to -0.58

-0.27 to -0.45

- These changes need to be considered in the breeding program

Hidalgo et al. (2020)

Practice

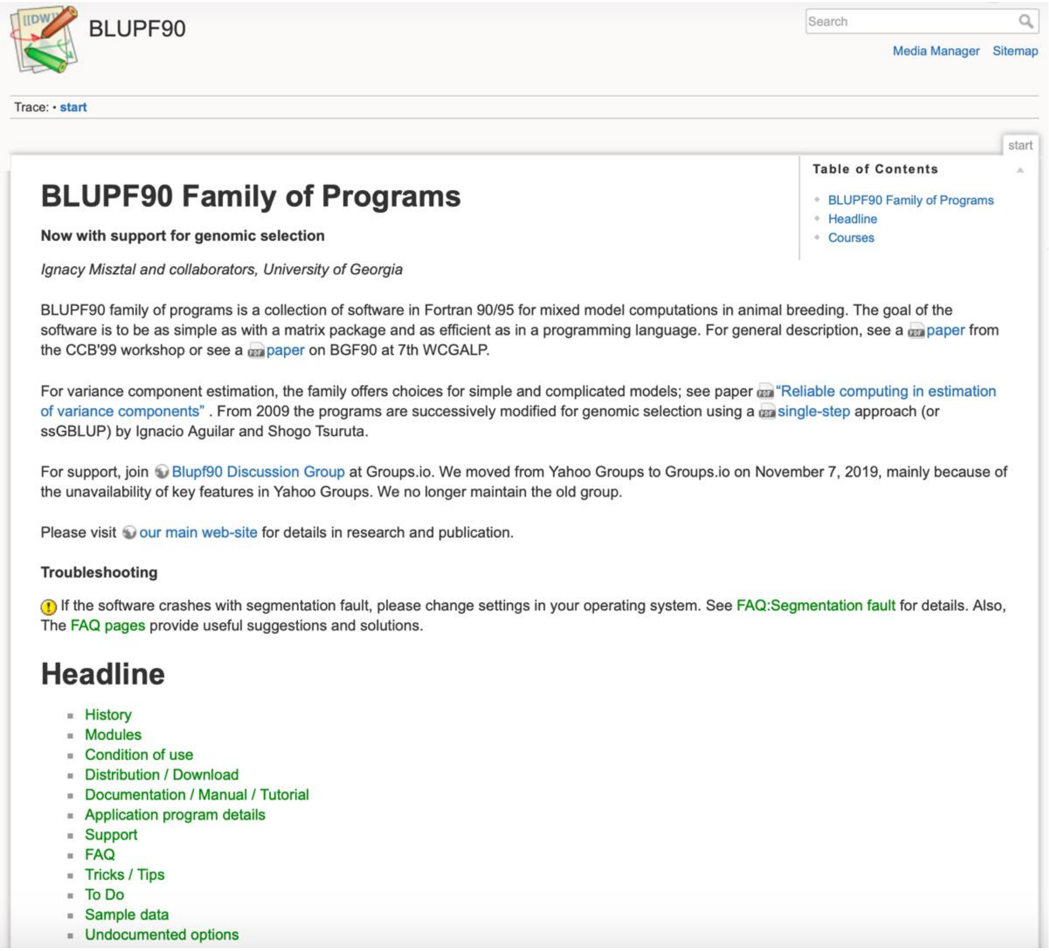


Table of Contents

- BLUPF90 Family of Programs
- Headline
- Courses

BLUPF90 Family of Programs

Now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a paper from the CCB'99 workshop or see a paper on BGF90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper "Reliable computing in estimation of variance components" . From 2009 the programs are successively modified for genomic selection using a single-step approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

For support, join Blupf90 Discussion Group at Groups.io. We moved from Yahoo Groups to Groups.io on November 7, 2019, mainly because of the unavailability of key features in Yahoo Groups. We no longer maintain the old group.

Please visit our main web-site for details in research and publication.

Troubleshooting

If the software crashes with segmentation fault, please change settings in your operating system. See FAQ:Segmentation fault for details. Also, The FAQ pages provide useful suggestions and solutions.

Headline

- History
- Modules
- Condition of use
- Distribution / Download
- Documentation / Manual / Tutorial
- Application program details
- Support
- FAQ
- Tricks / Tips
- To Do
- Sample data
- Undocumented options

- Collection of software
- Fortran ≥ 90
- Computations in AB & G
- Since 1997 by Ignacy Misztal
- Several developers + collaborators
- Simple, efficient, and comprehensive
- Very general models

https://nce.ads.uga.edu

https://nce.ads.uga.edu/software/

Blupf90 programs

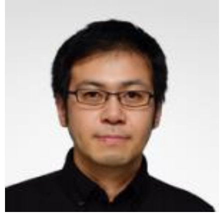
Practice

Ignacy
Misztal

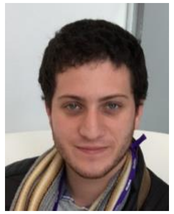
Shogo
Tsuruta

Andres
Legarra

Ignacio
Aguilar



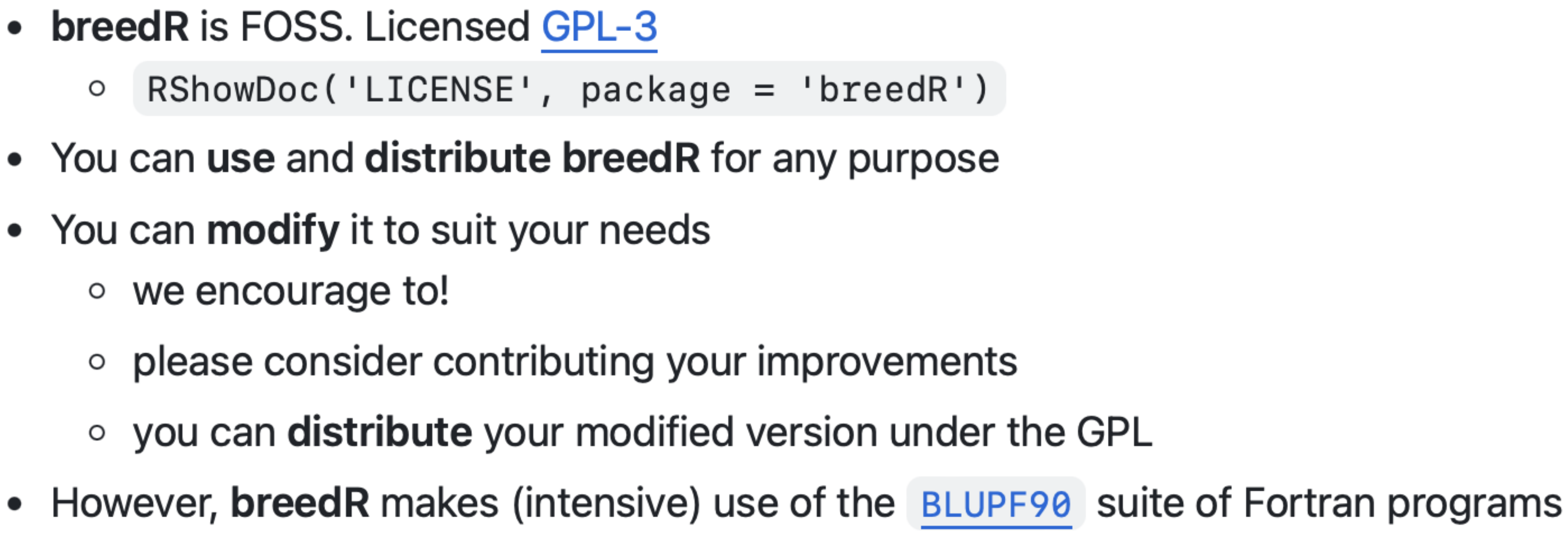
Yutaka
Masuda



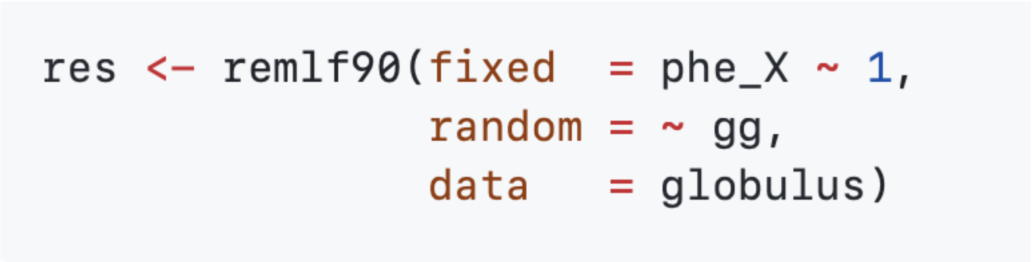
Matias
Bermann

- + Several contributors
- Research turns into code

Practice



- **breedR** is FOSS. Licensed GPL-3
  - `RShowDoc('LICENSE', package = 'breedR')`
- You can **use** and **distribute breedR** for any purpose
- You can **modify** it to suit your needs
  - we encourage to!
  - please consider contributing your improvements
  - you can **distribute** your modified version under the GPL
- However, **breedR** makes (intensive) use of the `BLUPF90` suite of Fortran programs



```
res <- remlf90(fixed  = phe_X ~ 1,
               random = ~ gg,
               data   = globulus)
```

https://github.com/famuvie/breedR/wiki/Overview

Blupf90 programs

Practice

blupf90

BLUP with explicit equations

remlf90

Expectation Maximization REML

airemI90

Average Information REML

gibbsf90

Bayesian Analyses – linear traits

thrgibbsf90

Bayesian Analyses – categorical traits

postgibbsf90

Post-analyses of Gibbs samples

blupf90+

gibbsf90+

renumf90

Renumbering + data QC

preGSf90

Processing of SNP data (QC + matrices)

QCf90 ✖

QC of large SNP data

postGS90

Estimation of SNP effects and GWAS

predf90 ✖

Prediction of GEBV based on SNP effects

seekparent90 ✖

Parentage verification (SNP and pedigree)

predictf90

Adjusted and predicted phenotypes + residuals

blup90iod

cblup90iod

accf90

accf90GS

✖ No need for the renumbering process

Practice

- Renumf90 parameter file

renumf90

- `renumf90 --help`
- `renumf90 --show-template`

- FAQ blupf90

https://nce.ads.uga.edu/wiki/doku.php?id=faq

```
DATAFILE
data3.txt
TRAITS
 4
FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE
0.60
EFFECT
3 cross alpha
EFFECT
1  cross alpha
RANDOM
animal
FILE
ped3.txt
FILE_POS
1 2 3 0 0
SNP_FILE
snp3.2k
PED_DEPTH
0
(CO)VARIANCES
0.40
OPTION map_file mrkmap.txt
OPTION use_yams
```

$$y = sex + animal + e$$

$$\hat{\sigma}_u^2 = 0.4$$

$$\hat{\sigma}_e^2 = 0.6$$

Practice

- Renumf90 parameter file

renumf90

```
DATAFILE
 renf90.dat
NUMBER_OF_TRAITS
           1
NUMBER_OF_EFFECTS
           2
OBSERVATION(S)
    1
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBER_OF_LEVELS TYPE_OF_EFFECT[EFFECT NESTED]
 2          2 cross
 3     12010 cross
RANDOM_RESIDUAL VALUES
  0.60000
 RANDOM_GROUP
     2
 RANDOM_TYPE
 add_an_upginb
 FILE
renadd02.ped
(CO)VARIANCES
  0.40000
OPTION SNP_file snp3.2k
OPTION map_file mrkmap.txt
OPTION use_yams
```

$$y = sex + animal + e$$

$$\hat{\sigma}_u^2 = 0.4$$

$$\hat{\sigma}_e^2 = 0.6$$

- FAQ blupf90

https://nce.ads.uga.edu/wiki/doku.php?id=faq

Practice

RANDOM_GROUP

Number of the effect(s) from list of effects

Correlated effects should be consecutive e.g. Maternal effects, Random Regression

RANDOM_TYPE

diagonal, add_animal, add_sire, add_an_upg, add_an_upginb, add_an_self, user_file, user_file_i, or par_domin

FILE

Pedigree file, parental dominance, or user file

(CO)VARIANCES

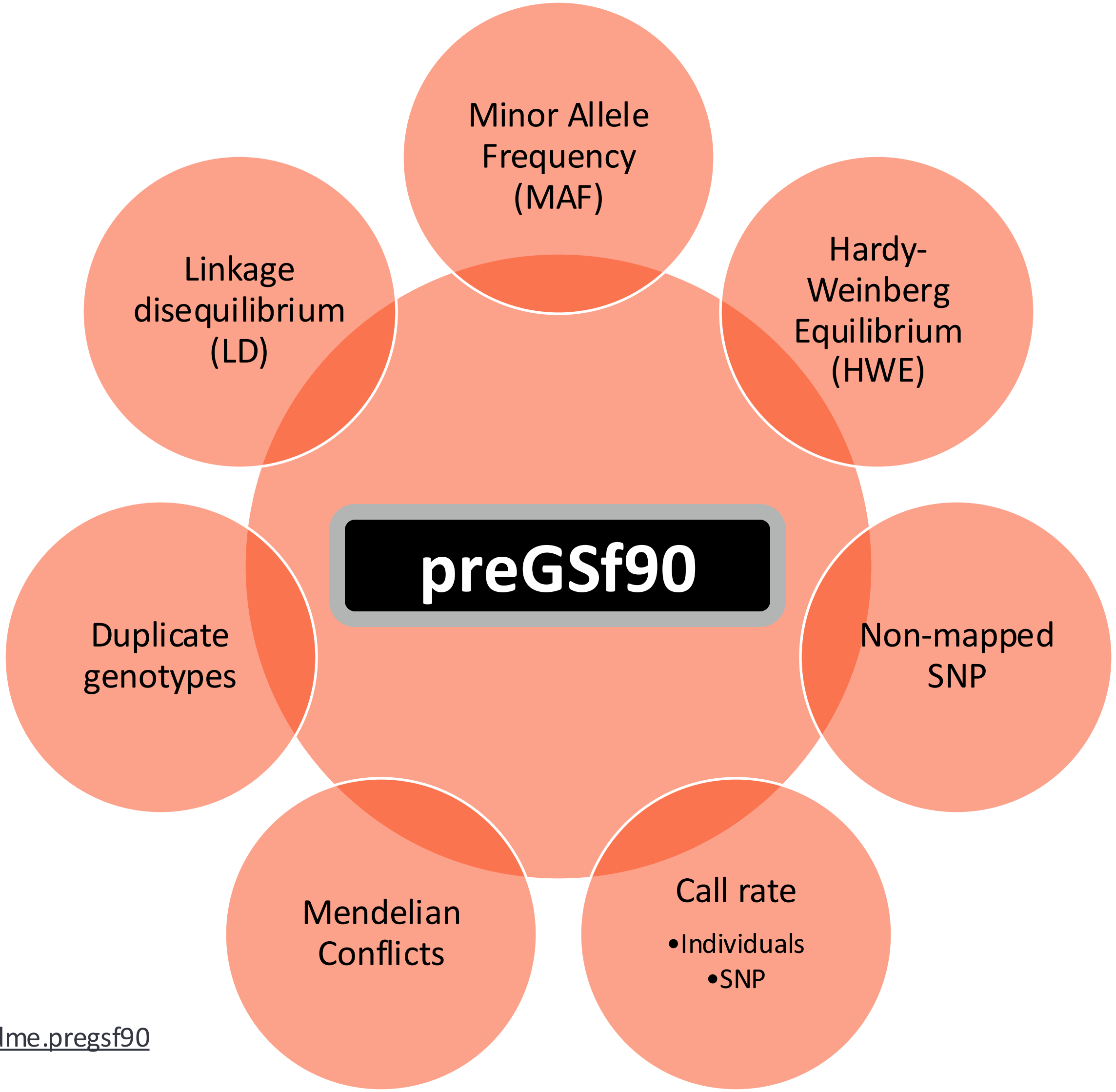
Square matrix with dimension equal to the number_of_traits*number_of_correlated_effects

- *Add_an_self*
  - *To create a relationship matrix when there is selfing*
  - Pedigree file:
    - `individual, parent 1, parent 2, number of selfing generations`
- *user_file*
  - An inverted matrix is read from the file
  - Matrix is stored only upper- or lower-triangular
  - Matrix file:
    - `row, col, value`
- *user_file_i*
  - As before but the matrix will be inverted by the program

preGSf90

Practice

- Quality control



Minor Allele Frequency (MAF)

Hardy-Weinberg Equilibrium (HWE)

Non-mapped SNP

Call rate
- Individuals
- SNP

Mendelian Conflicts

Duplicate genotypes

Linkage disequilibrium (LD)

preGSf90

https://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90

Practice

- Same parameter file as for all BLUPF90 programs
- Needs an extra OPTION in renf90.par
- OPTION SNP_file marker.geno

preGSf90

- Reads 2 extra files (besides data and pedigree):
- marker.geno
- marker.geno_XrefID(created by renumf90)

- `_XrefID` has 2 columns: Renumbered ID Original ID *preGSf90*

Practice

- preGSf90 saves $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ by default (file: GimA22i)
  - To save 'raw' genomic matrix:
  - OPTION saveG [all]
    - If the optional all is present all intermediate G matrices will be saved!!!

preGSf90

  - To save $\mathbf{G}^{-1}$
  - OPTION saveGInverse
  - Only the final G, after blending, scaling, etc. is inverted !!!

  - To save $\mathbf{A}_{22}$ and inverse
  - OPTION saveA22 and OPTION saveA22Inverse

Practice

- renf90.par
- OPTION method VCE
- OPTION EM-REML xx

blupf90+

gibbsf90+

```
DATAFILE
 renf90.dat
NUMBER_OF_TRAITS
           1
NUMBER_OF_EFFECTS
           2
OBSERVATION(S)
    1
WEIGHT(S)

EFFECTS: POSITIONS_IN_DATAFILE NUMBER_OF_LEVELS TYPE_OF_EFFECT[EFFECT NESTED]
 2          2 cross
 3     12010 cross
RANDOM_RESIDUAL VALUES
  0.60000
 RANDOM_GROUP
     2
 RANDOM_TYPE
 add_an_upginb
 FILE
renadd02.ped
(CO)VARIANCES
  0.40000
OPTION SNP_file snp3.2k_clean
OPTION map_file mrkmap.txt_clean
OPTION use_yams
```

$$y = sex + animal + e$$

$$\hat{\sigma}_u^2 = 0.4$$

$$\hat{\sigma}_e^2 = 0.6$$

Blupf90 programs

Practice

- EM-REML
  - OPTION SNP_file snp3.2k_clean
  - OPTION map_file mrkmap.txt_clean
  - OPTION no_quality_control
  - OPTION use_yams
  - OPTION method VCE
  - OPTION EM-REML

```
At round: 23 Converge in fewer rounds than EM-REML rounds: 10000
  Stop EM-REML at 23 and no runs with AI-REML


* END ITERATION: 07-17-2024  09h 54m 06s 649
solutions stored in file: "solutions"


Final Estimates
 Genetic variance(s) for effect  2
  0.35532
 Residual variance(s)
  0.61222
 *** Statistical Method: VCE
 * FINISHED (BLUPF90): 07-17-2024  09h 54m 06s 680
```

blupf90+

Practice

- AI-REML
  - OPTION SNP_file snp3.2k_clean
  - OPTION map_file mrkmap.txt_clean
  - OPTION no_quality_control
  - OPTION use_yams
  - OPTION method VCE

blupf90+

```
-2logL =    26720.6457620796      : AIC =    26724.6457620796
 In round           4  convergence=  7.833323538291451E-014
 delta convergence=  7.908716592526159E-008
new R
 0.61221
new G
 0.35534
* END ITERATION: 07-17-2024  10h 14m 55s 278
solutions stored in file: "solutions"


Final Estimates
 Genetic variance(s) for effect  2
 0.35534
 Residual variance(s)
 0.61221
 inverse of AI matrix (Sampling Variance)
 0.73121E-03 -0.37380E-03
-0.37380E-03  0.32167E-03
 Correlations from inverse of AI matrix
  1.0000     -0.77076
-0.77076       1.0000
 SE for G
 0.27041E-01
 SE for R
 0.17935E-01
*** Statistical Method: VCE
* FINISHED (BLUPF90): 07-17-2024  10h 14m 55s 315
```

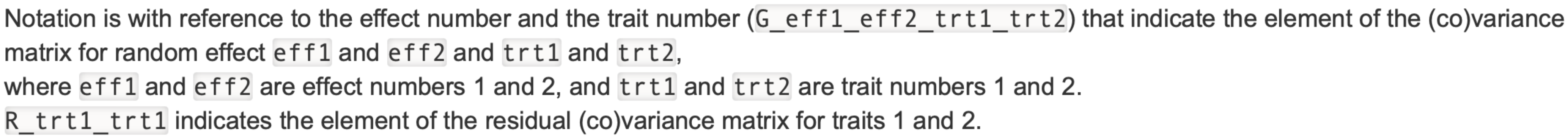
Blupf90 programs

Practice

- AI-REML

  SE for genetic parameters

- OPTION se_covar_function h2 G_2_2_1_1/(G_2_2_1_1+R_1_1)



Notation is with reference to the effect number and the trait number (`G_eff1_eff2_trt1_trt2`) that indicate the element of the (co)variance matrix for random effect `eff1` and `eff2` and `trt1` and `trt2`,
where `eff1` and `eff2` are effect numbers 1 and 2, and `trt1` and `trt2` are trait numbers 1 and 2.
`R_trt1_trt1` indicates the element of the residual (co)variance matrix for traits 1 and 2.

blupf90+

- https://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90

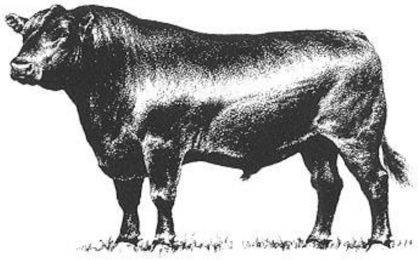
Practice

- MME solver (default)
  - OPTION SNP_file snp3.2k_clean
  - OPTION map_file mrkmap.txt_clean
  - OPTION no_quality_control
  - OPTION use_yams
  - OPTION store_accuracy eff orig

```
jorgehidalgo@endpoint-10-192-53-192 Data % head acc_bf90
trait/effect level original_id  solution acc
   1   2         1 UGA46217          0.05314548  0.5257
   1   2         2 UGA46272         -0.16554279  0.5903
   1   2         3 UGA43455         -1.22049127  0.5542
   1   2         4 UGA51333         -0.22292902  0.5449
   1   2         5 UGA42183         -0.15143591  0.7176
   1   2         6 UGA51501         -0.09200698  0.5224
   1   2         7 UGA43704         -0.12728916  0.5011
   1   2         8 UGA44900          0.49888989  0.5319
   1   2         9 UGA45303         -0.24224250  0.5009
```

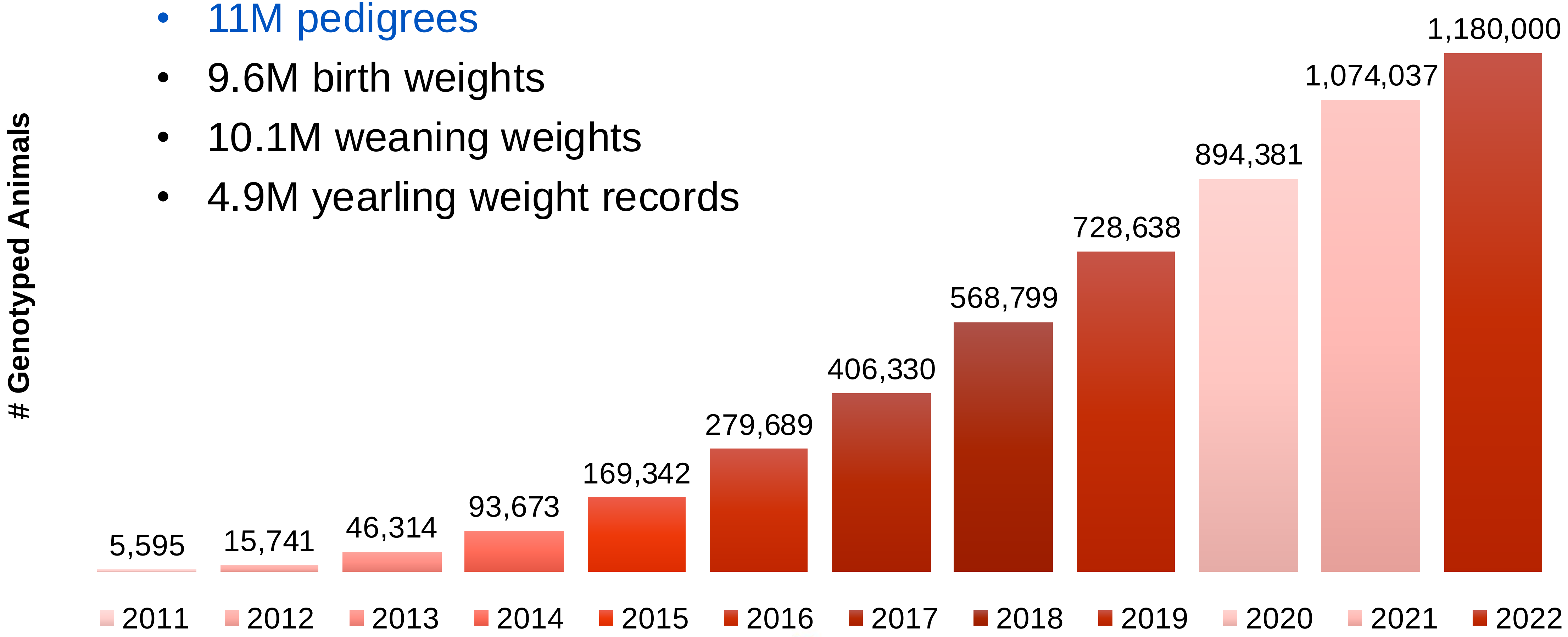
blupf90+

An application example



American Angus

- 11M pedigrees
- 9.6M birth weights
- 10.1M weaning weights
- 4.9M yearling weight records



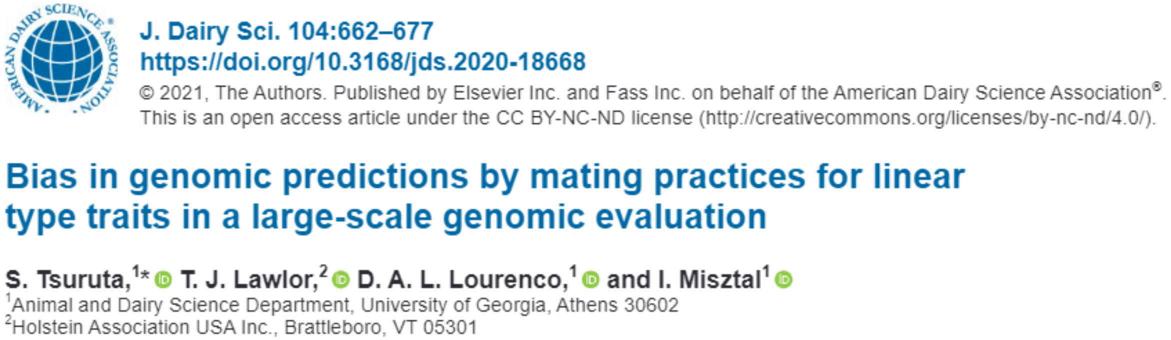
Blupf90 programs

An application example – largest ssGBLUP evaluation

- US Holstein Type trait data
  - 18 trait-model
  - 13.6M animals in pedigree
  - 10.2M phenotypes
  - 2.3M genotyped animals
  - 447,492,870 equations to solve

- APY ssGBLUP with 15k core
  - 1 day to build $\mathbf{G}_{\mathrm{APY}}^{-1}$ and $\mathbf{A}_{22}^{-1}$
  - ~ 2.5 days to converge
  - < 500 GB RAM with APY

  - > 30 TB RAM to compute $\mathbf{G}^{-1}$ without APY



J. Dairy Sci. 104:662–677
https://doi.org/10.3168/jds.2020-18668

© 2021, The Authors. Published by Elsevier Inc. and Fass Inc. on behalf of the American Dairy Science Association®. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

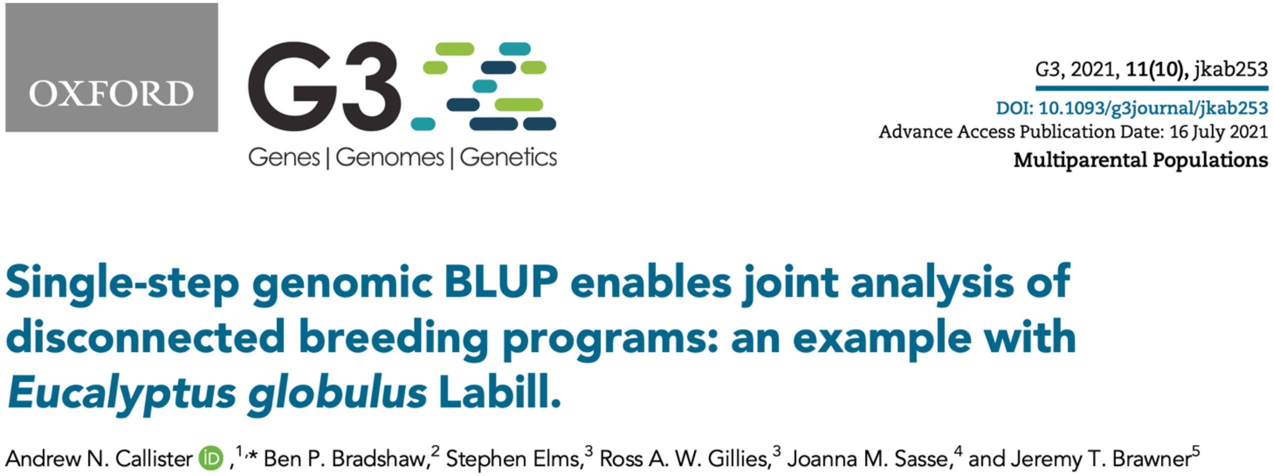
Bias in genomic predictions by mating practices for linear type traits in a large-scale genomic evaluation

S. Tsuruta,[1*] T. J. Lawlor,[2] D. A. L. Lourenco,[1] and I. Misztal[1]

[1] Animal and Dairy Science Department, University of Georgia, Athens 30602
[2] Holstein Association USA Inc., Brattleboro, VT 05301

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

An application example

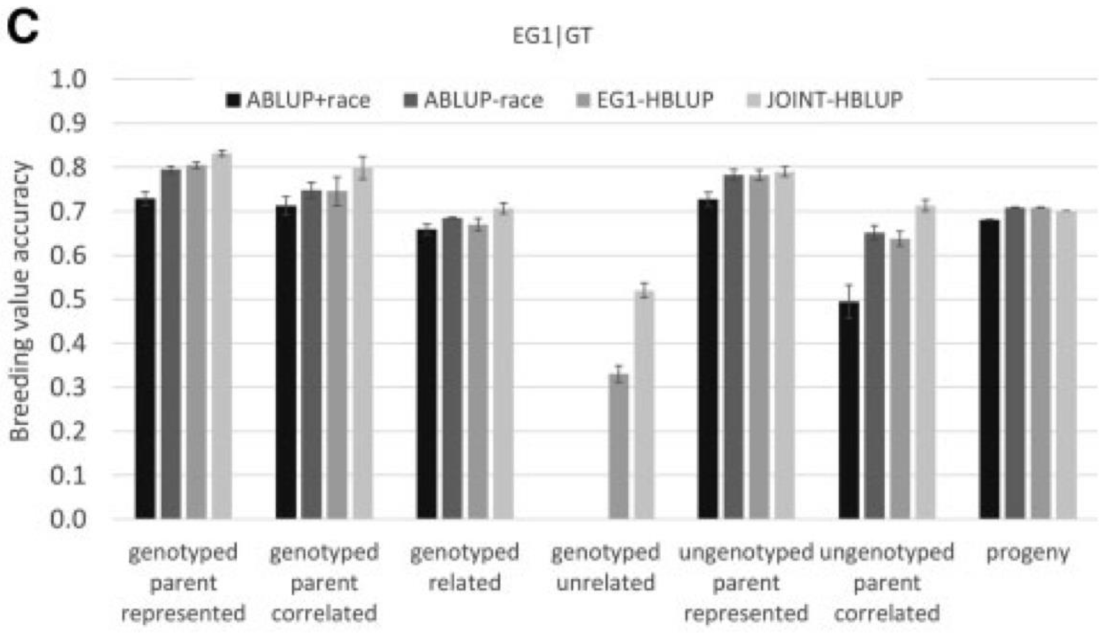


OXFORD G3 Genes | Genomes | Genetics

G3, 2021, **11(10)**, jkab253
DOI: 10.1093/g3journal/jkab253
Advance Access Publication Date: 16 July 2021
Multiparental Populations

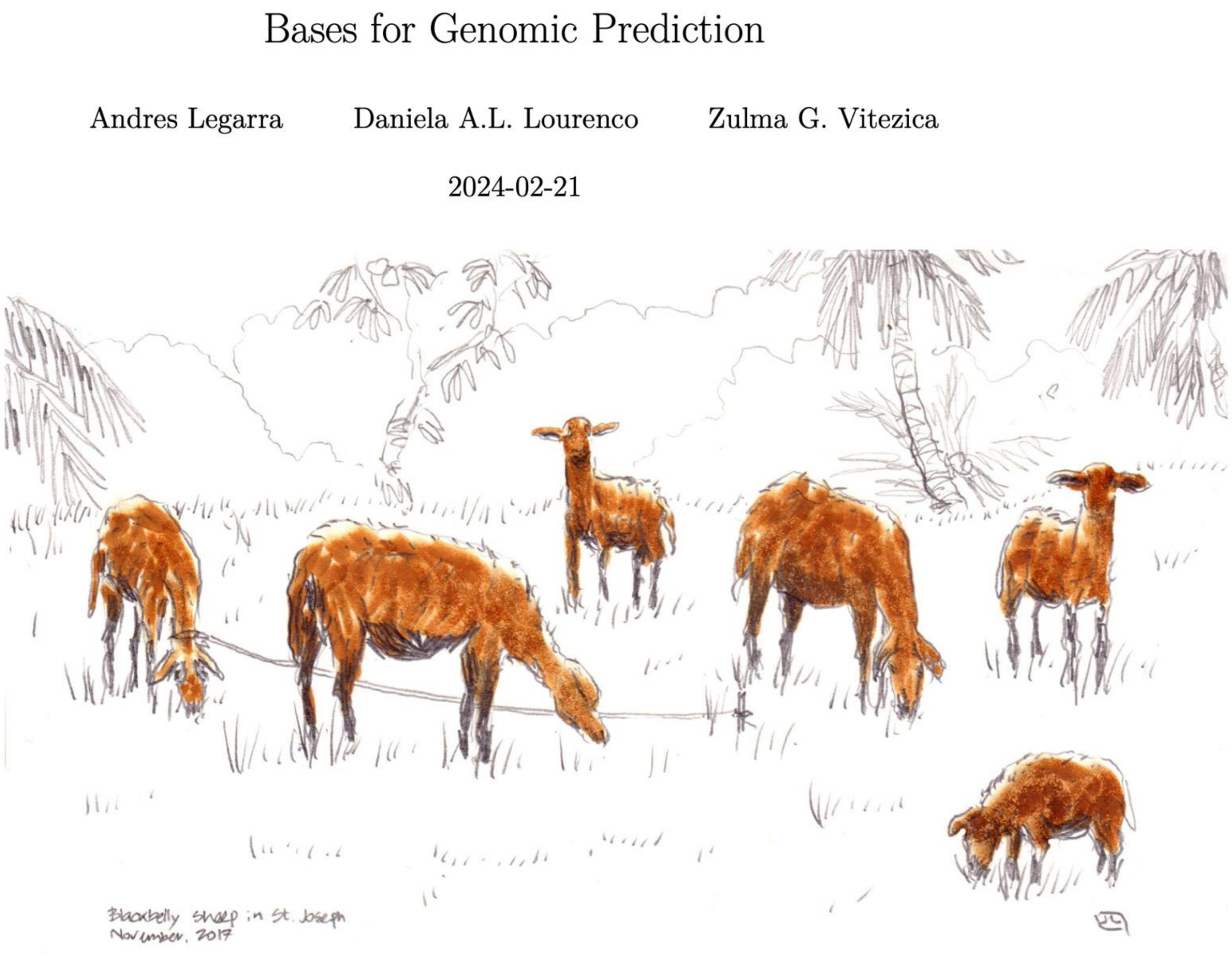
Single-step genomic BLUP enables joint analysis of disconnected breeding programs: an example with *Eucalyptus globulus* Labill.

Andrew N. Callister ,[1,]* Ben P. Bradshaw,[2] Stephen Elms,[3] Ross A. W. Gillies,[3] Joanna M. Sasse,[4] and Jeremy T. Brawner[5]



Blupf90 programs

Practice



Bases for Genomic Prediction

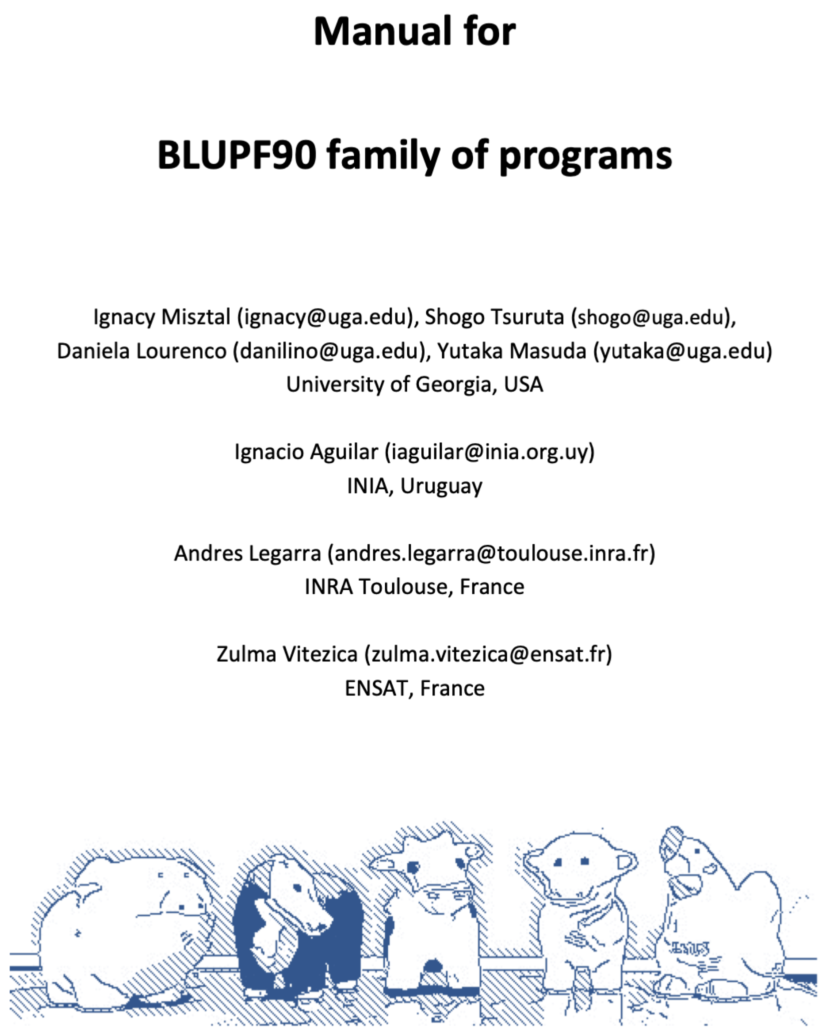
Andres Legarra Daniela A.L. Lourenco Zulma G. Vitezica

2024-02-21

https://genoweb.toulouse.inra.fr/~alegarra/GSIP.pdf

Thanks!

http://nce.ads.uga.edu/html/projects/programs/docs/blupf90_all8.pdf



Manual for

BLUPF90 family of programs

Ignacy Misztal (ignacy@uga.edu), Shogo Tsuruta (shogo@uga.edu),
Daniela Lourenco (danilino@uga.edu), Yutaka Masuda (yutaka@uga.edu)
University of Georgia, USA

Ignacio Aguilar (iaguilar@inia.org.uy)
INIA, Uruguay

Andres Legarra (andres.legarra@toulouse.inra.fr)
INRA Toulouse, France

Zulma Vitezica (zulma.vitezica@ensat.fr)
ENSAT, France