



UNIVERSITY OF
GEORGIA

ssGBLUP

Daniela Lourenco
UGA USA

Andres Legarra
INRA France

Ignacio Aguilar
INIA Uruguay

About prediction methods ...

No Genotypes – Only Pedigree

$$\text{BLUP} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

No Pedigree – Only Genotypes

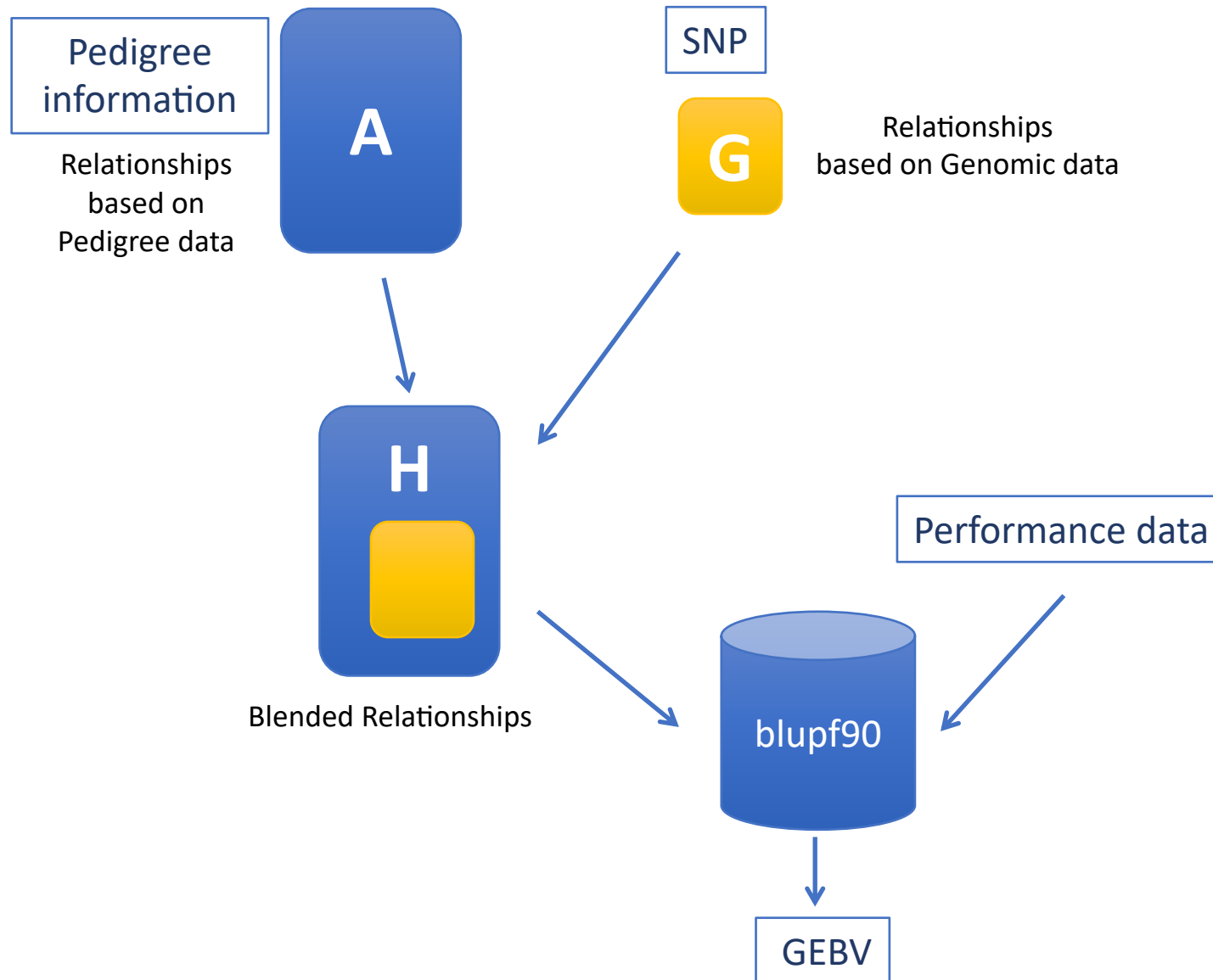
$$\text{GBLUP} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\text{SNP-BLUP} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{M} \\ \mathbf{M}'\mathbf{X} & \mathbf{M}'\mathbf{M} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{M}'\mathbf{y} \end{bmatrix}$$

Pedigree + Genotypes

$$\text{ssGBLUP} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{H}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Single-Step Genetic Evaluation



H adjusts relationships for ungenotyped animals

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

Pedigree
Relationship
Matrix (**A**)

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

Genomic
Relationship
Matrix (**G**)
for animals 3 and 4

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

Realized
Relationship
Matrix (**H**)

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

Understanding the H matrix

- It is a projection of **G** matrix on the rest of individuals “so that” **G** matrix makes sense
 - e.g. parents of two animals related in **G** should be related in **A**
- It is a Bayesian updating of the pedigree matrix based on new information from genotypes
- Typically
 - **A**⁻¹ in the millions but extremely sparse
 - **G** and **A**₂₂ in the thousands
 - Leads to a very efficient method of genomic evaluation:
 - Single Step GBLUP

Some properties of \mathbf{H}

- Semi-positive definite always
 - eigenvalues are always positive or zero
- Positive definite & invertible if \mathbf{G} is invertible
- In practice, if \mathbf{G} is too different (wrong pedigree or genotyping) from \mathbf{A}_{22} , this gives lots of numerical problems
- If everyone is genotyped, Single Step is GBLUP
- If no one is genotyped, Single Step is BLUP

H matrix from Legarra et al. (2009)

- Genomic evaluation would be simpler if all animals were genotyped (2)
- Genomic info can be extended to ungenotyped (1) animals
 - joint distribution of EBV for ungenotyped (u_1) and genotyped (u_2)

$$p(u_1, u_2) = p(u_2)p(u_1|u_2)$$

$$\mathbf{H} = \begin{pmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}$$

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

Combining two sources of relationship

- **A**
 - Contains expected relationships
 - Is limited by the pedigree depth and completeness
 - Depends on accuracy of recording pedigrees
- **G**
 - Contains number of alleles shared between animals weighted by heterozygosity
 - No limitations regarding to number of past generations
 - Depends on allele frequency and quality of genomic data

About the matrices

- Inverse of **H** is used in MME

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

Inverse of the regular pedigree
relationship matrix

Correcting for genomic
relationships

...and avoiding 'double counting'

Computing all matrices before 2016

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

Computed using Henderson-Quaas' algorithm without inbreeding

Computed using Colleau's algorithm, which considers inbreeding

Computed using VanRaden's formula, which considers inbreeding

Initial tests with ssGBLUP

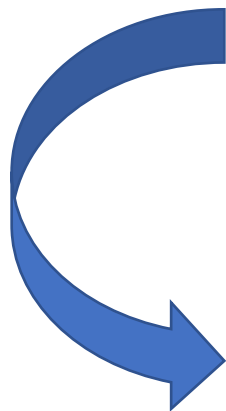
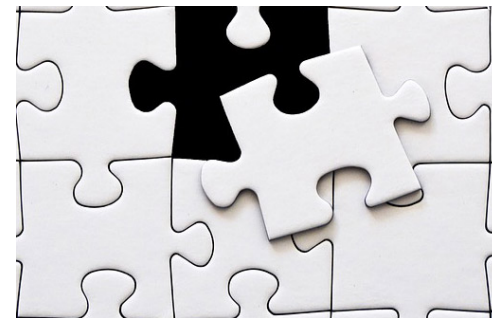
- Tsuruta et al. (2011)
 - US Holsteins final score
 - 8.9M phenotypes | 7.9M pedigree | 17.3k genotypes (6.9k validation)
 - Inflated GEBV for young bulls (validation)
 - Solution: to reduce \mathbf{A}_{22}^{-1}

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

- τ and ω to reduce inflation in GEBV

Experience with simulated data

- Pocrnice et al. (2016)
- Each of the 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
- **Convergence obtained with $\omega = 0.70$**

Computing all matrices after 2016

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

Computed using Henderson-Quaas' algorithm **with inbreeding**

Computed using Colleau's algorithm, which considers inbreeding

Computed using VanRaden's formula, which considers inbreeding

To prepare data for ssGBLUP with inbreeding in A^{-1}

- renumf90

```
EFFECT
1 1 1 cross alpha
RANDOM
animal
OPTIONAL
mat mpe
FILE
aaaped.dat
FILE_POS
1 2 3 4 5
SNP_FILE
allsnp.dat_clean
PED_DEPTH
4
INBREEDING
pedigree
(CO)VARIANCES
28.904825
```

Compatibility between G and A_{22}

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

$$G^{-1} - A_{22}^{-1} \quad \text{Inflation/deflation}$$

$$G^{-1} - A_{22}^{-1} \quad \text{Bias}$$

$$G^{-1} - A_{22}^{-1} \quad \text{Inflation/deflation???$$

Blending and compatibility

- These are two different things
- Many people do not understand this
- “compatibility” tries to put **G** and **A** in the same scale
- “blending” : $\mathbf{G} = \mathbf{G} * 0.95 + \mathbf{A}_{22} * 0.05$
 - used to have an invertible **G**
 - assigns part of the genetic variance to pedigree – not markers

Options for Blending and compatibility

- **Blending**
- OPTION AlphaBeta alpha beta
 - $G = \alpha * G + \beta * A_{22}$
- Compatibility
- OPTION tunedG
 - 0: no adjustment
 - 1: $\text{mean}(\text{diag}(G))=1$, $\text{mean}(\text{offdiag}(G))=0$
 - 2: $\text{mean}(\text{diag}(G))=\text{mean}(\text{diag}(A_{22}))$,
 $\text{mean}(\text{offdiag}(G))=\text{mean}(\text{offdiag}(A_{22}))$ (default)
 - 3: $\text{mean}(G)=\text{mean}(A_{22})$
 - 4: Use Fst adjustment Powell et al. (2010) & Vitezica et al. (2011)

$$\rho = \frac{1}{n^2} (\sum_i \sum_j A_{22\ ij} - \sum_i \sum_j G_{ij})$$

$$G^* = (1 - \rho / 2) G + \mathbf{1}\mathbf{1}' \rho$$

Forcing \mathbf{G} to be similar to \mathbf{A}_{22}

- Vitezica et al. (2011) and Christensen et al. (2012) provided an unbiased method that forces the same genetic base across \mathbf{G} and \mathbf{A}_{22} :

$$\mathbf{G}^* = a + b\mathbf{G}$$

- a accounts for old relationships among non-genotyped ancestors
- b accounts for reduction in the genetic variance

$$a + b \bar{\mathbf{G}} = \bar{\mathbf{A}}_{22}$$

$$a + b \overline{\text{diag}(\mathbf{G})} = \overline{\text{diag}(\mathbf{A}_{22})}$$

Forcing \mathbf{G} to be similar to \mathbf{A}

Recipe (default in blupf90)

- Compute \mathbf{G} with current allele frequencies
- Compute \mathbf{A}_{22}
- Solve equations $a + b \bar{\mathbf{G}} = \bar{\mathbf{A}}_{22}, a + b \overline{\text{diag}(\mathbf{G})} = \overline{(\text{diag}(\mathbf{A}_{22}))}$
- Get new $\mathbf{G}^* = a + b\mathbf{G}$
- Build final $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{*-1} - \mathbf{A}_{22}^{-1} \end{pmatrix}$

Does actually G resemble A_{22} ?

- If pedigree is good and genotyping is good, yes it does!
- Usually
 - $Cor(A_{22_{ij}}, G_{ij}) \approx 0.8$
 - If $Cor(A_{22_{ij}}, G_{ij}) > 0.95$ genomic is not so informative
 - If $Cor(A_{22_{ij}}, G_{ij}) < 0.5$ mislabeling of samples or heterogeneous population
 - $Cor(F_{pedigree_i}, F_{genomic_i}) \approx 0.5$
- Useful for quality control

Main scaling parameters in ssGBLUP

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

$$\mathbf{G} = (\alpha \mathbf{G} + \beta \mathbf{A}_{22})$$

- **Blending**
- makes \mathbf{G} positive-definite
- α = from 0.95 to 0.80
- Improves convergence

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau(\alpha \mathbf{G} + \beta \mathbf{A}_{22})^{-1} - \omega \mathbf{A}_{22}^{-1} \end{bmatrix}$$

- **Scaling**
- τ and ω
- Used for compatibility between \mathbf{G}^{-1} and \mathbf{A}_{22}^{-1}
- Reduces inflation

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau((\alpha \mathbf{G} + \beta \mathbf{A}_{22}) + \mathbf{1}\mathbf{1}'\mathbf{a})^{-1} - \omega \mathbf{A}_{22}^{-1} \end{bmatrix}$$

- **Tuning**

- tunedG

- Accounts for selection in \mathbf{A}

- Improves accuracy and reduces bias

$$\mathbf{a} = \frac{1}{n^2} \left(\sum_i \sum_j \mathbf{A}_{22} - \sum_i \sum_j \mathbf{G} \right)$$

Should τ and ω be used in ssGBLUP evaluations?

- Need for scaling parameters depend on compatibility among matrices
 - Pedigree, genomic, pedigree for genotyped animals
- Most of the compatibility problems are caused by
 - Ignoring inbreeding in \mathbf{A}^{-1}
 - Missing pedigrees for genotyped animals
 - Wrong definition of UPG
 - Ignoring inbreeding for UPG

Validation of genomic models