



# A fast method to fit the mean of unselected base animals in Single-Step SNP-BLUP

Tribout T <sup>1</sup>, Boichard D <sup>1</sup>, Ducrocq V <sup>1</sup>, Vandenplas J <sup>2</sup>

<sup>1</sup>GABI, INRA, AgroParisTech, Université Paris-Saclay, France

<sup>2</sup>Wageningen University & Research, Animal Breeding and Genomics, The Netherlands



# Context

Single-Step GBLUP: current reference method for genomic evaluation  
allows joint use of pedigree, phenotypic and genomic information

Breeding Values Model (BVM):

*Legarra et al (2009), Aguilar et al (2010), Christensen & Lund (2010)*

$$y = X\beta + Zu + e$$

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_g^2} H^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

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

$$A = \begin{bmatrix} A_{nn} & A_{ng} \\ A_{gn} & A_{gg} \end{bmatrix}$$

# Context

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \frac{\sigma_e^2}{\sigma_g^2} H^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \quad H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{gg}^{-1} \end{bmatrix}$$

$A$  : mean genetic value of the base population = 0

$G$  : genotyped animals = selected animals

→ center the observed genotypes to the base allele frequencies (**unknown !**)

Proposed solutions:

*Vitezica et al (2011)*:  $G^* = G + 11'\alpha$ , where  $\alpha = \text{mean } A_{gg}(i,j) - \text{mean } G(i,j)$

*Christensen et al (2012)*:  $G^* = \beta G + 11'\alpha$

...

# Marker Effects Models

Growing genotyping:  $G$  larger and larger !

→ Marker Effects Models (MEM) for Single-Step GBLUP

Fernando et al (2016), Taskinen et al (2017); Liu et al (2014), ...

$$\begin{bmatrix} X'X & X'_g Z'_g M_g & X'_n Z_n \\ M'_g Z'_g X_g & Q & M'_g A^{gn} \frac{\sigma_e^2}{\sigma_g^2} \\ Z'_n X_n & A^{ng} M_g \frac{\sigma_e^2}{\sigma_g^2} & Z'_n Z_n + A^{nn} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \\ \hat{u}_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'_g Z'_g y_g \\ Z'_n y_n \end{bmatrix} \quad A^{-1} = \begin{bmatrix} A^{nn} & A^{ng} \\ A^{gn} & A^{gg} \end{bmatrix}$$

Same problem:

- genotypes  $M_g$  should also be centered to the base allele frequencies
- $G$  and  $A_{gg}$  not built ?

# Marker Effects Models

Proposed solution: *Fernando et al (2014), Hsu et al (2017)*

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} X_n \\ X_g \end{bmatrix} \beta + \begin{bmatrix} Z_n J_n \\ Z_g J_g \end{bmatrix} \mu_g + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g M_g \end{bmatrix} \begin{bmatrix} u_n \\ \alpha \end{bmatrix} + \begin{bmatrix} e_n \\ e_g \end{bmatrix}$$

= GEBV

$$J_g = \begin{bmatrix} -\mathbf{1} \\ \vdots \\ -\mathbf{1} \end{bmatrix}$$

$$J_n = -A_{ng} A_{gg}^{-1} \mathbf{1}$$

$$= (A^{nn})^{-1} A^{ng} \mathbf{1}$$

$A^{nn}$  is very sparse  $\rightarrow (A^{nn})^{-1}$  : Cholesky factorization of  $A^{nn}$

$(A^{nn})^{-1}$  also needed in MEM-SS GBLUP:  $M'_n A^{nn} M_n = M'_g A^{gn} (A^{nn})^{-1} A^{ng} M_g$

$$\begin{bmatrix} X'X & X'_g Z_g M_g & X'_n Z_n \\ M'_g Z'_g X_g & \textcircled{Q} & M'_g A^{gn} \frac{\sigma_e^2}{\sigma_g^2} \\ Z'_n X_n & A^{ng} M_g \frac{\sigma_e^2}{\sigma_g^2} & Z'_n Z_n + A^{nn} \frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \\ \hat{u}_n \end{bmatrix} = \begin{bmatrix} X'y \\ M'_g Z'_g y_g \\ Z'_n y_n \end{bmatrix}$$

# Cholesky factorization of $A^{nn}$

In large popul.,  $A^{nn} > 10\text{M} \times 10\text{M}$ , but very sparse:  $\approx 4$  NZE / non-genot. indiv.

L is also sparse

	Montbéliarde	Holstein	Limousin	Charolais
Nb of non-genotyped animals	3.7M	19.7M	5.6M	9.7M
Nb of NZE in L / non-gen. ani.	28	62		

Holstein: Cholesky Factorization of  $A^{nn} = 45$  minutes with MKL on 6 CPU

# Cholesky factorization of $A^{nn}$

In large popul.,  $A^{nn} > 10\text{M} \times 10\text{M}$ , but very sparse:  $\approx 4$  NZE / non-genot. indiv.

L is also sparse, but L is much less sparse in **beef cattle** than in **dairy cattle**:

	Montbéliarde	Holstein	Limousin	Charolais
Nb of non-genotyped animals	3.7M	19.7M	5.6M	9.7M
Nb of NZE in L / non-gen. ani.	28	62	314	650

Holstein: Cholesky Factorization of  $A^{nn} = 45$  minutes with MKL on 6 CPU

Charolais: Cholesky Factorization of  $A^{nn} = 5$  hours with MKL on 8 CPU

J. Vandenplas, EAAP 2018:

→ Replace  $M'_n A^{nn} M_n v = M'_g (A^{gn} (A^{nn})^{-1} A^{ng}) M_g v$

by  $M'_g (A_{anc}^{gn} (A_{anc}^{nn})^{-1} A_{anc}^{ng} + \Delta) M_g v$

$$A_{anc}^{-1} = \begin{bmatrix} A_{anc}^{nn} & A_{anc}^{ng} \\ A_{anc}^{gn} & A_{anc}^{gg} \end{bmatrix}$$

**Population = genotyped animals  
+ ancestors**

$A_{anc}^{nn}$  : size(ancestors of genotyped animals) << total nb of non-genotyped indiv.

$$A_{anc}^{nn} \approx 10^5 \times 10^5 \quad A^{nn} \approx 10^7 \times 10^7$$

→ Cholesky factor. of  $A_{anc}^{nn}$  is very fast (sec), even in large beef cattle populations






but  $(A^{nn})^{-1}$  still needed to compute  $J_n = (A^{nn})^{-1} A^{ng} 1 \dots$

→ a more efficient way to calculate  $J_n$   
that does not require  $(A^{nn})^{-1}$

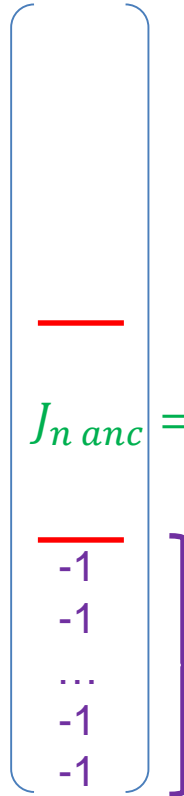
# A more efficient way to calculate $J_n$

Other non-genotyped animals = « oth »  
 2017 CHA : 9,500 K

youngest  
  
 oldest

Non-genotyped animals ancestors of ga = « anc »  
 2017 CHA : 155 K

Genotyped animals = « ga »  
 2017 CHA : 22 K

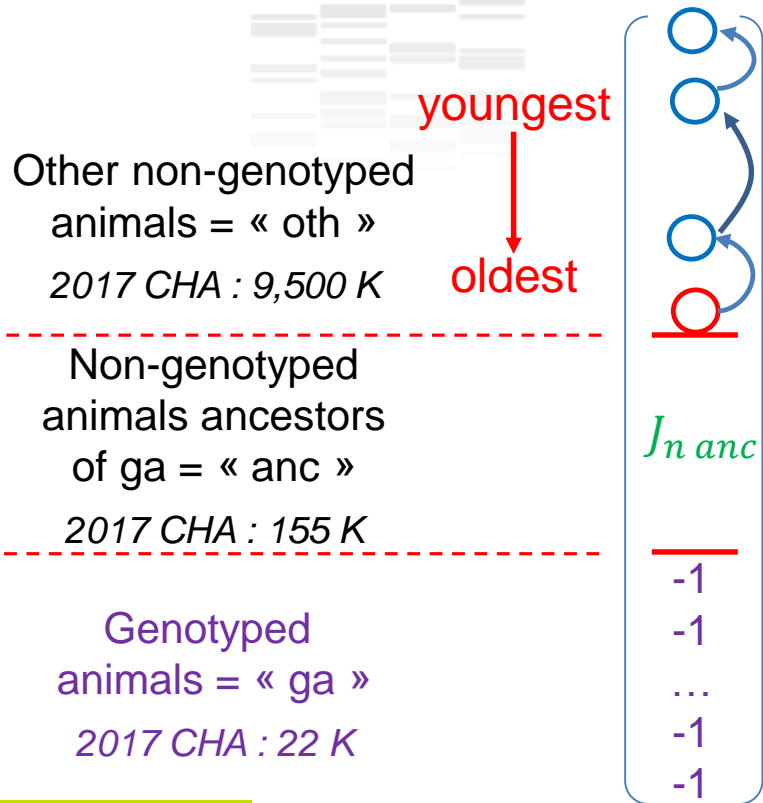


$$J_n^{anc} = \cancel{(A^{nn})^{-1}} A^{ng} 1 = (A_{anc}^{nn})^{-1} A_{anc}^{ng} 1 = (A^{nn})^{-1} A^{ng} 1$$



$$J_g = -1$$

# A more efficient way to calculate $J_n$



from the oldest to the youngest « oth » animal:

$$\gamma_{parent} = 0 \text{ if unknown parent}$$

$$\gamma_{parent} = -1 \text{ if genotyped parent}$$

$$\gamma_{parent} = J_n^{anc \ parent} \text{ if parent among } \langle anc \rangle$$

$$\gamma_{parent} = J_n^{oth \ parent} \text{ if parent among } \langle oth \rangle$$

$$\rightarrow J_n^{oth \ i} = 0.5 * (\gamma_s + \gamma_d)$$

$$= J_n \ i \text{ from } (A^{nn})^{-1} A^{ng} \mathbf{1}$$



# Conclusion

A simple and low cost method to calculate the covariate vector needed to fit the mean of unselected base animals in MEM SS-GBLUP models

- does not require the inverse of the complete  $A^{nn}$  matrix
- computing time saving: hours → seconds in large beef cattle populations
- memory saving

Contributes to make the MEM Single-Step GBLUP models even more independent of the size of the population