Genomic predictability of single-step GBLUP for production traits in US Holstein

Y. Masuda[1], P. M. VanRaden[2], H. L. Bradford[2], A. Legarra[3], I. Misztal[1], and T. J. Lawlor[4]

1 University of Georgia, USA; 2 AGIL, USDA, USA;
3 INRA, France; 4 Holstein Association USA, Inc., USA

Background

- Genomic prediction with single-step GBLUP (ssGBLUP)
  - Genotyped + non-genotyped animals
  - Accountability for pre-selection
  - APY: dimensionality reduction in marker genotypes
- Required: compatibility among relationship matrices

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

  - Reasonable in complete pedigree
  - Missing pedigree: adjustment of $\mathbf{A}_{22}^{-1}$ by ω
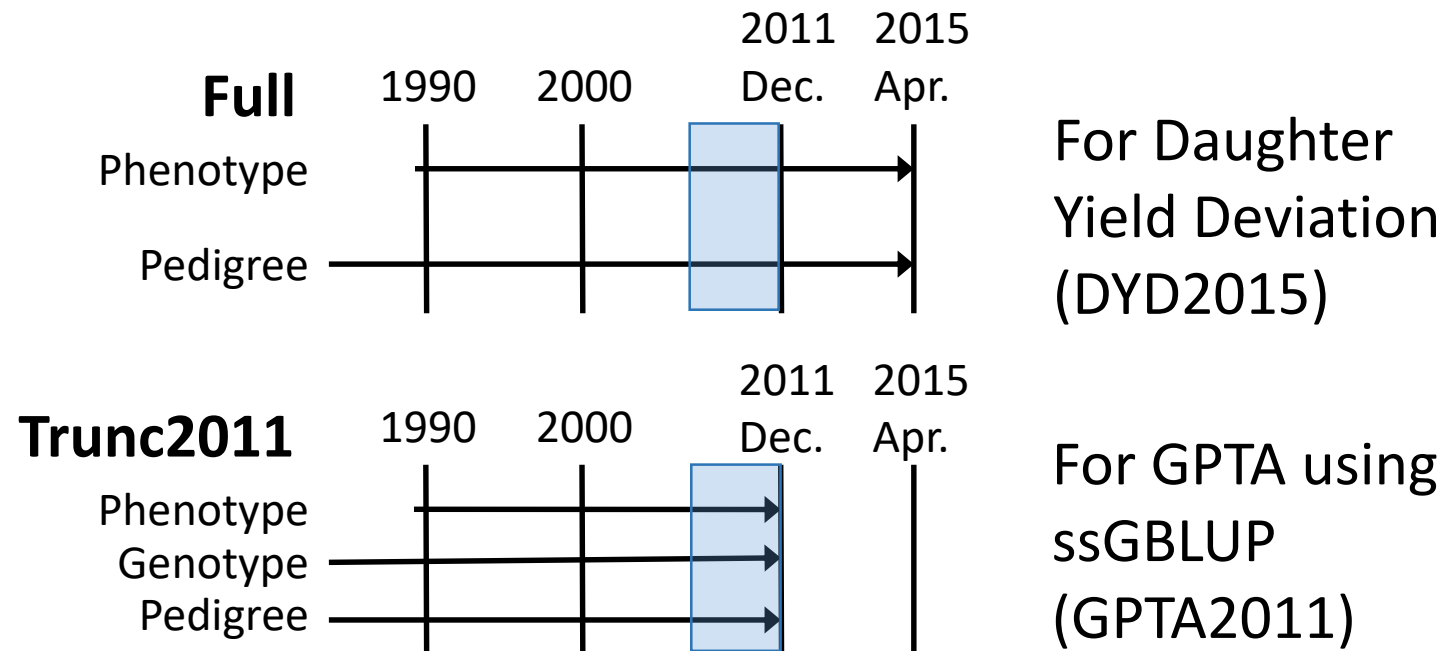 - How to incorporate unknown parent groups (UPG)?

Objectives

- To validate genomic predictions for young bulls by different configurations of UPG in US Holstein

- To discuss possible modifications on $\mathbf{H}^{-1}$ to handle UPG in ssGBLUP: a simulation study

Full data in Holstein

	Description	Number of records/animals
Phenotype	Milk, fat, and protein yield (305-d basis) for US Holstein cows recorded between Jan. 1990 and Apr. 2015	37,259,427
	Cows with phenotype(s)	15,891,366
Pedigree	Animals born in Apr. 2015 or earlier (3-gen. back from phenotyped cows) **185 UPGs**	22,963,255
Genotype	Animals born in Apr. 2015 or earlier (60,671 markers)	764,029

Validation study

Full

2011 2015
1990   2000   Dec.   Apr.

Phenotype

Pedigree

For Daughter
Yield Deviation
(DYD2015)

Trunc2011

2011 2015
1990   2000   Dec.   Apr.

Phenotype
Genotype
Pedigree

For GPTA using
ssGBLUP
(GPTA2011)

Validation Bulls:
Genotyped young bulls
with no tested daughters
in 2011 but with at least
50 tested daughters in
2015 (N=3,797)

$$DYD2015 = b_1 \times GPTA2011 + b_0$$

- R^2 : validation reliability
- Slope (b_1): Inflation of prediction

Different UPG in $\mathbf{H}^{-1}$

1. Weight (ω) on $\mathbf{A}_{22}^{-1}$: **0.9** or **1.0**

2. UPG: **pedigree + genomic UPG**, **pedigree UPG only**, **or no UPG**
   (genomic UPG)   **(pedigree UPG)**

$$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -(\mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1})\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'(\mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1}) & \mathbf{Q}_2'(\mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1})\mathbf{Q}_2 \end{bmatrix}$$

DYD2015 vs GPTA2011 (Protein)

Data					R2	b1
Official GPTA 2011					0.51	0.81

Data	UPG	ω=0.9 R2	b1	ω=1.0 R2	b1
GPTA2011	Genomic UPG	0.39	0.74	0.32	0.51

DYD2015 vs GPTA2011 (Protein)

Data					R2	b1
Official GPTA 2011					0.51	0.81

		ω=0.9 R2	b1	ω=1.0 R2	b1
Data	**UPG**				
GPTA2011	Genomic UPG	0.39	0.74	0.32	0.51
	Pedigree UPG	0.50	0.96	0.52	0.78
	No UPGs			0.50	0.78

Low accuracy with genomic UPG

- GPTA for young genotypes
  - Pedigree UPG: $\qquad GPTA = w_1 PA + w_2 DGV - w_3 PI \qquad\qquad \approx DGV$
  - Genomic UPG: $\qquad GPTA = w_1 PA + w_2 DGV - w_3 PI + w_4 UPG \quad \approx DGV + UPG$

  Larger weights with many genotypes **Too large for young animals**

- Possible solutions
  - Just using pedigree UPG
  - Discounting UPG effects
  - Removing double counting between DGV and UPG
  - Scaling $\mathbf{A}$ to $\mathbf{G}$ ("metafounders")

$$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \end{bmatrix}$$

Missing parents in ssGBLUP

- Genomic UPG

  - $\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{G}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{G}^{-1} & \mathbf{Q}_2'\mathbf{G}^{-1}\mathbf{Q}_2 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{A}_{22}^{-1} & \mathbf{Q}_2'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \end{bmatrix}$

- Genomic UPG without Q'G^{-1}Q

  - $\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}$ $\hspace{2cm} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{A}_{22}^{-1} & \mathbf{Q}_2'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \end{bmatrix}$

- Pedigree UPG

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

With A_{22}^*

- Genomic UPG
  - $$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A_{22}^{-1}} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{G}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{G}^{-1} & \mathbf{Q}_2'\mathbf{G}^{-1}\mathbf{Q}_2 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{A}_{22}^{-1} & \mathbf{Q}_2'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \end{bmatrix}$$

- Genomic UPG without Q'G⁻¹Q
  - $$\mathbf{H}^* = \mathbf{A}^* + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A_{22}^{-1}} & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad\qquad - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'\mathbf{A}_{22}^{-1} & \mathbf{Q}_2'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 \end{bmatrix}$$

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

  Obtained by absorbing UPG-related submatrices into $\mathbf{A}_{22}$

A^*_{22}: Modification with UPG

- Indirect inversion

$$A^{-1}_{22} = A^{22} - A^{21}(A^{11})^{-1}A^{12}$$

where $A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$

- With UPG

$$A^*_{22} = A^{22} - [A^{21} \quad A^{23}]\begin{bmatrix} A^{11} & A^{13} \\ A^{31} & A^{33} + I \end{bmatrix}\begin{bmatrix} A^{12} \\ A^{32} \end{bmatrix}$$

where $A^* = \begin{bmatrix} A^{11} & A^{12} & A^{13} \\ A^{21} & A^{22} & A^{23} \\ A^{31} & A^{32} & A^{33} \end{bmatrix} = \begin{bmatrix} A^{11} & A^{12} & -[A^{11} \quad A^{12}]Q \\ A^{21} & A^{22} & -[A^{21} \quad A^{22}]Q \\ -Q'\begin{bmatrix} A^{11} \\ A^{21} \end{bmatrix} & -Q'\begin{bmatrix} A^{12} \\ A^{22} \end{bmatrix} & Q'A^{-1}Q \end{bmatrix}$

Simulation study

- Structure
  - $h^2 = 0.3$
  - Sex-limited trait (n = 90,000)
  - EBV selection
  - 10 generations (n = 164,500)
  - Ne: 200 theoretical; 25 realized
  - Mean F in last generation: 0.11
- Genotypes
  - 18,674 total
  - 5108 in gen. 10 for validation

- Assignment of UPGs
  - UPG1 for generation 0-4
  - UPG2 for generation 5-7
  - UPG3 for generation 8-10

Category	Non genotyped	Genotyped
Top bulls	0	0
Top cows	5% (dam)	0
Bottom bulls	30% (dam)	10% (dam)
Bottom cows	30% (dam)	10% (dam)

Results from simulation

	Standard A_{22}^{-1}	
	R2	b1
Genomic UPG	0.53	0.86

* Genotyped young animals without records

Results from simulation

	Standard A_{22}^{-1}		Modified A_{22}^*	
	R2	**b1**	**R2**	**b1**
Genomic UPG	0.53	0.86		
without Q'G^{-1}Q	0.62	1.05		
Pedigree UPG	0.63	1.06	0.63	1.06

	R2	**b1**
Metafounders	0.63	1.08

* Genotyped young animals without records

Summary

- Missing pedigree may reduce the accuracy of genomic prediction in single-step GBLUP.
  - Specific data structure with many missing parents

- We have several options to discount the possible bias of UPG contributions in $\mathbf{H}^*$.
  - Removal of $\mathbf{G}^{-1}$ from the additional UPG contribution
  - Use of Modified $\mathbf{A}_{22}^{-1}$
  - Metafounders

Acknowledgement