# Pre-selection bias and validation in single-step GBLUP for production traits in US Holstein

Y. Masuda<sup>1</sup>, P. M. VanRaden<sup>2</sup>, I. Misztal<sup>1</sup>, and T. J. Lawlor<sup>3</sup>

1 University of Georgia, USA 2 AGIL, USDA, USA 3 Holstein Association USA, Inc., USA

WCGALP 2018, February 11-16, Auckland, New Zealand

## Genomic selection in practice

#### Implementation

#### **Statistical Methods**

#### Genomic tools

- Use of existing data-collection systems (phenotypes & pedigrees)
- Integration of genomic data
- Stabilization of genomic predictions
- Breeding value prediction
- Adjustments for bias-reduction
- Computing algorithms
- Efficient genotyping technique
- Affordable SNP chips

# Dairy cattle evaluation

Conventional animal-model BLUP





# Multi-step evaluation

#### Advantages

- Keeping the traditional systems
- Flexibly adjustable for GPTA (GEBV) in terms of bias
- Accumulated experience
- Concerns
  - Only for genotyped animals
  - Too many options for the second step (input values & methods)
  - "Pre-selection bias" in the traditional PTA





- Selection criteria not included in MME of animal model BLUP
- Bias down in the prediction



- Biased PTA (EBV) to GPTA
- GPTA biased down



• Possible result: underestimated genetic trend for genomically selected animals





# Single-step GBLUP

#### Single-step GBLUP



#### • Advantages

- Expected accountability for genomic pre-selection
- Use of genotyped & nongenotyped in the same equations
- Simplicity
- Concerns
  - Computational costs (solved)
  - Is it reliable for genomic prediction in dairy cattle?

# Production traits of US Holsteins

- Comparison of genetic trends
  - Single-step GBLUP GPTA vs the traditional PTA (Data up to 2015)
  - Multi-step official GPTA vs the corresponding PTA (Published in 2016)
- Validation reliability for young bulls
  - 4-year truncated data
  - DYD in 2015 vs GPTA in 2011
- Computational feasibility
  - APY: Algorithm for Proven and Young for  ${\bf G}^{-1}$



#### Full data

Data	Description	Number of records
Phenotypes	305-d Milk, fat, and protein yield from US Holsteins; from 1990 to 2015	50,970,954
Pedigree	3 generations back from phenotyped cows or genotyped animals; 215 unknown-parent groups (UPG)	29,651,623
Genotypes	Both male and female; including young bulls and heifers (#SNPs = 60671)	764,029

Three-trait repeatability model; same as in the official evaluation.

# $\mathbf{H}^{-1}$ and GPTA

- Mixed model equations  $\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{H}^{-1} \otimes \boldsymbol{\Sigma}_{g}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$
- Inverse relationship matrix

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

• GPTA of a young animal  $GPTA = w_1PA + w_2DGV - w_3PI$ 

•  $\omega$ : Constant to compensate for missing pedigrees ( $\omega = 1$  for the full data).



Aguilar et al. (2010)

## APY: Algorithm for Proven and Young

- Genotyped animals into two groups: "core" and "non-core"
  - Assumption:

$$\mathbf{u}_n = \mathbf{P}\mathbf{u}_c + \mathbf{\Phi}$$

- BV for non-core  $(\mathbf{u}_n)$  is a linear function of BV for core  $(\mathbf{u}_c)$ .
- APY G-inverse (Misztal et al. 2016)

$$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} + \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}^{-1}\mathbf{G}_{cn}'\mathbf{G}_{cc}^{-1} & \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}^{-1} \\ -\mathbf{M}^{-1}\mathbf{G}_{cn}'\mathbf{G}_{cc}^{-1} & \mathbf{M}^{-1} \end{bmatrix}$$



• **M**<sup>-1</sup>: Diagonal matrix

APY  $G^{-1}$ 



• Sparse



- Easy computations
- Gives the same GPTA as the regular G<sup>-1</sup> using a few core animals (Fragomeni et al., 2015)
- How to choose core animals?
  - How many? dimensionality of G
  - Which animals? random choice

# Dimensionality of ${\bf G}$



- Dim. of  $G \approx M_e$ 
  - *M<sub>e</sub>*: the number of independent chromosome segments
    - = the optimal number of core animals
- Estimate of dim. of G
  - *M<sub>e</sub>* ≈ the # of largest eigenvalues explaining the most (98%) of variation in G
  - 18,359 cores for 760K US Holsteins

# Which core animals?





Bradford et al. (2017)

- The best practice:
  - Core animals covering all generations.
  - Or, just randomly choose the core.
- Core animals represent independent chromosome segments in the populations.
- In this study:
  - 18,359 random core animals

# Inbreeding and UPG

• QP-transformation for  $A^{-1}$  (Westell et al., 1988; Quaas 1988)

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{Q} \\ -\mathbf{Q}'\mathbf{A}^{-1} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q} \end{bmatrix}$$
: Henderson's rule with inbreeding

• QP-transformation for  $H^{-1}$  (Misztal et al., 2013)

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

# Computing time

Preparation	Traditional BLUP	Single-step GBLUP	
$\mathbf{G}_{APY}^{-1}$	N/A	6 h 53 min	
Other	9 min	48 min	
Subtotal in preparation	9 min	7 h 41 min	
Iteration	Traditional BLUP	Single-step GBLUP	
Number of iterations	402	464	
Time per PCG iteration	51 sec	83 sec	
Post-processing	12 min	13 min	
Subtotal in iterations	5 h 53 min	10 h 54 min	

Intel Xeon X7650 (2.26 GHz; 20 cores for preparation and 6 cores for iterations)

**Computationally feasible** 

#### **Cows : ssGBLUP vs traditional PTA (protein)**



\*Cows with record(s)

#### **Cows : ssGBLUP vs traditional PTA (protein)**



**Bulls: ssGBLUP vs traditional PTA (protein)** 



\*Genotyped bulls with at least 10 daughters with record(s)

#### Bias in genotyped cows and bulls



Larger bias in cows: shorter generation interval & lower reliability of GPTA



# Adjustments on the official PTA

- Official PTA adjusted by Wiggans et al. (2012)
  - Cow trend aligned to bull trend (Reduction in bias for cows)
  - Same trend in PTA and GPTA
- Additional adjustments in the official evaluation
  - Breed difference
  - Inbreeding



# Validation study



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<sup>2</sup> value: validation reliability
- Slope (*b*<sub>1</sub>): Bias of prediction

# Configurations in $\mathbf{H}^{-1}$

- 1. Weight ( $\omega$ ) on  $A_{22}^{-1}$ : **0.9** or **1.0**
- 2. UPG: pedigree only, pedigree + genomic UPG, or no 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	<b>b1</b>
Official GPTA 2011				0.51	0.81
		ω=0.9		ω=1.0	
Data	UPG	R2	<b>b1</b>	<b>R2</b>	<b>b1</b>
Truncated 2011	Pedigree	0.50	0.96	0.52	0.78
	Ped. + Genomic	0.39	0.74	0.32	0.51
	No UPGs			0.50	0.78

**Different predictions by UPG** 

## Incomplete pedigree on accuracy & inflation







\* Simulated data (Bradford et al., 2017)

## Low accuracy with genomic UPG

- GPTA for young genotypes
  - No UPG:  $GPTA = w_1 PA + w_2 DGV w_3 PI \approx DGV$
  - With UPG:  $GPTA = w_1PA + w_2DGV w_3PI + w_4UPG \approx DGV + UPG$

Larger weights with many genotypes

Not needed for young animals

- Specific pattern of missing pedigree
  - Production traits: many grade animals
  - No problem in Finland (Koivula et al. 2017) or for US type traits (Tsuruta 2017)
- Solutions: research in progress
  - Option: only DGV for young genotypes
  - Metafounders



## Indirect prediction

- Optional step to ssGBLUP
  - 1. Compute DGV  $\widehat{\boldsymbol{u}}$  from ssGBLUP without young animals
  - 2. Compute SNP effects as  $\hat{\mathbf{a}} = k\mathbf{Z}'\mathbf{G}\hat{\mathbf{u}}$ .
  - 3. Compute DGV for young animals as  $\hat{\mathbf{u}}_{young} = \mathbf{Z}\hat{\mathbf{a}}$ .
- Successfully applied to Angus & simulated data



Bradford et al. (2017)

## Metafounders

- Regular ssGBLUP: scaling G to A; reasonable in complete pedigree
- Metafounders: scaling A to G
  - Treat UPG as metafounders
  - Estimate genomic relationships among metafounders ( $\boldsymbol{\Gamma}$ ) using  $\boldsymbol{G}$
  - Construct  $A^{-1}$  and  $A^{-1}_{22}$  with  $\Gamma$  using the Henderson's and Collau's methods
- Final form:

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



Legarra et al. (2015)

# Summary

- The traditional PTA for genotyped animals are likely underestimated; Needs adjustments in multi-step methods.
- Single-step GBLUP can account for the pre-selection bias.
- Single-step GBLUP may give a reasonable genetic trend without adjustments.
- Missing pedigree may reduce predictability of genomic predictions.
- We can recover the predictability for young animals; research in progress.

# Acknowledgement

- USDA NIFA (2015-67015-22936) and Holstein Association USA for financial support.
- Council of Dairy Cattle Breeding for phenotype, genotype, and pedigree data.
- John Cole and Melvin Tooker (USDA-AGIL) for preparing the initial data sets and a computing environment.





Daniela A. L. Lourenco



lgnacy Misztal



Yvette Steyn



Andre Garcia





Andres Legarra



Instituto Nacional de Investigación Agropecuaria U R U G U A Y



lgnacio Aguilar



Breno Fragomeni



lvan Pocrnic



Rafael

Silva



Shogo Tsuruta





Heather Bradford