Use of causative variants and SNP weighting in a single-step GBLUP context

Breno Fragomeni¹, Lourenco DAL¹, Legarra A², Tooker ME ³, VanRaden PM³, Lawlor, T.⁴, Misztal I¹

¹University of Georgia, Athens, USA
²INRA, Castanet-Tolosan, France
³AGIL ARS-USDA, Beltsville, USA
⁴American Holstein Association, Brattleboro, USA
Motivation

• Decreasing costs of whole genome sequence
• Revived interest in causative variants for prediction
• Several authors are finding and using causative variants
  • No improvement:
    • Binsbergen et al., 2015 and Erbe et al., 2016
  • Up to 5% improvement:
    • Brondum et al. 2015 and Vanraden et al., 2017
Motivation

• ssGBLUP was able to reach accuracies close to 1 in simulation
  • Simulated QTN position and effects known
  • GWA estimated weights had limited impact
• GWA Methodology – no limitation in minimum and maximum weights (Zhang et al., 2016)
Objective

• Test different SNP weighting methods in GBLUP and ssGBLUP in field data with the inclusion of causative variants
Field Data

- 4M Records for Stature
- 3M Cows
- 4.6M Animals in pedigree
- $h^2=0.44$
- 27k Genotyped Sires
  - 54k SNP
  - 54k SNP + 17k Causative Variants (VanRaden et al., 2017)
Analysis

• **GBLUP**
  - Multi-step approach
  - Daughter deviation as phenotypes
  - Genomic Relationship Matrix
  - Homogeneous or heterogeneous residual variance – different reliabilities

• **ssGBLUP**
  - Same model as national evaluation for type traits
  - No deregressions
  - Matrix combining pedigree and genomic information \((H)\)
Weighted genomic relationship matrix

\[ G = ZDZ' \frac{\sigma_s^2}{\sigma_a^2} = \frac{ZDZ'}{\sum_i 2p_i q_i} \]

- Default
  \[ \hat{d}_i = 1 \]
  \[ (\text{VanRaden et al., 2008}) \]

- Linear weights
  \[ \hat{d}_i \sim \sigma_{SNP_i}^2 \]
  \[ (\text{Zhang et al., 2010}) \]

- Non-linear A weights
  \[ \hat{d}_i = 1.125 \frac{|\hat{u}_i|}{sd(u)} - 2 \]
  \[ \text{Value capped at 10} \]
  \[ (\text{VanRaden et al., 2008}) \]

- Fast-Bayes A
  \[ \hat{d}_i = \frac{\hat{u}_i^2 + df*S^2}{df+1} \]
  \[ (\text{Sun et al., 2012}) \]
Weight matrix elements

- Linear
- Fast Bayes-A
- Non Linear A

SNP Effect/SNP SD vs Weight

- Linear
- Fast Bayes-A
- Non Linear A
Simulation results

![Graphs showing simulation results](image-url)
GBLUP – 54K SNP - Reliabilities

**HOMOGENEOUS RESIDUAL VARIANCE**

- No Weights: 54.9
- Linear: 54.5
- NonLinear: 56.6

**HETEROGENEOUS RESIDUAL VARIANCE**

- No Weights: 58.7
- Linear: 56.3
- NonLinear: 58.6
GBLUP and ssGBLUP – 54K SNP - Reliabilities

\[ GEBV = w_1PA + w_2DGV \]
Including causative variants

![Bar chart showing HETEROGENEOUS RESIDUALS for different models and datasets.](image)

- **GBLUP - Homogeneous**: 54.9, 55.5
- **GBLUP - No Weights**: 58.2, 58.7
- **GBLUP - Linear**: 56.4, 56.3
- **GBLUP - NonLinear**: 58.4, 58.9
- **ssGBLUP - GEBV**: 60.9, 60.8
- **ssGBLUP - DGV**: 59.5
Inflation coefficient: $b_1$
Conclusions

• Gains with causative variants have more impact in GBLUP than in ssGBLUP
  • More data is used in single-step methodology, therefore impact of prior is less important
  • Sequence data might mask or fix methodology problems

• Non-linear methodology is better for weighting marker effects than linear weights

• Estimating weights in single-step GBLUP is still a research topic