GenTORE

Genomic management Tools to Optimise Resilience and Efficiency

Grant agreement n°: 727213

H2020 - Research and Innovation Action

D4.5

Paper on genomic prediction of growth efficiency for beef bulls on different dairy breeds

Due date: M57 (February 2022)

Actual submission date: M57 (February 2022)

Project start date: 1st June 2017 **Duration:** 60 months

Workpackage concerned: WP4

Concerned workpackage leader: Mogens Sandø Lund

Lead Beneficiary: AU

Dissemination level:

X PU: Public (must be available on the website)

☐ **CO:** Confidential, only for members of the consortium (including the Commission Services)

☐ **CI:** Classified, as referred to in Commission Decision 2001/844/EC

Table of content

1. Summary

This paper corresponds to deliverable D4.5 of GenTORE workpackage 4 (*Genomic management tools to optimise resilience and efficiency).* This contribution shows the superiority of a genomic model developed in GenTORE WP4 for multibreed and crossbred animals when compared to state of the art methods on real industry data. The objective of this study was to evaluate genomic prediction for growth efficiency for a multi-breed Irish mixed beef and dairy cattle population, using a breed origin of alleles (BOA) model. Our BOA model uses one matrix with allele counts for each breed to predict breed-specific marker effects, enabling to use crossbred and purebred animals in the reference population. Pedigree-based best linear unbiased prediction (PBLUP), a genomic model assuming a homogeneous population (SNP-BLUP) and BOA model were compared. Accuracy was estimated as the correlation between breeding values and corrected phenotype divided by the square root of heritability. The accuracy of predictions using BOA increased by 82% and 42% for crossbred and purebred animals, respectively, compared with PBLUP, and improvements of BOA over SNP-BLUP were of 6% for crossbred animals.

2. Introduction

Genomic prediction (GP) has been rapidly implemented in cattle due to its ability to accurately predict the genomic values of selection candidates early in life. Accuracy of genomic prediction relies on the linkage disequilibrium (LD) between genetic markers, usually single nucleotide polymorphisms (SNPs), and causal mutations. The stronger the LD, the more accurately will be the prediction of the direct genomic breeding values (DGVs). The effects of causal mutations captured by the SNPs is not consistent across distantly related populations due to differences in LD patterns. Additionally, some causal mutations may not segregate in all breeds, or the allelic effects of causal mutations may differ in different breeds due to epistasis and differences in allele frequencies (Goddard et al, 2018). Therefore, multiple breed evaluation might benefit from a common multi-breed reference population when the models consider the differences in LD and/or marker effects among the involved breeds. The breed origin of alleles BOA model (Karaman et al., 2021) predicts breeding values based on breed-specific marker effects estimated with purebred and crossbred data, allowing for accounting for correlations between the SNP effects of the different breeds. The application of BOA model

can be advantageous in multi-breed populations to the extent that the breeds are distantly related. Otherwise, a joint analysis of all available data can be highly competitive. The Irish beef cattle population is a multi-breed population involving around 40 breeds, with some of them being dairy breeds. Therefore, it is expected that accounting for breed-specific marker effect would be more accurate than assuming homogeneous marker effects across the breeds. The objective of this study was to compare accuracy of predictions from a pedigree-based model (PBLUP) or using genomic information in a joint analysis (SNP-BLUP) or BOA model for the Irish beef cattle population.

3. Material and methods

3.1. Data

The data used in the present study belongs to an Irish cattle multi-breed population (ICBF, http://www.icbf.com). We kept as main breeds the four pure breeds which contributed the most to the crossbred genotyped animals in the population. The pure breeds were Limousin (LM), Charolais (CH), Angus (AA) and Holstein (HO). For BOA estimation, as a fifth group, we kept five breeds that followed the main and treated them as an "others" (OT) group. We selected genotyped crossbred animals for estimating BOA with more than 85% of the four main breeds (166,925 animals). Carcass weight (kg) data were available on 10,159,493 animals born between 2000 and 2020 slaughtered between 300 and 1,277 days of age. Purebred animals from minor breeds and animals without registered sire were removed, resulting in 9,947,707 phenotypic records. After quality control, 46,983 SNPs remained available. The pedigree contained 16,120,121 animals born between 1950 and 2021.

3.2. Breed origin estimation

We used the AllOr method (Eiríksson *et al*. 2021) to infer breed origin to each SNP allele. We included a donor genotype library with purebred animals of main breeds and OT, and we also included a pedigree file obtained by tracing the genotyped crossbreds back. On average, 96.95% of the alleles were assigned a breed of origin. The alleles not assigned to a particular breed were defined as coming from the OT group.

3.3. Prediction of breeding values

The phenotypes for the whole population were pre-corrected for the following non-genetic effects (*fixed*: birth year; type of birth; factory of slaughter; and age at slaughter linear, quadratic and cubic; *random*: contemporary group of herd of slaughter (CG1) and prior to slaughter; and dam permanent environment). Then, we selected a subset (as explained later in the validation subsection) to run PBLUP, SNP-BLUP and BOA models. For PBLUP we traced three generations back the pedigree of animals in the subset.

3.4. Models

PBLUP and SNP-BLUP models

$$\mathbf{y^*}=\mathbf{1}u+\mathbf{Xb}+\mathbf{Zu}+\mathbf{e} \quad (1)$$

where $\mathbf{y}^*$ is a vector of adjusted phenotypes of all animals, $\mathbf{1}$ is a vector of 1s, μ is the overall mean, $\mathbf{X}$ is the matrix of breed proportions computed from BOA assignments, $\mathbf{b}$ is the vector of fixed breed effects, and $\mathbf{e}$ is the vector of random residuals. Random residuals were assumed to follow a normal distribution, $\mathbf{e}\sim N(0,\sigma^2_e)$, where σ^2_e is the residual variance. $\mathbf{Z}$ and $\mathbf{u}$ changes depending on the model. For PBLUP $\mathbf{u}$ is a vector of EBVs, distributed as $\mathbf{u}\sim N(0,\mathbf{A}\sigma^2_g)$, where σ^2_g is the additive variance, and $\mathbf{A}$ is the numerator relationship matrix, $\mathbf{Z}$ relates phenotypes to animals. For the SNP-BLUP model, $\mathbf{u}$ is the vector of SNP effects and $\mathbf{Z}$ is the matrix of centered genotypes based on current allele frequency.

BOA model

$$\mathbf{y^*}=\mathbf{1}u+\mathbf{Xb}+\mathbf{Z}_{CH}\mathbf{u}_{CH}+\mathbf{Z}_{LM}\mathbf{u}_{LM}+\mathbf{Z}_{AA}\mathbf{u}_{AA}+\mathbf{Z}_{HO}\mathbf{u}_{HO}+\mathbf{Z}_{OT}\mathbf{u}_{OT}+\mathbf{e} \quad (2)$$

where $\mathbf{y^*}$,$\mathbf{1}$, μ, $\mathbf{X}$, $\mathbf{b}$, and $\mathbf{e}$ were as described in Equation 1, $\mathbf{Z}_{CH}$, $\mathbf{Z}_{LM}$, $\mathbf{Z}_{AA}$, $\mathbf{Z}_{HO}$, and $\mathbf{Z}_{OT}$ are the matrices of breed-specific content of SNP for CH, LM, AA, HO and OT, $\mathbf{u}_{CH}$, $\mathbf{u}_{LM}$, $\mathbf{u}_{AA}$, $\mathbf{u}_{HO}$ and $\mathbf{u}_{OT}$ are vectors of SNP effects for CH, LM, AA, HO, and OT, respectively. The **Z** matrices were formed by assigning each allele to a certain pure population for which the expected probability is the highest. The entry at a locus in an **Z** matrix, for instance, $\mathbf{Z}_{CH}$, were the number of "*A"* alleles (0, 1 or 2) originating from CH for an animal. Consequently, if an animal had "*aa*" genotype or had no allele originating from CH, the corresponding entry in $\mathbf{Z}_{CH}$ was zero. The **Z** matrices were centered prior to analysis. The analysis assumed correlations between SNP effects of the different breeds as follows: a multivariate normal distribution was assigned for the vectors of SNP effects: $[\mathbf{u'}_{CH}, \mathbf{u'}_{LM}, \mathbf{u'}_{AA}, \mathbf{u'}_{HO}, \mathbf{u'}_{OT}]' \mid \mathbf{B} \sim N(0, \mathbf{B} \otimes \mathbf{I})$, where **I** is an identity matrix, and **B** is as in (Karaman et al., 2021). Briefly, the diagonals are the breed-specific SNP variances and off-diagonals are covariances.

The PBLUP analysis was performed in DMU (Madsen et al., 2014). The SNP-BLUP and BOA analyses were written and carried out in Julia software (Bezanson et al., 2017).

3.5. Validation

The validation animals were selected from the whole population, using paternal half-sib families in which more than half of the animals were born after 2018. For SNP effects estimation we used a subset of animals in CG1 with >10 animals (28,138 animals) and we removed animals in the validation that were included in this subset. We kept 30,234 animals for validation, of which 24,805 were crossbred and 5,618 were purebred. Estimated breeding values were calculated as the summation of the breeding value and breed effects × breed proportions for each animal. The prediction accuracies were calculated as the correlation between adjusted phenotype (y*) and DGV divided by the square root of the heritability. We obtained a bootstrap sample to obtain 10,000 samples that were used to calculate accuracies, standard errors and to make the comparisons between the models with a paired t-test.

4. Results

Figure 1 shows the prediction accuracies obtained with PBLUP, SNP-BLUP and BOA.

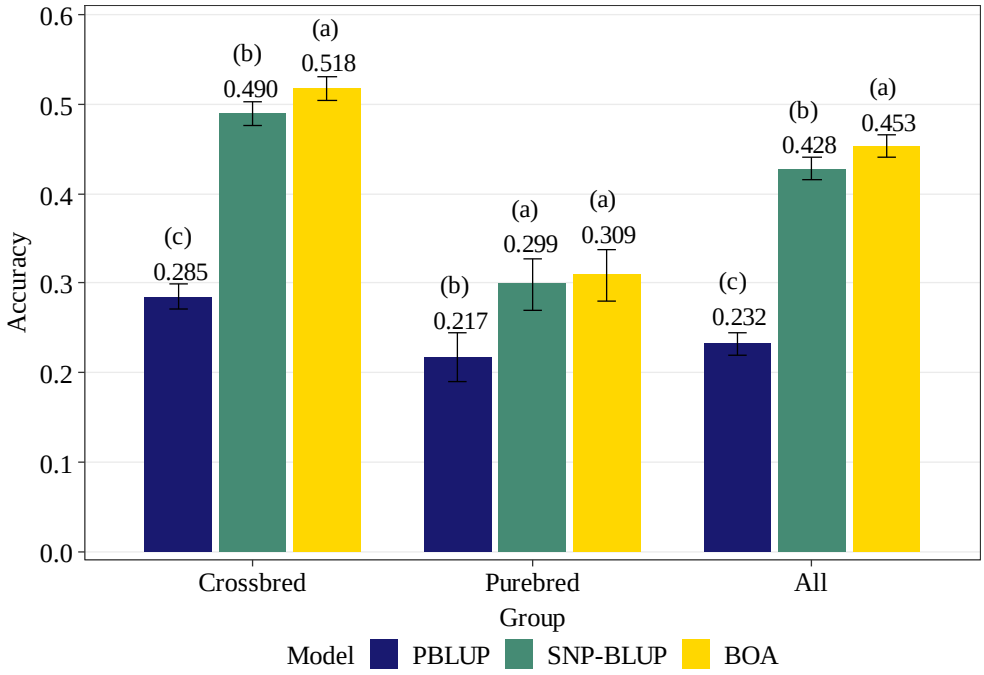

Figure 1 Accuracies and standard errors for validation animals.

Based on the validation procedure, gains in prediction accuracy with SNP-BLUP over PBLUP were in 72% and 38% for crossbred and purebred animals, respectively, and gains in prediction accuracy with BOA over PBLUP were 82% and 42% for crossbred and purebred animals, respectively. Regarding SNP-BLUP and BOA, there was no significant improvement in accuracy of purebred animals, but improvements of 6% for crossbred animals. Additionally, posterior distribution of B from BOA were used to calculate the correlation of SNP effects among the breeds. Correlations ranged from 0.45 (HO - LM) to 0.89 (HO - OT).

5. Discussion

The increases in accuracies when including genomic data were observed for both crossbred and purebred animals. Further improvements were obtained when using BOA. In general, the correlations between SNP effects of the OT group and the four main breeds were higher than between the main breeds (results not given). The estimates of high correlations between the SNP effects of the breeds may occur for different reasons: i) there were five breeds combined in the OT group (beef and dairy breeds) which may lead to problems in assignments of alleles, ii) unassigned alleles were included in the OT group, iii) the imputation of genotypes was carried out jointly in a multi-breed setting for the entire population which might lead to errors in genotypes, and therefore in predictions using BOA. It has been shown that errors in assignments of alleles lead to an increase in the correlation of SNP effects between the breeds (Guillenea et al., 2022). Even with correlations higher than expected, the BOA model outperformed PBLUP but more importantly, also out-performed the SNP-BLUP model for animals of the Irish cattle population. Similar results were reported in a simulation study

comparing BOA with a model that combines the genomic information of the different breeds (Karaman et al., 2021).

The Irish beef cattle evaluation is a very large diverse population in which the vast majority of the animals are crossbred and due to the levels of older phenotypes most are non-genotyped. This implies that a model needs to be developed that can handle both situations. We are currently working on an extension of the BOA model to implement it in populations where some animals are non-genotyped, in a combination of the BOA model with a single-step SNPBLUP model (Liu et al., 2014).

6. References

Bezanson J., Edelman A., Karpinski S., and Shah. V. (2017) SIAM rev. 59(1):65-98. https://doi.org/10.1137/141000671

Guillenea A., Su G., Lund M.S., Karaman E. (2022). J. Dairy Sci. (in press) https://doi.org/10.3168/jds.2021-21173

Eiríksson J.H., Karaman E., Su G., and Christensen O.F. (2021) GSE 53(1), 1-13. https://doi.org/10.1186/s12711-021-00678-3

Goddard M., MacLeod I., Kemper K., Xiang R., Van den Berg I., et al. (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Karaman E., Su G., Croue I., and Lund M.S. (2021) GSE 53(1), 1-13. https://doi.org/10.1186/s12711-021-00637-y

Liu Z., Goddard M., Reinhardt F., and Reents R. (2014) J. Dairy Sci. 97, 5833-50. https://doi.org/10.3168/jds.2014-7924

Madsen P., Jensen J., Labouriau R., Christensen O., Sahana G., DMU. (2014) Proc. of the 12th WCGALP, Vancouver, Canada.

7. Partners involved in the work

AU, Teagasc (ICBF).

8. Annex

These results are presented in a communication by Guillenea et al. (2022) submitted to the World Congress of Genetics Applied to Livestock Production (Rotterdam, NL):

Genomic prediction in multi-breed Irish beef cattle population using a breed origin of alleles model

A. Guillenea[1*], M.S. Lund[1], R. Evans[2], V. Boerner[1] and E. Karaman[1]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark; [2] Irish Cattle Breeding Federation, Bandon, Co. Cork, Ireland; [*]ana.guillenea@qgg.au.dk

Genomic prediction in multi-breed Irish beef cattle population using a breed origin of alleles model

A. Guillenea[1*], M.S. Lund[1], R. Evans[2], V. Boerner[1] and E. Karaman[1]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark; [2] Irish Cattle Breeding Federation, Bandon, Co. Cork, Ireland; [*]ana.guillenea@qgg.au.dk

Abstract

The objective of this study was to evaluate genomic prediction using a breed origin of alleles (BOA) model for a multi-breed Irish beef cattle population. Our BOA model uses one matrix with allele counts for each breed to predict breed-specific marker effects, enabling to use crossbred and purebred animals in the reference population. Pedigree-based best linear unbiased prediction (PBLUP), a genomic model assuming a homogeneous population (SNP-BLUP) and BOA model were compared. Accuracy was estimated as the correlation between breeding values and corrected phenotype divided by the square root of heritability. The accuracy of predictions using BOA increased by 82% and 42% for crossbred and purebred animals, respectively, compared with PBLUP, and improvements of BOA over SNP-BLUP were of 6% for crossbred animals.

Introduction

Genomic prediction (GP) has been rapidly implemented in cattle due to its ability to accurately predict the genomic values of selection candidates early in life. Accuracy of genomic prediction relies on the linkage disequilibrium (LD) between genetic markers, usually single nucleotide polymorphisms (SNPs), and causal mutations. The stronger the LD, the more accurately will be the prediction of the direct genomic breeding values (DGVs). The effects of causal mutations captured by the SNPs is not consistent across distantly related populations due to differences in LD patterns. Additionally, some causal mutations may not segregate in all breeds, or the allelic effects of causal mutations may differ in different breeds due to epistasis and differences in allele frequencies (Goddard et al, 2018). Therefore, multiple breed evaluation might benefit from a common multi-breed reference population when the models consider the differences in LD and/or marker effects among the involved breeds. The breed origin of alleles BOA model (Karaman et al., 2021) predicts breeding values based on breed-specific marker effects estimated with purebred and crossbred data, allowing for accounting for correlations between the SNP effects of the different breeds. The application of BOA model can be advantageous in multi-breed populations to the extent that the breeds are distantly related. Otherwise, a joint analysis of all available data can be highly competitive. The Irish beef cattle population is a multi-breed population involving around 40 breeds, with some of them being dairy breeds. Therefore, it is expected that accounting for breed-specific marker effect would be more accurate than assuming homogeneous marker effects across the breeds. The objective of this study was to compare accuracy of predictions from a pedigree-based model (PBLUP) or using genomic information in a joint analysis (SNP-BLUP) or BOA model for the Irish beef cattle population.

Materials & Methods

Data.

The data used in the present study belongs to an Irish cattle multi-breed population (ICBF, http://www.icbf.com). We kept as main breeds the four pure breeds which contributed the most to the crossbred genotyped animals in the population. The pure breeds were Limousin (LM), Charolais (CH), Angus (AA) and Holstein (HO). For BOA estimation, as a fifth group, we kept five breeds that followed the main and treated them as an "others" (OT) group. We selected genotyped crossbred animals for estimating BOA with more than 85% of the four main breeds (166,925 animals). Carcass weight (kg) data were available on 10,159,493 animals born between 2000 and 2020 slaughtered between 300 and 1,277 days of age. Purebred animals from minor breeds and animals without registered sire were removed, resulting in 9,947,707 phenotypic records. After quality control, 46,983 SNPs remained available. The pedigree contained 16,120,121 animals born between 1950 and 2021.

Breed origin estimation.

We used the AllOr method (Eiríksson *et al*. 2021) to infer breed origin to each SNP allele. We included a donor genotype library with purebred animals of main breeds and OT, and we also included a pedigree file obtained by tracing the genotyped crossbreds back. On average, 96.95% of the alleles were assigned a breed of origin. The alleles not assigned to a particular breed were defined as coming from the OT group.

Prediction of breeding values.

The phenotypes for the whole population were pre-corrected for the following non-genetic effects (*fixed*: birth year; type of birth; factory of slaughter; and age at slaughter linear, quadratic and cubic; *random*: contemporary group of herd of slaughter (CG1) and prior to slaughter; and dam permanent environment). Then, we selected a subset (as explained later in the validation subsection) to run PBLUP, SNP-BLUP and BOA models. For PBLUP we traced three generations back the pedigree of animals in the subset.

Models.

PBLUP and SNP-BLUP models

$$\mathbf{y^*}=\mathbf{1}u+\mathbf{Xb}+\mathbf{Zu}+\mathbf{e} \quad (1)$$

where **y*** is a vector of adjusted phenotypes of all animals, **1** is a vector of 1s, μ is the overall mean, **X** is the matrix of breed proportions computed from BOA assignments, **b** is the vector of fixed breed effects, and **e** is the vector of random residuals. Random residuals were assumed to follow a normal distribution, $\mathbf{e}\sim N(0,\sigma^2_e)$, where σ^2_e is the residual variance. **Z** and **u** changes depending on the model. For PBLUP **u** is a vector of EBVs, distributed as $\mathbf{u}\sim N(0,\mathbf{A}\sigma^2_g)$, where σ^2_g is the additive variance, and **A** is the numerator relationship matrix, **Z** relates phenotypes to animals. For the SNP-BLUP model, **u** is the vector of SNP effects and **Z** is the matrix of centered genotypes based on current allele frequency.

BOA model

$$\mathbf{y^*}=\mathbf{1}u+\mathbf{Xb}+\mathbf{Z}_{CH}\mathbf{u}_{CH}+\mathbf{Z}_{LM}\mathbf{u}_{LM}+\mathbf{Z}_{AA}\mathbf{u}_{AA}+\mathbf{Z}_{HO}\mathbf{u}_{HO}+\mathbf{Z}_{OT}\mathbf{u}_{OT}+\mathbf{e} \quad (2)$$

where **y*,1,** μ, **X**, **b**, and **e** were as described in Equation 1, $\mathbf{Z}_{CH}$, $\mathbf{Z}_{LM}$, $\mathbf{Z}_{AA}$, $\mathbf{Z}_{HO}$, and $\mathbf{Z}_{OT}$ are the matrices of breed-specific content of SNP for CH, LM, AA, HO and OT, $\mathbf{u}_{CH}$, $\mathbf{u}_{LM}$, $\mathbf{u}_{AA}$, $\mathbf{u}_{HO}$ and $\mathbf{u}_{OT}$ are vectors of SNP effects for CH, LM, AA, HO, and OT, respectively. The **Z** matrices were formed by assigning each allele to a certain pure population for which the expected probability is the highest. The entry at a locus in an **Z** matrix, for instance, $\mathbf{Z}_{CH}$, were the number of "*A*" alleles (0, 1 or 2) originating from CH for an animal. Consequently, if an animal had "*aa*" genotype or had no allele originating from CH, the corresponding entry in $\mathbf{Z}_{CH}$ was zero. The **Z** matrices were centered prior to analysis. The analysis assumed correlations between SNP effects of the different breeds as follows:

a multivariate normal distribution was assigned for the vectors of SNP effects: [$\mathbf{u}'_{CH}$, $\mathbf{u}'_{LM}$, $\mathbf{u}'_{AA}$, $\mathbf{u}'_{HO}$, $\mathbf{u}'_{OT}$]' | $\mathbf{B}$~N(0,$\mathbf{B}\otimes\mathbf{I}$), where $\mathbf{I}$ is an identity matrix, and $\mathbf{B}$ is as in (Karaman et al., 2021). Briefly, the diagonals are the breed-specific SNP variances and off-diagonals are covariances.

The PBLUP analysis was performed in DMU (Madsen et al., 2014). The SNP-BLUP and BOA analyses were written and carried out in Julia software (Bezanson et al., 2017).

Validation.

The validation animals were selected from the whole population, using paternal half-sib families in which more than half of the animals were born after 2018. For SNP effects estimation we used a subset of animals in CG1 with >10 animals (28,138 animals) and we removed animals in the validation that were included in this subset. We kept 30,234 animals for validation, of which 24,805 were crossbred and 5,618 were purebred. Estimated breeding values were calculated as the summation of the breeding value and breed effects × breed proportions for each animal. The prediction accuracies were calculated as the correlation between adjusted phenotype (y*) and DGV divided by the square root of the heritability. We obtained a bootstrap sample to obtain 10,000 samples that were used to calculate accuracies, standard errors and to make the comparisons between the models with a paired t-test.

Results

Figure 1 shows the prediction accuracies obtained with PBLUP, SNP-BLUP and BOA.

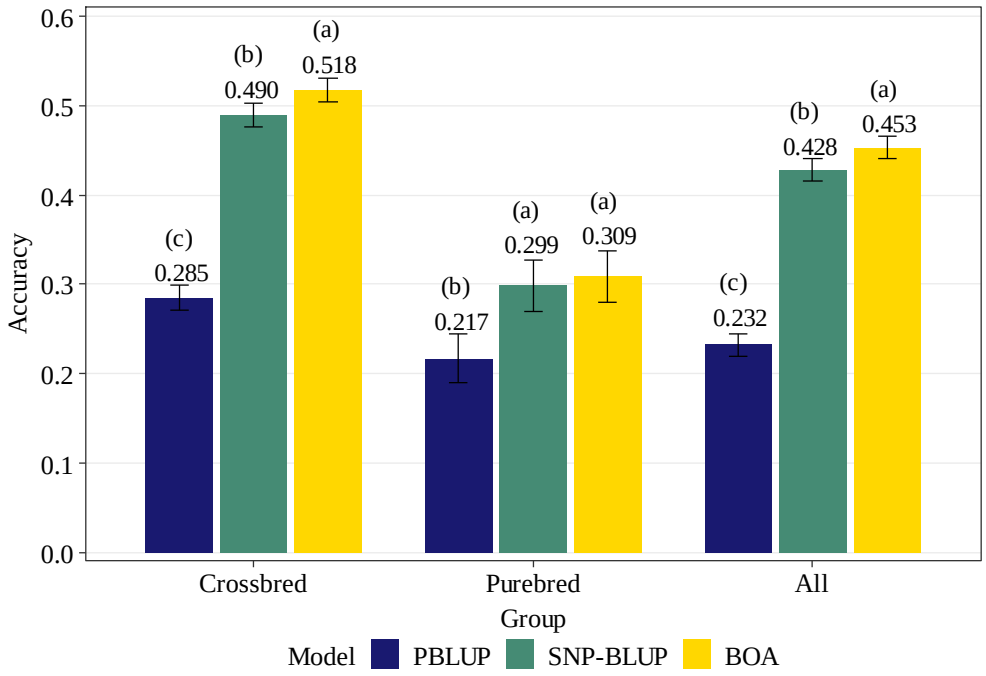

Figure 1. Accuracies and standard errors for validation animals.

Based on the validation procedure, gains in prediction accuracy with SNP-BLUP over PBLUP were in 72% and 38% for crossbred and purebred animals, respectively, and gains in prediction accuracy with BOA over PBLUP were 82% and 42% for crossbred and purebred animals, respectively. Regarding SNP-BLUP and BOA, there was no significant improvement in accuracy of purebred animals, but improvements of 6% for crossbred animals. Additionally, posterior distribution of **B** from BOA were used to calculate the correlation of SNP effects among the breeds.

Correlations ranged from 0.45 (HO - LM) to 0.89 (HO - OT).

Discussion

The increases in accuracies when including genomic data were observed for both crossbred and purebred animals. Further improvements were obtained when using BOA. In general, the correlations between SNP effects of the OT group and the four main breeds were higher than between the main breeds (results not given). The estimates of high correlations between the SNP effects of the breeds may occur for different reasons: i) there were five breeds combined in the OT group (beef and dairy breeds) which may lead to problems in assignments of alleles, ii) unassigned alleles were included in the OT group, iii) the imputation of genotypes was carried out jointly in a multi-breed setting for the entire population which might lead to errors in genotypes, and therefore in predictions using BOA. It has been shown that errors in assignments of alleles lead to an increase in the correlation of SNP effects between the breeds (Guillenea et al., 2022). Even with correlations higher than expected, the BOA model outperformed PBLUP but more importantly, also out-performed the SNP-BLUP model for animals of the Irish cattle population. Similar results were reported in a simulation study comparing BOA with a model that combines the genomic information of the different breeds (Karaman et al., 2021). The Irish beef cattle evaluation is a very large diverse population in which the vast majority of the animals are crossbred and due to the levels of older phenotypes most are non-genotyped. This implies that a model needs to be developed that can handle both situations. We are currently working on an extension of the BOA model to implement it in populations where some animals are non-genotyped, in a combination of the BOA model with a single-step SNPBLUP model (Liu et al., 2014).

Acknowledgments

We thank GentTORE and ANII Uruguay (POS_EXT_2018_1_154296) for financial support.

References

Bezanson J., Edelman A., Karpinski S., and Shah. V. (2017) SIAM rev. 59(1):65-98. https://doi.org/10.1137/141000671

Guillenea A., Su G., Lund M.S., Karaman E. (2022). J. Dairy Sci. (in press) https://doi.org/10.3168/jds.2021-21173

Eiríksson J.H., Karaman E., Su G., and Christensen O.F. (2021) GSE 53(1), 1-13. https://doi.org/10.1186/s12711-021-00678-3

Goddard M., MacLeod I., Kemper K., Xiang R., Van den Berg I., *et al*. (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Karaman E., Su G., Croue I., and Lund M.S. (2021) GSE 53(1), 1-13. https://doi.org/10.1186/s12711-021-00637-y

Liu Z., Goddard M., Reinhardt F., and Reents R. (2014) J. Dairy Sci. 97, 5833-50. https://doi.org/10.3168/jds.2014-7924

Madsen P., Jensen J., Labouriau R., Christensen O., Sahana G., DMU. (2014) Proc. of the 12th WCGALP, Vancouver, Canada.