Genomic prediction using breed origin of alleles model accounting for probabilities in the assignment of the alleles

A. Guillenea*, S. Guosheng, M.S. Lund and E. Karaman

Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark;
*ana.guillenea@qgg.au.dk

Abstract

This study investigates genomic prediction using a breed origin of alleles (BOA) model which accounts for BOA with a (i) definitive assignment to a breed, or (ii) probabilities of assignments to each breed. Our BOA model estimates breed-specific marker effects based on genotypic and phenotypic information from the purebred and crossbred animals. A traditional combined analysis of all breeds' data implicitly assumes a correlation of one between the marker effects of the breeds, whereas our BOA model allows the estimation of these correlations. We used de-regresed proof of production traits from the admixed Nordic Red Cattle population to evaluate the model and performed the analysis assuming marker effects between breeds are correlated or assuming they are uncorrelated. We found that using probabilities in the BOA model outperformed the BOA model assuming the origin of each marker is known with certainty, especially when the marker effects were assumed uncorrelated between breeds.

Keywords: admixed animals, breed-specific allele effects, breed origin assignment

Introduction

The implementation of genomic prediction (GP) is challenged in admixed populations since the same allele of a given marker might have a different effect on the phenotype according to its breed of origin. Accuracy of GP relies on the linkage disequilibrium (LD) between single nucleotide polymorphisms (SNP) and quantitative trait loci (QTL). In pooled multi-breed reference populations, accuracy of GP relies on the LD between SNP and QTL across breeds, but LD may be different, or even the phase be reversed across breeds. QTL effects might also vary between breeds if there are interactions between QTL and genetic background (Calus et al., 2018). Several methods have been proposed recently to account for breed origin of alleles (BOA) in GP, and several methodologies/softwares have been developed to infer BOA (Lawson et al., 2012; Vandelplas et al., 2016; Eiriksson et al. 2021). For GP, Karaman et al. (2021) developed a method that allows to include purebred and crossbred animals, estimating SNP effects considering correlation of SNP effects between the breeds. This method includes breed-specific matrices and thus, relies on the correct assignment of the alleles. Some of the existing softwares to infer BOA provide the probabilities of belonging to each breed for each allele, and others provide the probabilities for the unassigned alleles. Including these probabilities into the models, instead of assignment of an allele to one particular breed, could improve reliabilities. Because they may account for some uncertainty in the assignments, especially for highly admixture populations. The objective of this study was to compare GP for a BOA model using either assignment of an allele to a specific breed or probabilities of an allele from the possible breeds, based on data from the RDC population, which is an admixture of Danish Red (RDM), Swedish Red (SRB) and Finnish Ayrshire (FAY) dairy

cattle breeds, which have strong ties due to the use of common bulls.

Materials & Methods

Genotypic and phenotypic data.

We used de-regress proofs (DRP) of 51,336 animals of the RDC population for milk, fat and protein as response variables to predict direct genomic breeding values (DGV). The Nordic Cattle Genetic Evaluation provided the data. The bulls in the data set were genotyped using a conventional 50K chip. Some cows were genotyped with 50K and some with a customized LD, including the standard LD SNPs and additional SNPs. HOL was included as an ancestral population as a preliminary pedigree analysis showing ties of HOL with RDM. The genotypes of all animals were imputed to 50K and phased using FImpute v3.0 software, and 42,470 SNPs common to RDC and HOL were used.

Estimation of breed origin.

First, we used all genotyped RDC bulls to perform an ancestry analysis to identify pure animals, regardless of the presence of DRP. In total, 8,356 bulls were included in this analysis (5,356 RDC and 3,000 HOL). The bulls were allocated to purebred and admixed groups based on the predictions of breed proportions performed in ADMIXTURE v1.30 software $k = 4$ (k = ancestral populations). Second, breed origin of alleles was inferred using ChromoPainterV2 software (Lawson et al., 2012), separately for each chromosome, using haplotypes of purebred animals selected based on ADMIXTURE analysis. The rest of the RDC animals were defined as admixed, and their BOA were estimated. The expected probability of each SNP coming from each of the purebred populations was used to infer the BOA in two ways: (1) the allele was assigned to a certain pure population for which the expected probability is the highest, or (2) the allele was assigned to all populations as a product of the allele count and the expected probabilities. Breed proportions were inferred by summing the alleles from each breed.

Statistical model.

The BOA model to estimate breed-specific SNP effects is as follows (Karaman et al., 2021):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{Z}_D\mathbf{u}_D + \mathbf{Z}_F\mathbf{u}_F + \mathbf{Z}_S\mathbf{u}_S + \mathbf{Z}_H\mathbf{u}_H + e$$

where $\mathbf{y}$ is the vector of DRP of milk, fat or protein for the reference animals, $\mathbf{1}$ is the vector of 1s, μ is the overall mean, $\mathbf{X}$ is the matrix of breed proportions, $\mathbf{b}$ is the vector of fixed breed effects, and $\mathbf{e}$ is the vector of random residuals. $\mathbf{Z}_D$, $\mathbf{Z}_F$, $\mathbf{Z}_S$, and $\mathbf{Z}_H$ are the matrices of breed-specific content of SNP for RDM, FAY, SRB, and HOL, $\mathbf{u}_D$, $\mathbf{u}_F$, $\mathbf{u}_S$ and $\mathbf{u}_H$ are vectors of SNP effects for RDM, FAY, SRB, and HOL, respectively.

Two approaches to construct $\mathbf{Z}$ matrices were used. For the first approach, the $\mathbf{Z}$ matrices were formed by assigning each allele to a certain pure population for which the expected probability is the highest (**BOAg**). The entry at a locus in an $\mathbf{Z}$ matrix, for instance, $\mathbf{Z}_D$, were the number of "*A*" alleles (0, 1 or 2) originating from RDM for an animal. Consequently, if an animal had "*aa*" genotype or had no allele originating from RDM, the corresponding entry in $\mathbf{Z}_D$ was zero. The same applies to other $\mathbf{Z}$ matrices for each breed. For the alternative approach, each allele was assigned to all original populations according to expected probabilities (**BOAp**). Thus, the $\mathbf{Z}$ matrices were achieved by taking the sum of the product of the allele count (0 or 1) and the expected probabilities over the two alleles at each locus. The $\mathbf{Z}$ matrices were centered by subtracting the mean of the column prior to analysis.

For the analysis assuming breed-specific SNP effects uncorrelated across the breeds

(Uncor), each vector of SNP was assigned a prior of a normal distribution separately for each breed (i = D,F,S,H): $\mathbf{u}_i|\sigma^2_i \sim N(0, \sigma^2_i)$. Here, σ^2_i is the SNP variance for breed i, obtained from a preliminary analysis using only bulls' data. For the analysis assuming correlated breed-specific SNP effects (Cor), a multivariate normal distribution was assigned for the vectors of SNP effects: $[\mathbf{u}'_D, \mathbf{u}'_F, \mathbf{u}'_S, \mathbf{u}'_H]' \mid \mathbf{B} \sim N(0, \mathbf{B} \otimes I)$, where **I** is an identity matrix, and **B** is as in (Guillenea et al., 2022). Briefly, the diagonals are the breed-specific SNP variances and off-diagonals of **B** are covariances. Random residuals were assumed to follow a normal distribution, $e \sim N(0, \mathbf{D}\sigma^2_e)$, where σ^2_e is the residual variance, and **D** is a diagonal matrix with elements $d_{ii} = 1/w$, w is the weighting factor for the ith DRP. More detail of priors are given in Guillenea et al. (2022). Fixed effects were assigned to having flat priors. The analyses were written and carried out in a software with Julia programming language (Bezanson et al., 2017).

Model validation.

The validation consisted of 11,786 cows, while reference consisted of 5,104 bulls and 34,446 cows. Both populations were made based on groups of half-sibs with a cut-off birth date July 1, 2015. The DGVs of validation animals were calculated by multiplying the breed-specific matrices with breed-specific SNP effects and adding a fixed breed contribution. The prediction reliability was calculated as the squared correlation between DGV and DRP divided by the reliability of DRP for the validation animals. We used a bootstrap procedure sampling 10,000 times to calculate reliabilities and the standard errors of each model.

Results

Reliabilities of GP are presented in Figure 1.

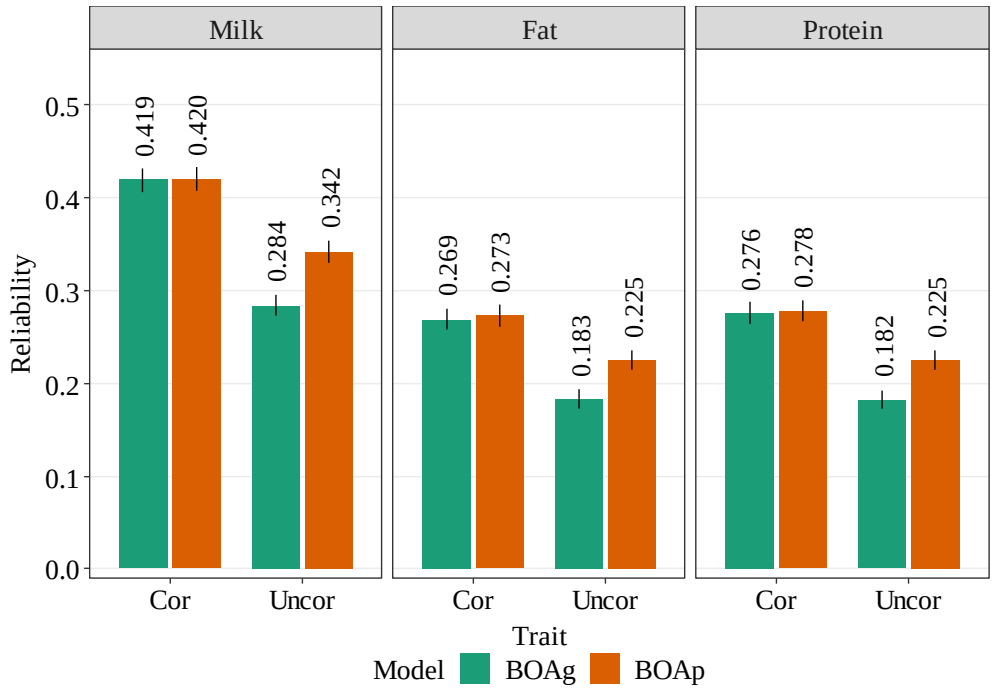

Figure 1. Reliabilities and standard errors of BOA with definitive assignment to a breed (BOAg) (Guillenea et al., 2022) or probabilities of assignments to possible breeds (BOAp) for scenario of correlated (Cor) or scenario of uncorrelated (Uncor) SNP effects between breeds.

Using probabilities of assignments improves prediction reliabilities for the three traits in Uncor, but not in Cor. On average, the increase of BOAp over BOAg was 1% for the correlated analysis and 33% for the uncorrelated analysis. Regarding the production traits under study, milk presented the highest reliabilities and the pattern over the three traits were similar.

Discussion

The performance of models based on the breed origin of the alleles depends, among other factors, on the precise estimation of the origin of each allele, which is used to construct the breed-specific allele matrices. The RDC is a highly admixed breed, which made it difficult to define purebred animals of the three Nordic original breeds, and consequently, the assignments of the alleles in ChromoPainterV2 showed a distribution of probabilities, in which only 40% were assigned to their breed origin with probability >0.9. This means that in our first approach (BOAg), if one allele with a probability of, for instance, 30% of coming from RDM and the probabilities for the other breeds three were lower, the allele was assigned to RDM. In the second approach (BOAp), the same allele was assigned to all breeds with the expected probabilities. The uncorrelated analysis obtains more gains for using the probabilities because it depends more on intra-breed information. In this population, in which breeds were found to be highly related, the impact of accounting for probabilities was small for the correlated analysis. However, it is expected that the improvement of predictions by using probabilities would be more in a correlated analysis for the breeds which more distantly related. For such breeds the use of BOA models would also become more important (Karaman et al., 2021).

Conclusions

This study compared GP of BOA with two types of breed-specific genotypic matrices: (i) each allele was assigned to only one specific breed, and (ii) probabilities in the assignements were taken into account in the matrices. We conclude that the BOA model benefits from the use of probabilities in the construction of the breed-specific matrices, so it is a more precise approach to assign alleles that can be implemented in any models with breed-specific matrices.

References

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

Calus M., Goddard M., Wientjes Y., Bowman P., and Hayes. B. (2018) J Dairy Sci 101(5):4279-4294. https://doi.org/10.3168/jds.2017-13366

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

Guillenea A., Su G., Lund M.S., Karaman E. (2022). J. Dairy Sci. (in press)

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

Lawson D.J., Hellenthal G., Myers S., and Falush D. (2012). PLoS Genet 8(1):e1002453. https://doi.org/10.1371/journal.pgen.1002453

Vandenplas J., Calus M.P., Sevillano C.A., Windig J.J. and Bastiaansen J.W. (2016). GSE 48(1):61. https://doi.org/10.1186/s12711-017-0350-1