Including local genomic breed proportions in genomic predictions for crossbred

J.H. Eiríksson*, G. Su and O.F. Christensen

Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark; *jonh@qgg.au.dk

Abstract

Genomic evaluations for crossbred animals introduce challenges not encountered for purebred animals. We propose to include effects of local breed proportions (LBP) based on detected breed of origin of alleles in a genomic model for crossbred dairy cows. We tested different segment length for modelling the LBP: 1, 20, 100 and 500 markers, for genomic prediction based on data of first lactation protein yields of 5,176 Danish crossbred dairy cows with contribution from three dairy breeds. The models that included LBP had slightly higher cross-validation predictive ability, 0.390 for all segment lengths compared with 0.385 for a model without LBP effects. The LBP models also had less dispersion bias. Inclusion of LBP in genomic predictions can therefore improve predictions even when considering the local effects in segments of hundreds of markers.

Introduction

Genomic evaluations for crossbred introduce challenges not encountered in genomic evaluation within the pure breeds. Among those challenges is the lower level of linkage disequilibrium between markers and quantitative trait loci across breeds than within breed levels (Ibánẽz-Escriche *et al.*, 2009), genetic background (Falconer, 1989), and breed segregation, i.e. the additional genetic variation in offspring of crossbred parents (Christensen *et al.*, 2015).

Models that split the breeding values of crossbreds into breed specific terms and account for breed origin of alleles (BOA) can in principle account for the different linkage between breeds (Ibánẽz-Escriche *et al.*, 2009). However, these models do not always result in more accurate predictions than models assuming common marker effects (Sevillano *et al.,* 2017). Applications of BOA models (e.g. Sevillano *et al.* 2017) have generally not accounted for the breed segregation term which Christensen *et al.*, (2015) presented in their derivation of the models for three-way crosses.

For BOA models with training on purebred data, intercepts for correction of breed levels are necessary for the prediction of breeding values of crossbreds (Eiríksson *et al.*, 2021) and might not be readily available from the genomic evaluations of purebred. Phenotypes of crossbreds along with BOA information could give more precise estimates of the breed level intercepts than purebred information alone. Additionally, BOA could give information on the local breed proportions (LBP), i.e. the contribution of the breeds to different regions of the genome, which can vary in offspring of crossbred animals and similarly to the breed segregation term. The LBP can be considered at individual marker level, or for a number of consecutive markers because of the high correlation between BOA of proximate markers.

Therefore, the objectives of this study were to compare predictive ability of genomic models including global genomic breed proportions (GBP) with a model that additionally includes local breed proportions (LBP) for genomic prediction for crossbred cows. We considered different segment lengths for the LBP model (LBPM).

Materials & Methods

Data. Single nucleotide polymorphism (SNP) genotypes from the EUROG MD custom chip and phenotypes of 305-day first lactation protein yields of 5,176 Danish crossbred dairy cows from 74 herds were used in this study. These cows were crosses of Holstein (H), Jersey (J) and Nordic Red Dairy Cattle (R) in various combinations. We imputed the genotypes to a set of 50,684 markers, to fill in missing genotypes, and phased them to two haplotypes using FImpute (Sargolzaei *et al*., 2014), for the genotyped crossbred and 2,500 genotyped animals from each of the pure breeds. We detected BOA using the AllOr method (Eiríksson *et al*., 2021) with a purebred reference genotypes of 2,500 cows from each of the breeds, a segment length of 100 SNPs and shifting five SNPs between rounds in the method.
For investigating crossbreed-specific genetic effects, we constructed a corrected phenotype, i.e. corrected for both fixed effects and genomic estimated breeding values computed from purebred information. We calculated the corrected phenotype as:
$\mathbf{y}^{*}=\mathbf{y}-\hat{\mathbf{a}}_{\mathbf{BOM}}$,
where **y** is a vector of phenotypes of first lactation 305-day protein yields that were corrected for herd-year and calving age effects, and $\hat{\mathbf{a}}_{\mathbf{BOM}}$ is a vector of genomic estimated breeding values. The $\hat{\mathbf{a}}_{\mathbf{BOM}}$ were calculated with a breed of origin model using solutions from each of the three purebred genomic evaluations, as described in Eiríksson *et al.* (2021), but without the intercepts for breed level corrections. The estimated effects from purebred evaluations, i.e. estimated marker effects and residual polygenic effects, were from the February 2021 run of the genomic evaluations of the three pure breeds from the Nordic Cattle Genetic Evaluation.

Models. We compared models with and without including local breed proportion effects.
The GBP model (GBPM) included regressions on global breed proportions and genomic breed heterozygosity rates:
$$\mathbf{y}^{*}=\mathbf{1}u+\mathbf{f_H}c_H+\mathbf{f_J}c_J+\mathbf{k_{H,J}}h_{H,J}+\mathbf{k_{H,R}}h_{H,R}+\mathbf{k_{J,R}}h_{J,R}+\mathbf{e}, \quad (1)$$
where u is an intercept, **1** is a vector of ones, $\mathbf{f_b}$ contains global proportions of alleles assigned to breed b (b=H, J ,R), c_b is a regression coefficient on breed proportions, $\mathbf{k}_{b,b'}$ contains genomic breed heterozygosity rate, i.e. proportion of loci with alleles assigned to two different breeds, b and b', $h_{b,b'}$ is the regression coefficients on genomic breed heterozygosity, and **e** is the random residual $\mathbf{e}\sim N(0,\mathbf{I}\sigma^2_e)$. The values in $\mathbf{f_b}$ were calculated as $f_{i,b}=\Sigma_j(x_{b,i,1,j}+x_{b,i,2,j})/(2m)$, where $x_{b,i,g,j}$ is 1 for marker j with allele on haplotype g assigned to breed b, 0 if the allele is assigned to another breed, and a value between 0 and 1 if the allele was not assigned to definite breed and m is the number of markers. Note that since $\mathbf{f_H}+\mathbf{f_J}+\mathbf{f_R}=\mathbf{1}$, one of the global breed proportion effects can be left out of the model, given that an intercept is included. The values in $\mathbf{k_{H,J}}$ were calculated as $k_{i,b,b'}=\Sigma_j(x_{b,i,1,j}\circ x_{b',i,2,j}+x_{b',i,1,j}\circ x_{b,i,2,j})/m$ where $\circ$ denotes element wise multiplication.
The LBPM additionally included random LBP effects:
$$\mathbf{y}^{*}=u+\mathbf{f_H}c_H+\mathbf{f_J}c_J+\mathbf{k_{H,J}}h_{H,J}+\mathbf{k_{H,R}}h_{H,R}+\mathbf{k_{J,R}}h_{J,R}+\mathbf{Z}^{*}_H\mathbf{p}_H+\mathbf{Z}^{*}_J\mathbf{p}_J+\mathbf{Z}^{*}_R\mathbf{p}_R+\mathbf{e}, \quad (2)$$
where $\mathbf{Z}^{*}_b=\mathbf{Z}_b-\mathbf{f}_b\mathbf{1}'$, with $\mathbf{Z}_b$ containing the local breed proportions, i.e. the proportion of alleles on a segment that were assigned to breed b. Furthermore, $\mathbf{p}_b$ is a vector with random effects of local proportion of breed b, with $\mathbf{p}_b\sim N(0,\mathbf{I}\sigma^2_{p,b})$, where $\sigma^2_{p,b}$ is the variance for segment LBP effect for breed b (similar to SNP variance for SNP-BLUP).
An equivalent model of the LBPM is to construct relationship matrices for the local effects, which is computationally more efficient for short segments. The LBPM with relationship matrix is then:
$$\mathbf{y}^{*}=u+\mathbf{f_H}c_H+\mathbf{f_J}c_J+\mathbf{k_{H,J}}h_{H,J}+\mathbf{k_{H,R}}h_{H,R}+\mathbf{k_{J,R}}h_{J,R}+\mathbf{v}_H+\mathbf{v}_J+\mathbf{v}_R+\mathbf{e}, \quad (3)$$
where $\mathbf{v}_b$ is a vector of local breed proportion animal effects with, $\mathbf{v}_b\sim N(0, \mathbf{Q}_b\sigma^2_{v,b})$ where $\mathbf{Q}_b=\mathbf{Z}^{*}_b\mathbf{Z}^{*}_b{}'/(\mathrm{tr}(\mathbf{Z}^{*}_b\mathbf{Z}^{*}_b{}')/n)$ where tr stands for trace and n is number of animals.

Implementation. We estimated the variance components for LBPM with segment length of 500, 100, 20, and 1 SNPs. For the 1 SNP estimation we constructed a relationship matrix and added a low value, 0.0001, to the diagonal of $\mathbf{Q}_b$ to make it invertible. We used the AI-REML procedure from the DMU package (Madsen and Jensen, 2013) for variance component estimation.

We tested the predictive ability of the LBPMs and GBPM using five-fold cross validation, where we randomly split the herds into five groups, and made the prediction for animals in each group based on records from the other four groups. For all models, the predictions $\hat{\mathbf{y}}$ included all the effects in the model plus $\hat{\mathbf{a}}_{\mathbf{BOM}}$. We compared the predictions with $\mathbf{y}$, i.e. the phenotypes corrected for fixed herd-year and calving age effects.

Results

The genomic breed proportions according to the BOA assignment were 49% H, 15% J, and 36% RDC. According to the results from GBPM, a 10% increase in genomic breed heterozygosity increased protein yield by 1.6 kg (SE=1.0), 1.7 kg (SE=0.4), and 2.5 kg (SE=1.1), for H/J, H/R and J/R respectively. For the LBPMs, the local breed proportion effects together explained around 2% of the variance in $\mathbf{y}^*$, with the highest proportion for the shortest segments (Table 1). The variance related to H proportions was the largest, ranged from 20.8 to 25.8, compared with the variances for J and R proportions, which ranged from 5.0 to 5.7 and 3.1 to 4.8, respectively (Table 1). The predictive ability of the models from the cross-validation was equal for all LBPM, and 0.005 higher than the correlation from GBPM (Table 1). The LBPM had less dispersion bias than GBPM, irrespective of segment length.

Table 1. Variance components for breed proportions as well as prediction accuracy and dispersion bias for first lactation protein yield.

	Segm.	Variance				Cross-validation	
Model[3]	length	$\sigma^2_{v,H}$	$\sigma^2_{v,J}$	$\sigma^2_{v,R}$	σ^2_e	Cor($\hat{\mathbf{y}}$,$\mathbf{y}$)	b_1[2]
GBPM						0.385	0.880
LBPM	500	20.8[1]	5.0[1]	3.1[1]	1578.7	0.390	0.901
LBPM	100	24.8[1]	5.6[1]	4.8[1]	1572.8	0.390	0.901
LBPM	20	25.7[1]	5.7[1]	4.7[1]	1571.9	0.390	0.901
LBPM	1	25.8	5.7	4.8	1571.6	0.390	0.901

[1] Calculated as $\sigma^2_{v,b} = \sigma^2_{p,b} * \text{tr}(\mathbf{Z}^*_b{}' \mathbf{Z}^*_b)/m$

[2] Dispersion bias, estimated as the slope of the regression of $\mathbf{y}$ on $\hat{\mathbf{y}}$

[3] GBPM = Global breed proportion model, LBPM = Local breed proportion model

Discussion

The $\sigma^2_{v,H}$, $\sigma^2_{v,J}$, and $\sigma^2_{v,R}$ estimates are dependent on the breed composition in the datasets. Local breed proportions are for example constant throughout the genome of F1 animals, which were around 30% of our data, and they do therefore not contribute to this variance. Alternative standardization of the LPM relationship matrices, e.g. based on setting the expected diagonal values for a specific type of cross to be 1, could give estimated variances that do not suffer from this. However, in our study that would make comparison across segment lengths skewed. The estimated variances suggest that there is variation related to LBP in the data and the magnitude is about 1-2% of the phenotypic variance and around 5% of the additive genetic variance.

The improvement in predictive ability from LBPM compared to GBPM was marginal in our study (Table 1). However, improvement in prediction was not expected for all the animals in the data, LBP could be more important for groups with few or no F1 animals. Additionally, the training set for estimating LBP effects was not large, only around 4,000 animals for estimating

LBP effects for the three breeds. The relatively small difference in variance explained and predictive ability between alternative segment lengths suggests that local breed proportion effects can be included for longer segments, e.g. 100 SNPs, and therefore reduce the number of parameters without compromising accuracy.

The $\mathbf{Q}_b$ matrices in equation 3 have similar structure as the genomic breed-segregation partial relationship matrix presented by Christensen *et al*. (2015). Both have off-diagonal indicating if animals share origin in fewer or more loci than that is expected based on breed proportions. Christensen *et al*. (2015) only considered three-way crosses with the same terminal sire breed and the diagonal of the breed-segregation partial relationship matrix is therefore constant, equal to 0.5. In our case, with crossbred animals of varying breed compositions, the diagonal elements were highly variable, but with the average equal to one because we standardized the matrix with the trace.

Detected BOA has to our knowledge not before been applied to model heterozygosity in crossbreds. If BOA information is available, modelling heterozygosity in this way is simple, and can be simpler than having to rely on pedigree information.

The LBPM presented here is an add-on to the BOA model with training in purebred population (Eiríksson *et al.* 2021). A potential complication with the application of the BOA model in that paper is that the intercepts for correction for breed level differences may not be available from routine genomic evaluation. The global breed proportion effects could be used for accounting for breed levels in that model. Other authors have presented BOA models with training in both purebreds and crossbreds (Karaman *et al*., 2021) without accounting for breed segregation. Whether the inclusion of local breed proportion effects in those models improves prediction is an interesting area for further development.

Inclusion of LBP effects can moderately increase the predictive ability and reduce bias even when the local effects are considered in segments of hundreds of markers. They should therefore be considered in future development of models for genomic prediction for crossbred.

References

Christensen O.F., Legarra A., Lund M.S., and Su G. (2015) Genet. Sel. Evol. 47:98. https://doi.org/10.1186/s12711-015-0177-6

Eiríksson J.H., Karaman E., Su G. and Christensen O.F. (2021) Genet. Sel. Evol. 53:84. https://doi.org/10.1186/s12711-021-00678-3

Falconer D.S. (1989) Introduction to quantitative genetics. Longman Scientific & Technical, Harlow, United Kingdom.

Ibánẽz-Escriche N., Fernando R.L., Toosi A., and Dekkers J.C. (2009) Genet. Sel. Evol. 41:12. https://doi.org/10.1186/1297-9686-41-12

Karaman E., Su G., Croue I., and Lund M.S. (2021) Genet. Sel. Evol. 53:46. https://doi.org/10.1186/s12711-021-00637-y

Madsen P., and Jensen J. (2013) DMU, a package of analysing multivariate mixed models. Ver. 6, rel 5.2. Available at: https://dmu.ghpc.au.dk/dmu/DMU/Doc/Current/dmuv6_guide.5.2.pdf

Sargolzaei M., Chesnais J.P., and Schenkel F.S. (2014) BMC Gen. 15:478. https://doi.org/10.1186/1471-2164-15-478

Sevillano C.A., Vandenplas J., Bastiaansen J.W.M., Bergsma R., and Calus M.P.L. (2017) Genet. Sel. Evol. 49:75. https://doi.org/10.1186/s12711-016-0234-9