J. Dairy Sci. 105:9822–9836
https://doi.org/10.3168/jds.2022-22225

Local breed proportions and local breed heterozygosity in genomic predictions for crossbred dairy cows

Jón H. Eiríksson,[1]* Ismo Strandén,[2] Guosheng Su,[1] Esa A. Mäntysaari,[2] and Ole F. Christensen[1]
[1]Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark
[2]Natural Resources Institute Finland (Luke), 31600 Jokioinen, Finland

ABSTRACT

For genomic prediction of crossbred animals, models that account for the breed origin of alleles (BOA) in marker genotypes can allow the effects of marker alleles to differ depending on their ancestral breed. Previous studies have shown that genomic estimated breeding values for crossbred cows can be calculated using the marker effects that are estimated in the contributing pure breeds and combined based on estimated BOA in the genotypes of the crossbred cows. In the presented study, we further exploit the BOA information for improving the prediction of genomic breeding values of crossbred dairy cows. We investigated 2 types of BOA-derived breed proportions: global breed proportions, defined as the proportion of marker alleles assigned to each breed across the whole genome; and local breed proportions (LBP), defined as the proportions of alleles on chromosome segments which were assigned to each breed. Further, we investigated 2 BOA-derived measures of heterozygosity for the prediction of total genetic value. First, global breed heterozygosity, defined as the proportion of marker loci that have alleles originating in 2 different breeds over the whole genome. Second, local breed heterozygosity (LBH), defined as proportions of marker loci on chromosome segments that had alleles originating in 2 different breeds. We estimated variance related to LBP and LBH on the remaining variation after accounting for prediction with solutions from the genomic evaluations of the pure breeds and validated alternative models for production traits in 5,214 Danish crossbred dairy cows. The estimated LBP variances were 0.9, 1.2, and 1.0% of phenotypic variance for milk, fat, and protein yield, respectively. We observed no clear LBH effect. Cross-validation showed that models with LBP effects had a numerically small but statistically significantly higher predictive ability than models only including global breed proportions. We observed similar improvement in accuracy by the model having an across crossbred residual additive genetic effect, accounting for the additive genetic variation that was not accounted for by the solutions from purebred. For genomic predictions of crossbred animals, estimated BOA can give useful information on breed proportions, both globally in the genome and locally in genome regions, and on breed heterozygosity.

Key words: crossbreeding, genomic selection, breed of origin of alleles, heterozygosity, heterosis

INTRODUCTION

Crossbreeding (i.e., the mating of animals from different breeds) is a common practice in many livestock production programs. Crossbred animals show superior performance for many important traits compared with purebred animals, so-called heterosis (Falconer and Mackay, 1996). Traditionally, crossbreeding has been more common in meat production systems, such as for beef cattle, pigs, and poultry, than for dairy cattle. However, the benefit of crossbreeding for dairy cattle has gained interest in recent decades (Sørensen et al., 2008; Kargo et al., 2014; Clasen et al., 2020).

Crossbreeding gives challenges to genomic prediction. Among those is the difference in the linkage disequilibrium (**LD**) between markers and QTL between breeds, and thus higher LD within breeds than between breeds (Ibáñez-Escriche et al., 2009; Lund et al., 2014). To accommodate for this difference, models that account for the breed origin of alleles (**BOA**) in genotypes of crossbred animals were proposed (Ibáñez-Escriche et al., 2009; Christensen et al., 2014). However, these models have not always resulted in more accurate predictions than the models that assume the same marker allele effects across breeds (Ibáñez-Escriche et al., 2009; Sevillano et al., 2017; Guillenea et al., 2022).

For dairy cattle, VanRaden et al. (2020) calculated genomic EBV (**GEBV**) for crossbred animals using solutions from purebred evaluations weighted by breed proportions. Eiríksson et al. (2021, 2022) presented a method based on BOA where the GEBV of crossbred

Received April 25, 2022.
Accepted July 19, 2022.
*Corresponding author: jonh@qgg.au.dk

were also based on solutions from purebred evaluations. Such methods require breed level estimates, weighted by the breed proportions, to account for the different genetic levels of the breeds (VanRaden et al., 2020; Eiríksson et al., 2021). VanRaden et al. (2020) used multiple breed genetic evaluation models, which estimated the breed levels. However, it is practical to use solutions from routine genetic evaluations. The routine evaluations are often performed for each of the pure breeds separately, and these do not provide breed level estimates. Eiríksson et al. (2022) suggested estimating the breed levels from purebred phenotypes. Their approach, however, requires assumptions that are unlikely to hold in all instances. In particular, they assumed that the level differences between breeds were only genetic and not related to the different breeds being raised in different environments. An alternative is to estimate the breed levels from phenotypes of crossbred animals.

Estimated BOA (Vandenplas et al., 2016) of crossbred animals can be used to calculate 2 types of estimates for each breed contribution: global breed level and local breed level. The global breed proportion (**GBP**) is the proportion of alleles throughout the genome that originate from the breed in question, and the local breed proportion (**LBP**) is the proportion of alleles locally for segments of the genome that originate from the breed. Thus, a GBP effect attempts to account for the different overall levels of the breeds contributing to crossbred animals, and an LBP effect attempts to indicate segment-specific breed levels contributing to crossbred animals.

In addition to the differences in LD and breed levels between breeds, the mixture of genetic backgrounds complicates predictions for crossbred animals. In particular, the increased heterozygosity in the genomes of crossbred animals, which contributes to heterosis, is of interest to study. Genomic predictions for heifers and cows can be used for 2 types of selection. First, which type of semen should be used in their insemination, and second, whether the animal should enter the herd or be sold (Calus et al., 2015; Hjortø et al., 2015; Clasen et al., 2021). In the latter case, it is not only the breeding value that matters, but the level of heterosis also contributes to the genetic potential for production. Therefore, estimated total genetic value (**ETGV**), including heterosis, for crossbred heifers can be useful. Expected heterosis based on pedigree information can be fitted in genetic evaluations (e.g., Lidauer et al., 2006). This kind of heterozygosity estimate accounts for the difference in the average performance of crossbred animals due to the expected increase in heterozygosity when alleles come from different breeds. However, for animals with at least 1 crossbred parent (e.g., 3-way crosses and backcrosses), the true number of loci that carry alleles from different breeds may differ from the expectation. Thus, more accurate estimates of heterozygosity, based on genomic information could be useful for predicting heterosis for crossbred dairy heifers.

Chromosome segments that have different BOA of the 2 haplotypes are more likely to have heterozygous QTL, than when both haplotypes have the same BOA. Estimated BOA can therefore give important information about increased heterozygosity of QTL in crossbred animals. First, estimated BOA gives the proportion of marker loci over the whole genome that have alleles originating in different breeds [i.e., global breed heterozygosity (**GBH**)]. Second, BOA could give information on whether the 2 alleles in marker loci or chromosome segments originated from the same breed or from 2 different breeds. We call this local breed heterozygosity (**LBH**). The LBH allows estimation of the effect of genome region-specific breed heterozygosity. An alternative could be to include genotype heterozygosity with dominance model (Xiang et al., 2016; Doekes et al., 2020).

The aims of this study were, first, to present how estimated BOA can be used to model the effects of GBP, LBP, GBH, and LBH in the genome of crossbred animals. Second, to compare model fit of genomic models with or without LBP or LBH effects. Third, to compare cross-validation predictive ability for GEBV and ETGV calculation using the same models. The investigation was made for milk production traits in Danish crossbred dairy cows. All models used solutions from the genomic evaluations of the contributing pure breeds as a starting point, but our aim was to investigate whether additional effects (e.g., effects of LBP and LBH) estimated from phenotypes and genotypes of the crossbred cows could improve the model fit and predictive ability.

MATERIALS AND METHODS

No animals were used in this study, and ethical approval for the use of animals was thus deemed unnecessary.

Data

Data had 5,214 Danish crossbred dairy cows born in the years 2012 to 2018 from 74 herds. The study focused on crosses of the Holstein (**H**), Jersey (**J**), and Nordic Red dairy cattle (**R**) breeds. Therefore, we excluded crossbred cows with contribution of more than 1/16 from other breeds according to registered pedigree. The cows were genotyped using EuroG MD chip (EuroGenomics, 2019). In addition, genotypes of 7,500 purebred animals, 2,500 from each of the 3 breeds, H, J, and

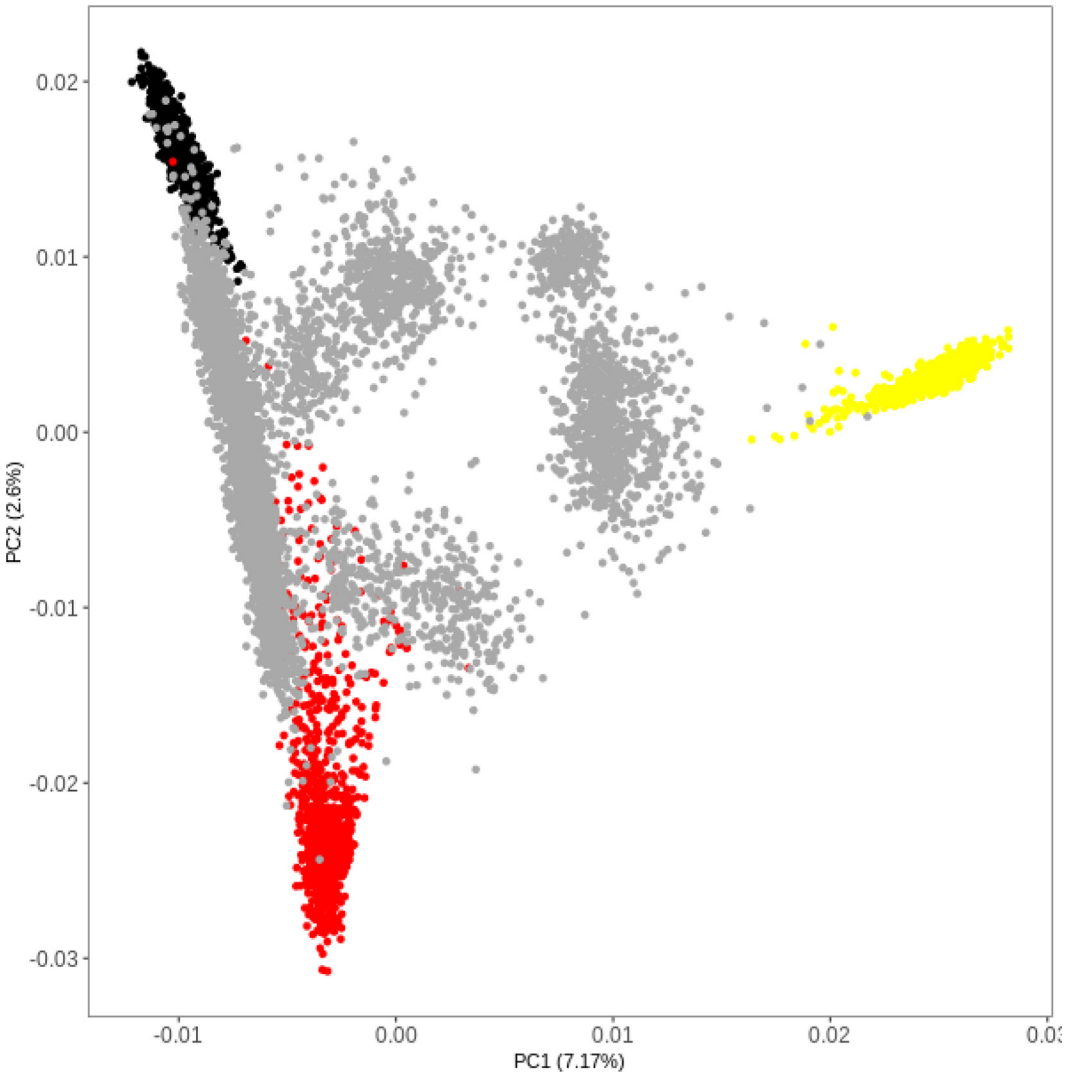

Figure 1. First (PC1) and second (PC2) principle components of genomic relationships for 1,000 Holstein (black), 1,000 Jersey (yellow), 1,000 Red dairy cattle (red) and 5,214 crossbred (dark gray) animals. Percentages are proportion of variance explained by the respective principle components.

R, genotyped on various 50k SNP chips, were included for the imputing, phasing, and assigning BOA steps. Figure 1 has a principal component plot, showing the first 2 principal components of the genomic relationship matrix between the crossbred cows, as well as 3,000 purebred animals, 1,000 from each of the 3 breeds. We phased and imputed the genotypes to a set of 50,684 markers using FImpute (Sargolzaei et al., 2014). We estimated BOA in the genotypes of the crossbred cows using the AllOr method (Eiríksson et al., 2021). Most of the cows included in this study were also included in Eiríksson et al. (2022), which contains further details on genotypes, imputation, and the BOA assignment. The phenotypes were 305-d lactation milk (**MY**), fat (**FY**), and protein (**PY**) yields for the first 3 parities of the cows, a total of 11,001 records.

We calculated GEBV for the crossbred cows based on BOA as described in (Eiríksson et al., 2022). In short, the GEBV were calculated using solutions from the separate genomic evaluations of the pure breeds. Using estimated BOA, we split the gene content of the crossbred cows into components, each component counting the reference alleles with assigned origin in one of the pure breeds. We multiplied each component with estimated marker effects from the respective breed. We call these GEBV primary genomic estimated breeding values (**pGEBV**). The separate purebred evaluations came from the milk production breeding value estima-

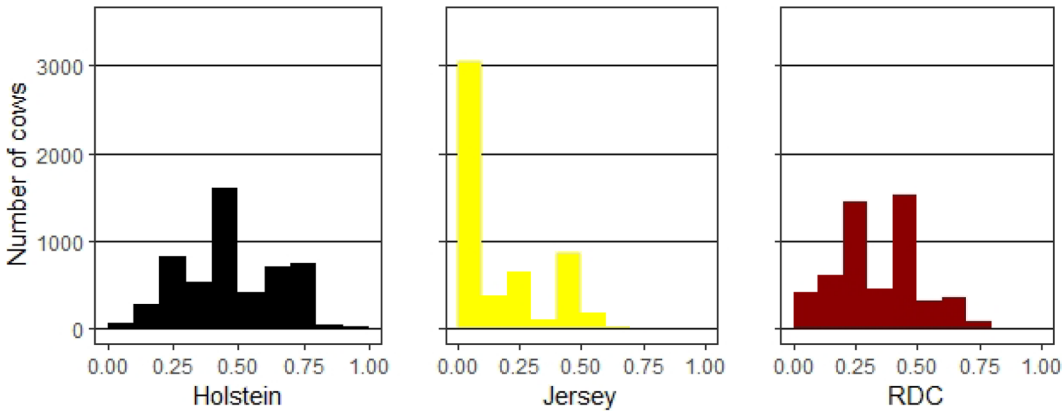

Figure 2. The distribution of breed proportions of Holstein, Jersey, and Red dairy cattle (RDC) in the 5,214 crossbred cows based on assigned breed of origin of alleles.

tions computed by Nordic Cattle Genetic Evaluation in December 2021 (NAV 2021) for each of the breeds (H, J, and R). Because we wanted to investigate if useful breed level correction factors could be estimated from the phenotypes of crossbred, we did not correct pGEBV for differences in breed levels.

We constructed 2 sets of corrected phenotypes, $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$, with and without correction for breed heterozygosity, respectively. In both cases, we corrected the original phenotypes for fixed effects and pGEBV. The phenotype corrected for heterozygosity was

$$\mathbf{y}_1^* = \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}\hat{\mathbf{a}}_{pGEBV},$$

where the $\mathbf{y}$ vector contains the original phenotypes, the $\mathbf{X}_1$ matrix connects the fixed effects to the phenotypes, the $\hat{\boldsymbol{\beta}}_1$ vector contains the fixed effect solutions of parity, herd-year, calving age, and breed heterozygosity from an animal model, $\mathbf{Z}$ is an incidence matrix connecting phenotypes to cows, and $\hat{\mathbf{a}}_{pGEBV}$ is the vector of pGEBV for the cows. The $\hat{\boldsymbol{\beta}}_1$ solutions were estimated with a pedigree-based animal model using a larger dataset, which had 207,116 lactation yields from 96,798 crossbred and purebred cows. The cows were in the same herds as the genotyped crossbred cows with phenotypes in $\mathbf{y}$. In that way, information from the purebred cows, and the crossbred cows without genotypes, strengthened the estimation of the fixed effects. Similarly, for the phenotypes not corrected for heterozygosity, we had

$$\mathbf{y}_2^* = \mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 - \mathbf{Z}\hat{\mathbf{a}}_{pGEBV},$$

where $\mathbf{X}_2\hat{\boldsymbol{\beta}}_2$ includes the same effects as $\mathbf{X}_1\hat{\boldsymbol{\beta}}_1$, except the breed heterozygosity effect was left out. The $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ fixed effects were estimates from the same model.

Global and Local Effects

Here, we explain the construction of global (across all loci) and local (for individual marker loci or chromosome segments) indicators of breed proportions and breed heterozygosity. The BOA output from AllOr contains, separate for the 2 gametes of the crossbred cows, an estimate of which breed the allele originated. From that, we built vectors of length equal to the number of markers ($m = 50{,}684$) for each animal i, gamete j ($j = 1, 2$) and breed b ($b = $ H, J, R), denoted $\mathbf{s}_{i,j}^b$ (i.e., 6 vectors for each animal). A marker in vector $\mathbf{s}_{i,j}^b$ had the value 1 for allele j assigned to breed b but 0 for allele j assigned to another breed. If the allele could not be assigned to a definite BOA, the corresponding value in $\mathbf{s}_{i,j}^b$ was a number between 0 and 1, according to the rules described for the output of AllOr (Eiríksson et al., 2021). For each marker and gamete, the sum across breeds was 1; that is, $\sum_{b\in\{H,J,R\}} \mathbf{s}_{i,j}^b = 1$.

Based on vectors $\mathbf{s}_{i,j}^b$, we built 2 types of estimates of the breed composition for each animal: GBP and LBP. We defined the LBP vector for animal i as $\mathbf{p}_i^b = (\mathbf{s}_{i,1}^b + \mathbf{s}_{i,2}^b)/2$. The GBP of animal i, denoted as $\overline{p_i^b}$, was the average of the values in $\mathbf{p}_i^b$. Figure 2 presents the distribution of GBP $\left(\overline{p_i^b}\right)$ in the crossbred animals. Additionally, to show the deviation in GBP from the expectation based on pedigree, we looked separately at the 695 3-way crossbred with H maternal grandsire or maternal granddam in the data. Figure 3 presents the distribution of GBP of H in this subset.

In addition to the single marker-based LBP vector, we defined LBP considering segments of 100 adjacent markers. For each animal i and breed b, the segment-based LBP vectors were constructed as an average of

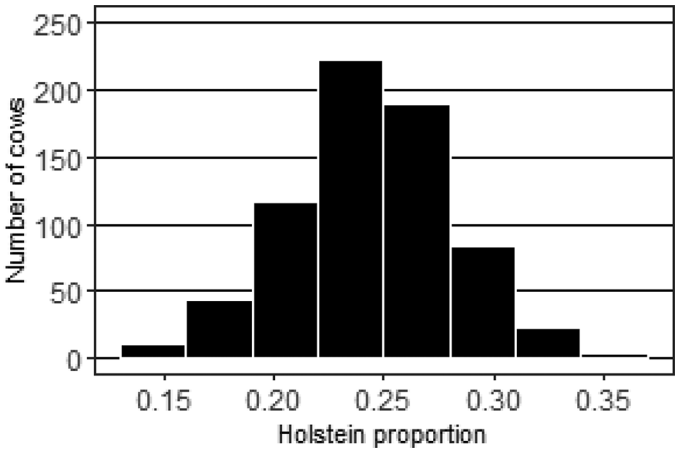

Figure 3. Distribution of global breed proportions of Holstein in 695 3-way crossbred cows with Holstein maternal grandsire or maternal granddam.

nonoverlapping chromosome segments of 100 adjacent values from $\mathbf{p}_i^b$, resulting in vector $\mathbf{p}_{(100),i}^b$ of length 506 marker segments. The last segment of each chromosome was combined with the second-to-last segment if its length was less than half of the defined segment length.

Similarly to the LBP vector, we constructed LBH for animal i between breeds b and b' as $\mathbf{t}_i^{b,b'} = \mathbf{s}_{i,1}^b \circ \mathbf{s}_{i,2}^{b'} + \mathbf{s}_{i,1}^{b'} \circ \mathbf{s}_{i,2}^b$, where $\circ$ is the element-wise multiplication and $b \neq b'$. We estimated GBH as the average of the values in $\mathbf{t}_i^{b,b'}$, denoted $\overline{\mathbf{t}_i^{b,b'}}$, which represents the proportion of loci that are assigned to 2 different breeds over the whole genome. Further, we constructed LBH of chromosome segments of length 100, denoted $\mathbf{t}_{(100),i}^{b,b'}$, which were based on the average of values from $\mathbf{t}_i^{b,b'}$ as described for LBP. Figure 4 presents the distribution of GBH values for the crossbred animals.

Models for GEBV

Four types of models for predicting the part of the breeding values of crossbred cows that was not captured by pGEBV are presented. First, a simple model having GBP, second, models having GBP and LBP effects, third, a model having GBP and a residual additive genetic (**RA**) effect by integrating an across crossbred genomic relationship matrix, and fourth, a model having GBP, LBP, and RA effects.

The simplest model for predicting GEBV had only regression on GBP and was named base model (**BaseM**). The model was

$$\mathbf{y}_1^* = \mathbf{1}\mu + \mathbf{Z}\mathbf{f}^H\eta^H + \mathbf{Z}\mathbf{f}^J\eta^J + \mathbf{Z}\mathbf{e}_{pe} + \mathbf{e}, \qquad [1]$$

where μ is the intercept, $\mathbf{1}$ is a vector of ones, vector $\mathbf{f}^b$ contains breed proportions p_i^b for breed $b \in \{H, J, R\}$ for all animals, η^b is the fixed effect of GBP of breed b, $\mathbf{e}_{pe}$ is vector of random permanent environment effects and $\mathbf{e}$ is the vector of random residuals. It was assumed that $\mathbf{e}_{pe} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_{pe}^2)$ and $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_{pe}^2 is the permanent environment variance and σ_e^2 is the residual variance. Note that $\mathbf{f}^b$ were only included for 2 breeds (H and J), because when an intercept is included in the model and $\mathbf{f}^H + \mathbf{f}^J + \mathbf{f}^R = \mathbf{1}$, there is no need to include fixed regression on all 3 breed proportions.

The local breed proportion model (**LBPM**) extended BaseM by including regression on LBP as random effects in a manner similar to having marker effects in a SNP-BLUP model. The LBP effects were included either for individual markers [LBPM(1)] or segments

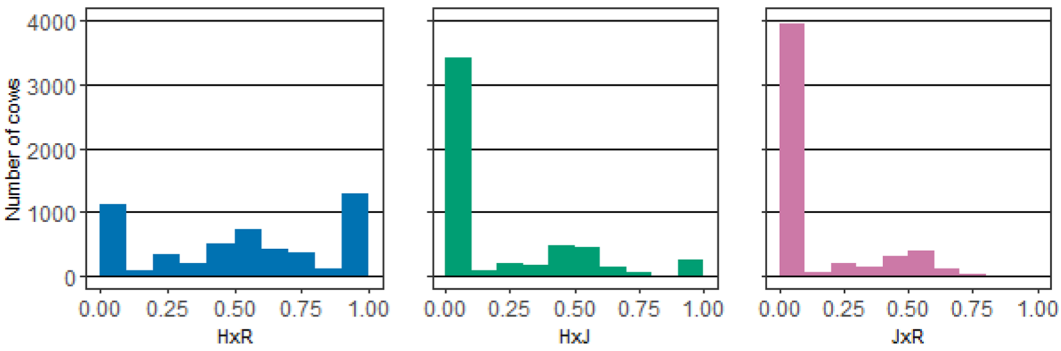

Figure 4. The distribution of global breed heterozygosity for Holstein/Red dairy cattle pairs (H×R), Holstein/Jersey pairs (H×J), and Jersey/Red dairy cattle pairs (J×R), for the 5,214 crossbred cows based on assigned breed of origin of alleles.

of 100 [LBPM(100)] markers. The model had the same general structure whether LBP was modeled for individual markers or for marker segments:

$$\mathbf{y}_1^* = \mathbf{1}\mu + \mathbf{Zf}^H\eta^H + \mathbf{Zf}^J\eta^J + \mathbf{ZP}^{*H}\boldsymbol{\gamma}^H + \mathbf{ZP}^{*J}\boldsymbol{\gamma}^J + \mathbf{ZP}^{*R}\boldsymbol{\gamma}^R + \mathbf{Ze}_{pe} + \mathbf{e}, \quad [2]$$

where $\mathbf{P}^{*b} = \mathbf{P}^b - \mathbf{f}^b\mathbf{1}'$, and $\boldsymbol{\gamma}^b$ is a vector with the regression coefficients on LBP for breed b. Here, matrix $\mathbf{P}^b$ contains the $\mathbf{p}_i^b\left(\mathbf{p}_{(100),i}^b\right)$ for each animal as a row. Consequently, $\mathbf{P}^{*b}$ contains the deviations from the breed proportion for each marker (segment). It was assumed that $\boldsymbol{\gamma}^b \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2_{\gamma,b})$, where $\sigma^2_{\gamma,b}$ is LBP marker (segment) variance for breed b.

When there are more marker segments than animals, a computationally more efficient equivalent LBPM can be constructed using LBP similarity matrices. The size of the resulting system of equations becomes proportional to the number of animals, rather than the number of markers, which is computationally advantageous, and the solutions are invariant to the transformation. Then, the LBP effect of breed b for the crossbred cows is $\mathbf{v}^b = \mathbf{P}^{*b}\boldsymbol{\gamma}^b$, and the LBP similarity matrix $\mathbf{Q}^b$ for breed b was constructed as $\mathbf{Q}^b = \dfrac{\mathbf{P}^{*b}\mathbf{P}^{*b\prime}}{\lambda_v}$, where λ_v is a normalizing constant defined as $\lambda_v = \dfrac{\mathrm{tr}(\mathbf{P}^{*b}\mathbf{P}^{*b\prime})}{n}$, where n is the number of animals. The related LBP variance for breed b is $\sigma^2_{v,b} = \sigma^2_{\gamma,b}\lambda_v$.

An alternative to using LBP to model the remaining genetic variation in $\mathbf{y}_1^*$, unexplained by pGEBV, is to use the genotypes directly. Thus, we investigated whether there was an additive genetic effect left in the crossbred data, here named RA genetic effect. The theoretical background for coding and scaling based on allele frequencies of base population for genetic relationship matrices, such as VanRaden (2008), are not applicable here, because, first, most of the additive genetic variance has already been accounted for, and second, the cows are crossbred rather than from a uniform population. Therefore, we constructed the RA relationship matrix as the normalized genomic additive relationship matrix (Vitezica et al., 2016): $\mathbf{G} = \dfrac{\mathbf{MM}'}{\mathrm{tr}(\mathbf{MM}')/n}$, where the $\mathbf{M}$ matrix contains the marker genotypes of the crossbred cows (irrespective of BOA) coded as −1, 0, and 1 for genotypes aa, Aa, and AA, respectively. In that way, we make no assumptions on the allele frequency of base population animals. The residual additive effects model (**RAM**) is

$$\mathbf{y}_1^* = \mathbf{1}\mu + \mathbf{Zf}^H\eta^H + \mathbf{Zf}^J\eta^J + \mathbf{Za} + \mathbf{Ze}_{pe} + \mathbf{e}, \quad [3]$$

where $\mathbf{a} \sim \mathrm{N}(\mathbf{0}, \mathbf{G}\sigma^2_a)$, and σ^2_a is the RA variance. Note that σ^2_a cannot be interpreted as traditional additive genetic variance.

Further, we investigated a model that combined the LBP effects and the RA effects (RA+LBPM). Thus, the model included the **Za** term from Equation 3 in addition to all the terms in Equation 2. We only tested RA+LBPM for segment length of 100 for LBP, denoted RA+LBPM(100).

Models for ETGV

We calculated ETGV as the estimated breeding value plus effects of heterozygosity, based on 2 main models with heterozygosity from genomic information. First, we tested models based on BOA (i.e., with GBH, and with and without accounting for LBH). Second, we modeled heterozygosity based on the genotype directly, either by considering genome-wide genotype heterozygosity, or by additionally accounting for local genotype heterozygosity using a dominance relationship matrix.

The global BOA heterozygosity model (**BOA-HM**) included both GBH from BOA information and the terms in BaseM in Equation 1:

$$\mathbf{y}_2^* = \mathbf{1}\mu + \mathbf{Zf}^H\eta^H + \mathbf{Zf}^J\eta^J + \mathbf{Zh}^{H,J}\tau^{H,J} + \mathbf{Zh}^{H,R}\tau^{H,R} + \mathbf{Zh}^{J,R}\tau^{J,R} + \mathbf{Ze}_{pe} + \mathbf{e}, \quad [4]$$

where $\mathbf{h}^{b,b'}$ is a vector with $\overline{t_i^{b,b'}}$ for each animal i, and $\boldsymbol{\tau}^{b,b'}$ is the fixed effect of proportions of loci with 1 allele assigned to breeds b and 1 allele assigned to breed b'.

For comparison, we tested the pedigree heterozygosity model (Ped-HM), where we replaced $\mathbf{h}^{b,b'}$ in Equation 4 with a pedigree-based estimate of breed heterozygosity, $\mathbf{h}^{b,b'}_{ped}$. We calculated the elements of $\mathbf{h}^{b,b'}_{ped}$ as $h^{b,b'}_{ped,i} = f^b_{ped,s} \times f^{b'}_{ped,d} + f^{b'}_{ped,s} \times f^b_{ped,d}$, where $f^b_{ped,s}$ and $f^b_{ped,d}$ are pedigree-based breed proportions of sire and dam of animal i, respectively.

The local breed heterozygosity model (**LB-HM**) extends BOA-HM by including random LBH effects. The LB-HM has the following form, regardless of whether LBH is modeled by single markers [**LB-HM(1)**] or marker segments [**LB-HM(100)**]:

$$\mathbf{y}_2^* = \mathbf{1}\mu + \mathbf{Zf}^H\eta^H + \mathbf{Zf}^J\eta^J + \mathbf{Zh}^{H,J}\tau^{H,J} + \mathbf{Zh}^{H,R}\tau^{H,R} + \mathbf{Zh}^{J,R}\tau^{J,R} + \mathbf{ZT}^{*H,J}\boldsymbol{\delta}^{H,J} + \mathbf{ZT}^{*H,R}\boldsymbol{\delta}^{H,R} + \mathbf{ZT}^{*J,R}\boldsymbol{\delta}^{J,R} + \mathbf{Ze}_{pe} + \mathbf{e}, \quad [5]$$

where $\mathbf{T}^{*b,b'} = \mathbf{T}^{b,b'} - \mathbf{h}^{b,b'}\mathbf{1}'$. The rows of the $\mathbf{T}^{b,b'}$ matrix contains the LBH vectors, $\mathbf{t}_i^{b,b'}\left(\mathbf{t}_{(100),i}^{b,b'}\right)$ for each animal, and therefore $\mathbf{T}^{*b,b'}$ contains the local deviations from the GBH of each cow. Further, $\boldsymbol{\delta}^{b,b'} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{I}\sigma^2_{\delta,b,b'}\right)$ are the effects of LBH between breeds b and b', where $\sigma^2_{\delta,b,b'}$ is the marker (segment) LBH variance. The LB-HM has an equivalent model with LBH similarity matrices instead of the $\mathbf{T}^{*b,b'}\boldsymbol{\delta}^{b,b'}$ terms. In that case, the LBH effect of breed pair b and b' for the crossbred cows is $\mathbf{u}_{b,b'} = \mathbf{T}^{*b,b'}\boldsymbol{\delta}^{b,b'}$ and the LBP similarity matrices were constructed as $\mathbf{K}^{b,b'} = \dfrac{\mathbf{T}^{*b,b'}\mathbf{T}^{*b,b'\prime}}{\lambda_u}$, where λ_u is a normalizing constant defined as $\lambda_u = \dfrac{\mathrm{tr}(\mathbf{T}^{*b,b'}\mathbf{T}^{*b,b'\prime})}{n}$. The related LBH variance is $\sigma^2_{u,b,b'} = \sigma^2_{\delta,b,b'}\lambda_u$.

In addition to considering breed heterozygosity based on the estimated BOA, we tested models that included the genotype heterozygosity instead of breed heterozygosity. Thus, the model had the marker genotype heterozygosity, irrespective of BOA (i.e., assuming heterozygosity of a locus was the same for all pairs of breeds). First, we considered a genotype heterozygosity model (**GT-HM**). Instead of including GBH effect $\mathbf{h}^{b,b'}\tau^{b,b'}$ as in Equation 4, the GT-HM had an effect of the proportion of loci that are heterozygous across the genome:

$$\mathbf{y}_2^* = \mathbf{1}\mu + \mathbf{Z}\mathbf{f}^H\eta^H + \mathbf{Z}\mathbf{f}^J\eta^J + \mathbf{Z}\mathbf{k}\kappa + \mathbf{Z}\mathbf{e}_{pe} + \mathbf{e}, \quad [6]$$

where vector $\mathbf{k}$ contains the proportion of marker loci that are heterozygous for each animal and κ is the fixed effect of genotype heterozygosity. Genotype heterozygosity can also be considered using the dominance relationship matrix, which accounts for heterozygosity at individual markers. The dominance heterozygosity model (**D-HM**) extends GT-HM:

$$\mathbf{y}_2^* = \mathbf{1}\mu + \mathbf{Z}\mathbf{f}^H\eta^H + \mathbf{Z}\mathbf{f}^J\eta^J + \mathbf{Z}\mathbf{k}\kappa + \mathbf{Z}\mathbf{d} + \mathbf{Z}\mathbf{e}_{pe} + \mathbf{e}, \quad [7]$$

where the $\mathbf{d}$ vector has the random dominance deviation effects for each animal. It was assumed that $\mathbf{d} \sim \mathrm{N}(\mathbf{0}, \mathbf{D}\sigma^2_d)$, where $\mathbf{D} = \dfrac{\mathbf{W}\mathbf{W}'}{\mathrm{tr}(\mathbf{W}\mathbf{W}')/n}$ is the genomic dominance relationship matrix following Vitezica et al. (2016). The $\mathbf{W}$ matrix has the genotypes coded as 1 for heterozygote and 0 for either homozygote, and σ^2_d is the dominance variance. Similar to the estimated RA variance in RAM, the σ^2_d estimated here cannot be interpreted as a population dominance variance (Vitezica et al., 2016).

Model Fit and Variance Components

We tested the models on the data presented above. First, we estimated variance components for the models using all data and compared the goodness-of-fit of the models.

For the LBPM models, we used the marker-based SNP-BLUP type models (Equation [2]) for segment lengths 100 markers, but the equivalent similarity matrix model for LBPM(1). Similarly, we used the SNP-BLUP version in Equation [5] for LB-HM(100), and the equivalent similarity matrix models for LB-HM(1).

We built and inverted the relationship and similarity matrices using Julia (Bezanson et al., 2017). We added a small value, 0.0001, to the diagonal of the LBP and LBH similarity matrices to ensure that they were invertible. We estimated REML variance components for all models based on all data using the average-information REML algorithm in the DMU package (Madsen and Jensen, 2013). However, for a few models, the algorithm did not converge. In those instances, we used the EM-REML algorithm instead, also using the DMU package.

We compared the goodness-of-fit of the more complicated models to the models having fixed global effects (i.e., BaseM, BOA-HM or GT-HM) with the likelihood ratio test. Significance in likelihood ratio test was determined by comparing difference in $-2\log(\mathrm{L})$, between models with a mixture (50/50) of 2 chi-squared distributions with N_m-N_{base} and N_m-N_{base} -1 degrees of freedom. Here, L is the REML likelihood, N_m is the number of parameters in the more advanced model and N_{base} is the number of parameters in the nested model. Additionally, we computed Bayesian information criterion [**BIC**; $\mathrm{BIC} = -2\log(\mathrm{L}) + N_m \times \log(n\text{-}N_{fixed})$], where n is the number of animals and N_{fixed} is the number of fixed effects (including intercept), and Akaike information criterion [**AIC**; $\mathrm{AIC} = -2\log(\mathrm{L}) + 2N_m$] for all the models (Meyer, 2001). Consequently, lower AIC or BIC indicates models with higher likelihood, with a penalty for the number of parameters. Therefore, the models with the lowest AIC or BIC are preferred.

Cross-Validation

We assessed the predictive ability of the models using random 5-fold cross-validation using similar settings as in Aliloo et al. (2016). We constructed 5 nonoverlapping validation sets out of the 5,214 crossbred cows of approximately equal size. The split was done randomly with 2 restrictions: first, paternal half-sib groups were kept in the same set, and second, cows that had daughters among the crossbred cows were not included in the

validation set to avoid the unrealistic scenario of predicting dams based on information from their daughters. For each of the validation sets, we set the phenotypes in $\mathbf{y}_1^*\left(\mathbf{y}_2^*\right)$ as missing for the validation animals and estimated the effects in the models based on the remaining data. The total GEBV for animal i was subsequently constructed as $GEBV_i = \hat{a}_{pGEBV,i} + \hat{y}^*_{1,i}$, where $\hat{y}^*_{1,i}$ is the sum of all estimated effects, except the permanent environment, in each model, including the intercept.

Predictive ability of the models for predicting GEBV, denoted $r_{GEBV,y}$, was estimated as weighted correlation between total GEBV and corrected phenotypes adjusted for their respective fixed effects. If a cow had yield records from multiple lactations (i.e., from second and third in addition to the first lactation), we used an average of the 2 or 3 lactation yields. The weighted correlation used the reliability based on the number of records and assuming the heritability of 0.30 and the repeatability of 0.45 as weights (the weights ranged from 0.3 to 0.47). Using reliability as weights follows the Interbull standard validation test for genomic predictions (Mäntysaari et al., 2010). We also estimated the dispersion bias of the prediction with the b_1 coefficient from the weighted linear regression of $\mathbf{y}_1$ on GEBV.

We evaluated the predicted total genetic values from the ETGV models in a similar manner as for the GEBV models. We calculated the ETGV for animal i as $ETGV_i = \hat{a}_{pGEBV,i} + \hat{y}^*_{2,i}$, where $\hat{y}^*_{2,i}$ is the sum of all estimated effects, except permanent environment, from the respective model for cow i, including the intercept. The ETGV were compared to the corrected phenotypes that were not corrected for heterozygosity, $\mathbf{y}_2 = \mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2$. The predictive ability correlation $\left(r_{ETGV,y}\right)$ and b_1 was assessed in the same manner as for GEBV.

To get an estimate of the sampling variance of the correlation estimates and to assess whether they were different, we used 10,000 bootstrap samples of the validation animal GEBV and corrected phenotypes for each trait. From each sample, we calculated $r_{GEBV,y}$ and b_1 from the models. We tested statistical significance at $P < 0.05$ for each pair of models from the distribution of the difference of $r_{GEBV,y}$ and b_1 between the alternative models [i.e., whether 0 (no difference) was within the 95% CI of the differences in $r_{GEBV,y}$ and b_1]. We performed the bootstrap procedure in the same way to test for differences in the models for ETGV. Note that this testing of difference between 2 models takes into account that the estimated $r_{GEBV,y}$ and b_1 are strongly correlated between the alternative models.

RESULTS

Model Fit and Variance Components

Table 1 shows the estimated variance components, −2log(L), BIC, and AIC for models for the BaseM, LBPM(1), LBPM(100), RAM, and RA+LBPM(100) models for GEBV. The LBPM(1), LBPM(100), RAM, and RA+LBPM(100) models fitted the data significantly ($P < 0.01$) better than the BaseM for all 3 traits based on likelihood ratio test. For all 3 traits, BIC was lowest for RAM whereas AIC was lowest for RA+LBPM(100).

The total of estimated LBP variance summed over the 3 breeds for MY from the LBPM(1) was 30,600 kg^2, which equals 2.9% of the estimated genetic variance from the pedigree-based animal model. The LBP variance for MY accounted for 0.9% of the phenotypic variance. The estimated LBP variances for all traits were slightly lower when we considered the LBP for chromosome segments rather than individual markers. For MY, the estimated LBP variances were 29,600 kg^2 for LBPM(100). The total estimated LBP variances from LBPM(1) were 57.9 and 31.2 kg^2 for FY and PY, respectively, summed over the 3 breeds. The total estimated LBP variances were 1.2 and 1.0% of the phenotypic variance for FY and PY, respectively.

Table 2 has the estimated variance components, −2log(L), BIC and AIC for Ped-HM, BOA-HM, LB-HM(1), LB-HM(100), GT-HM, and D-HM models for ETGV. For PY and MY, inclusion of LBH effects did not improve goodness-of-fit. Further, BIC and AIC were lowest for BOA-HM and the estimated LBH variances were not significantly different from zero for any of the models for MY and PY. For FY, however, the LB-HM models fitted the data significantly ($P < 0.01$) better than the BOA-HM based on likelihood ratio test. The AIC was also lower for the full LB-HM models, irrespective of segment length, than for BOA-HM for FY. Because of the different fixed effects in the models, the GT-HM and D-HM models could not be compared with the Ped-HM, BOA-HM, and LB-HM models based on the REML likelihoods. For all 3 traits, the D-HM fitted the data significantly ($P < 0.01$) better than the GT-HM based on likelihood ratio test and had lower BIC and AIC.

Cross-Validation

The cross-validation results for calculating GEBV are in Table 3. Differences between the models were generally small, with the largest difference between the highest and the lowest $r_{GEBV,y}$ being 0.012 for PY. The

Table 1. Estimated variance components, the log of the maximum likelihood [−2log(L)], Bayesian information criterion (BIC), and Akaike information criterion (AIC) in models for predicting breeding value[1]

Model[2]	$\sigma^2_{v,H}$	$\sigma^2_{v,J}$	$\sigma^2_{v,R}$	σ^2_a	σ^2_{pe}	σ^2_e	−2log(L)	BIC	AIC
Milk yield[3]									
BaseM					9,014	16,550	0	0	0
LBPM(1)	207	0	99		8,720	16,552	−19*	7	−13
LBPM(100)	198	1	98		8,722	16,552	−19*	7	−13
RAM				3,664	7,258	16,515	−56*	−47	−54
RA+LBPM(100)	154	0	85	3,304	7,186	16,518	−67*	−33	−59
Fat yield[4]									
BaseM					1,555	2,943	0	0	0
LBPM(1)	51	7	0		1,503	2,943	−16*	10	−10
LBPM(100)	48	7	0		1,506	2,943	−15*	10	−9
RAM				558	1,292	2,937	−40*	−31	−38
RA+LBPM(100)	45	0	5	516	1,268	2,937	−51*	−17	−43
Protein yield[4]									
BaseM					1,039	1,761	0	0	0
LBPM(1)	25	0	7		1,009	1,761	−18*	8	−12
LBPM(100)	23	0	7		1,010	1,761	−18*	8	−12
RAM				462	821	1,756	−62*	−54	−60
RA+LBPM(100)	19	0	6	430	811	1,756	−74*	−39	−66

[1]$\sigma^2_{v,H}$ = Local breed proportion variance for Holstein; $\sigma^2_{v,J}$ = local breed proportion variance for Jersey; $\sigma^2_{v,R}$ = local breed proportion variance for Red dairy cattle; σ^2_a = residual additive genetic variance; σ^2_{pe} = permanent environment variance; σ^2_e = residual variance. The value for BaseM was subtracted from all −2log(L), BIC, and AIC.

[2]BaseM = base model, fitting global breed proportions; LBPM(m) = local breed proportion model with segment length of m markers; RAM = residual additive effect model; RA+LBPM(100) = residual additive effect model including local breed proportion effects with segment length of 100 markers.

[3]Variances for milk yield are given in (10 kg)^2.

[4]Variances for fat yield and protein yield are given in kg^2.

*Models have significantly ($P < 0.01$) better fit than the BaseM model based on likelihood ratio test.

highest $r_{GEBV,y}$ was from the RA+LBPM(100) for all 3 traits, 0.581, 0.396, and 0.432 for MY, FY, and PY, respectively. The standard deviations of the $r_{GEBV,y}$ across the 10,000 bootstrap samples were 0.010, 0.012, and 0.012, for MY, FY, and PY, respectively, with only very minor differences between models. For all 3 traits, the LBPM models gave significantly ($P < 0.05$) higher $r_{GEBV,y}$ than BaseM. The RAM and the LBPM models had similar $r_{GEBV,y}$ for all 3 traits, significantly ($P < 0.05$) higher $r_{GEBV,y}$ than from BaseM with the exception of RAM for FY where the difference was not significant.

In general, the b_1 coefficients for GEBV were close to 1 for all models for MY and PY, indicating low dispersion bias. The values ranged from 0.983 to 1.011 for MY, and 0.944 to 0.982 for PY. The predictions were somewhat inflated for FY, with b_1 ranging from 0.881 to 0.907 for FY across models. For all 3 traits, the lowest b_1 was from RAM and the highest from LBPM(100).

The cross-validation results for ETGV are in Table 4. The $r_{ETGV,y}$ and b_1 values across the models were almost identical for all traits. The range in $r_{ETGV,y}$ was from 0.568 to 0.573 for MY, 0.425 to 0.429 for FY, and 0.399 to 0.407 for PY. Standard deviations of $r_{ETGV,y}$ across the 10,000 bootstrap samples were 0.010, 0.012, and 0.012, for MY, FY, and PY, respectively, for all tested models. Among the models that only included fixed regressions on genome-wide heterozygosity and no random effects (i.e., Ped-HM, BOA-HM, and GT-HM), the BOA-HM had the highest $r_{ETGV,y}$ for all the traits. However, the differences in $r_{ETGV,y}$ and b_1 between these models were not statistically significant. Similar to the model fit results, the cross-validation results for FY from the LB-HM models differed from those for MY and PY. For MY and PY, the $r_{ETGV,y}$ from BOA-HM and both LB-HM models were almost identical. The highest $r_{ETGV,y}$ for MY and PY was achieved by D-HM, significantly ($P < 0.05$) higher than from GT-HM. However, the differences were not significant when compared with the BOA-HM and the LB-HM models. For FY, the LB-HM models had the highest $r_{ETGV,y}$ for both segment lengths, significantly higher than BOA-HM. Further, the D-HM had similar $r_{ETGV,y}$ as the LB-HM models, and significantly higher than GT-HM. Similar to the effects on $r_{ETGV,y}$, the inclusion of LBH effects did not affect b_1 for MY and PY. The b_1 for FY were lower than for the other traits, but without considerable differences between the studied models.

Table 2. Estimated variance components, the log of the maximum likelihood [−2log(L)], Bayesian information criterion (BIC), and Akaike information criterion (AIC) in models for predicting total genetic value[1]

Model[2]	$\sigma^2_{u,H,R}$	$\sigma^2_{u,H,J}$	$\sigma^2_{u,J,R}$	σ^2_d	σ^2_{pe}	σ^2_e	−2log(L)	BIC	AIC
Milk yield[3]									
Ped-HM					8,978	16,547	10	10	10
BOA-HM					8,947	16,546	0	0	0
LB-HM(1)	8	0	0		8,940	16,546	0	26	6
LB-HM(100)	9	4	0		8,939	16,546	0	26	6
GT-HM					8,326	17,498	300[5]	300[5]	300[5]
D-HM				2,776	5,852	17,478	265[5]**	273[5]	267[5]
Fat yield[4]									
Ped-HM					1,543	2,943	11	11	11
BOA-HM					1,537	2,943	0	0	0
LB-HM(1)	6	1	32		1,496	2,944	−11*	15	−5
LB-HM(100)	6	1	31		1,497	2,944	−11*	15	−5
GT-HM					1,394	3,182	422[5]	422[5]	422[5]
D-HM				673	1,011	3,178	398[5]**	406[5]	400[5]
Protein yield[4]									
Ped-HM					1,002	1,786	48	48	48
BOA-HM					1,028	1,760	0	0	0
LB-HM(1)	0	0	1		1,027	1,760	0	26	6
LB-HM(100)	0	0	1		1,028	1,760	0	26	6
GT-HM					1,002	1,786	55[5]	55[5]	55[5]
D-HM				548	689	1,783	15[5]**	23[5]	17[5]

[1]$\sigma^2_{v,H,R}$ = Local breed heterozygosity variance for Holstein/Red dairy cattle cross; $\sigma^2_{v,H,J}$ = local breed heterozygosity variance for Holstein/Jersey cross; $\sigma^2_{v,J,R}$ = local breed heterozygosity variance for Jersey/Red dairy cattle cross; σ^2_d = dominance variance; σ^2_{pe} = permanent environment variance; σ^2_e = residual variance. The value for BOA-HM was subtracted from all −2log(L), BIC, and AIC.

[2]Ped-HM = pedigree global heterozygosity model. BOA-HM = breed of origin global heterozygosity model. LB-HM(m) = local breed heterozygosity model with segment length of m markers. GT-HM = global genotype heterozygosity model. D-HM = dominance model.

[3]Variances are given in (10 kg)^2.

[4]Variances are given in kg^2.

[5]Not comparable to Ped-HM, BOA-HM, and LB-HM models because the fixed effects are different.

*Models have significantly ($P < 0.01$) better fit than the BOA-HM model based on likelihood ratio test.

**Models have significantly ($P < 0.01$) better fit than the GT-HM model based on likelihood ratio test.

DISCUSSION

We have presented an investigation about whether the assigned BOA could provide information on breed proportions and breed heterozygosity that are useful for genomic prediction, in addition to BOA allowing for breed-specific marker effects. We found that inclusion of either LBP or RA did improve model fit for genomic models for production traits of Danish crossbred dairy cows, and resulted in a slight improvement of predictive ability, of up to 1 percentage point. Local effects of heterozygosity did not improve fit or predictive ability for ETGV for MY and PY, but for FY a statistically significant LBH effect of J × R was found.

Table 3. Correlations between genomic estimated breeding values and corrected phenotypes ($r_{GEBV,y}$) and dispersion bias (b_1) of prediction for milk yield, fat yield, and protein yield

	Milk yield		Fat yield		Protein yield	
Model[1]	$r_{GEBV,y}$	b_1	$r_{GEBV,y}$	b_1	$r_{GEBV,y}$	b_1
BaseM	0.574[c]	1.000[b]	0.387[c]	0.885[b]	0.420[c]	0.961[b]
LBPM(1)	0.577[b]	1.009[a]	0.392[ab]	0.906[a]	0.425[ab]	0.981[a]
LBPM(100)	0.578[ab]	1.011[a]	0.392[ab]	0.906[a]	0.426[ab]	0.982[a]
RAM	0.578[b]	0.983[c]	0.391[bc]	0.881[b]	0.428[b]	0.944[c]
RA+LBPM(100)	0.581[a]	0.993[b]	0.396[a]	0.901[ab]	0.432[a]	0.961[b]

[a–c]Differences in correlations and b_1 between models with a common superscript are not statistically significant ($P < 0.05$) based on the confidence interval of the differences between models in 10,000 bootstrap samples.

[1]BaseM = base model, fitting global breed proportions; LBPM(m) = local breed proportion model with segment length of m markers; RAM = residual additive effect model; RA+LBPM(100) = residual additive effect model including local breed proportion effects with segment length of 100 markers.

Table 4. Correlations between estimated total genetic value and corrected phenotypes ($r_{ETGV,y}$) and dispersion bias (b_1) of prediction for milk yield, fat yield, and protein yield

	Milk yield		Fat yield		Protein yield	
Model[1]	$r_{ETGV,y}$	b_1	$r_{ETGV,y}$	b_1	$r_{ETGV,y}$	b_1
Ped-HM	0.568^{ab}	1.000^{a}	0.425^{cd}	0.907^{a}	0.400^{b}	0.960^{a}
BOA-HM	0.569^{ab}	0.999^{ab}	0.426^{bcd}	0.907^{a}	0.402^{ab}	0.961^{a}
LB-HM(1)	0.569^{ab}	0.999^{ab}	0.429^{a}	0.913^{a}	0.402^{ab}	0.962^{a}
LB-HM(100)	0.569^{ab}	0.999^{ab}	0.429^{ab}	0.913^{a}	0.402^{ab}	0.962^{a}
GT-HM	0.568^{b}	1.000^{a}	0.423^{d}	0.906^{a}	0.399^{b}	0.958^{a}
D-HM	0.573^{a}	0.992^{b}	0.429^{abc}	0.905^{a}	0.407^{a}	0.953^{a}

$^{a–d}$Differences in correlations and b_1 between models with a common superscript are not statistically significant ($P < 0.05$) based on the confidence interval of the differences between models in 10,000 bootstrap samples.

[1]Ped-HM = Pedigree heterozygosity model; BOA-HM = global breed heterozygosity model; LB-HM(m) = local breed heterozygosity model with segment length of m markers; GT-HM = global genotype heterozygosity model; D-HM = dominance model.

Breed Proportions

In BOA models for the calculation of GEBV for the crossbred animals, Eiríksson et al. (2022) and Guillenea et al. (2022) accounted for breed levels based on estimates of breed proportions from BOA assignment, similarly as GBP. Other studies have used pedigree-based estimates of breed proportions (Makgahlela et al., 2013), breed base representation (VanRaden et al., 2020), or the output from the Admixture software (Khansefid et al., 2020) to account for breed proportions in crossbred animals for genomic prediction. Genotype-based approaches should be able to account for deviations in proportion of genome from the expectation based on breed composition of parents. This is evident for 3-way terminal crosses, where the crossbred parent, typically the dam, has a breed proportion of 0.50 for both contributing breeds, but the genomic breed proportion of the offspring can vary considerably. Figure 3 shows how this distribution looks for the 3-way crosses in our data. Among the 695 cows with H grandsire or granddam, 18% had GBP of H outside the 0.20 to 0.30 range. Based on pedigree information they would all get the breed proportion 0.25.

We are not aware of any published studies using BOA to investigate the effects of LBP as done in this study. Differences in the QTL allele frequencies between breeds can lead to differences in genetic levels of breeds. Modeling the breed levels globally, as in BaseM, only accounts for the effects of the QTL allele frequency difference averaged over the genome (i.e., it assumes that these effects are evenly distributed across the genome). When LBP is included, local differences in QTL allele frequencies are accommodated for. The estimated marker effects from the pure breeds, which we used for pGEBV, can only predict the effects of QTL that are both segregating within breed and in LD with markers that are also segregating within the breed. Consequently, QTL that are fixed in one breed, but segregating or fixed with another allele in the other breeds, are not contributing to pGEBV. However, the LBP effects capture the effect of having the chromosome segment with the QTL from the breed with the fixed QTL.

The results in this study (Table 1) indicate a significant LBP variance in crossbred cows from H, J, and R breeds for production traits, particularly for the H proportion. However, for comparison of the estimated variances, the contribution of each breed to the crossbred group has to be considered. Of the 3 breeds considered, the contribution of H to the crossbred animals was the largest (Figure 2). The presented LBP variances, $\sigma^2_{v,b}$, depend on λ_v, a normalizing constant including the trace of $\mathbf{P}^{*b}\mathbf{P}^{*b'}$, and hence on the diagonal of the matrix. An animal without the breed b contribution has 0 as diagonal element in the $\mathbf{Q}^b$ matrix. The same is true if both of its parents are purebred, such that the LBP is constant (0.5) throughout the genome. More than half of the animals in this study had no J contribution, which could partly explain the low LBP variance estimated for J compared to H and R. However, J is less related to the H and R breeds than the relationship between the H and R breeds (Figure 1 and Gautason et al., 2020). Therefore, larger LBP effects could be expected for J, if the breed contributions were the same for all breeds, which was not the case in this study (Figure 2).

Christensen et al. (2015) constructed the segregation partial relationship matrix between the maternal breeds of 3-way crossbred animals based on BOA in a similar manner as we constructed the LBP effects in our study. Segregation between breeds such as LBP is related to difference in allele frequencies of QTL (Lo et al., 1993). The segregation term is, however, defined

between pairs of breeds (Lo et al., 1993; García-Cortés and Toro, 2006), whereas LBP is defined only for a given breed. The relationship between LBP and breed segregation needs further investigation.

The estimated LBP variances decreased slightly with increasing chromosome segment length for the LBP from 1 to 100 markers (Table 1). However, the segment length did not affect $r_{GEBV,y}$. The longer the segment length, the more likely there are recombination points within segments that show difference from one breed origin to another. Consequently, the values in the $\mathbf{p}^{b}_{(100),i}$ vector becomes closer to GBP and, therefore, dilute the variation in LBP. Additionally, QTL with opposite effects are more likely to be within the same segments when the segments are long, also diluting possible LBP effects. Majority of the crossbred animals were from simple crosses such as F_1, 3-way crosses, and first generations of rotational crossbreeding. Therefore, the genomes of the animals were expected to be constructed from relatively long segments originating from the pure breeds [Further information on the genotyped cows is in Eiríksson et al. (2022)]. After applying crossbreeding for many generations, the chromosome segments with the same BOA are expected to be shorter. The differences between model fit and predictive ability between LBPM(1) and LBPM(100) might thus be larger for more complicated crossbreeding scenarios. We made additional tests using LBP models with segment length of 20 and 500 markers. For segment length of 20 markers, the results were almost identical to those from LBPM(1). For segment length of 500 markers, the estimated LBP variances were somewhat lower than those from LBPM(100), but the predictive ability was similar.

In this study, we calculated pGEBV from estimated marker effects from separate purebred evaluations that used data from different purebred animals. Other authors (Karaman et al., 2021; Guillenea et al., 2022) have estimated within-breed marker effects in BOA models from combined data set of crossbred and purebred animals. In theory, LBP would also be relevant for such models because the marker effects depend on the within-breed-allele frequency.

Residual Additive Effects

The RAM considers the additive genetic effects twice. First, the within-breed genetic effect was predicted (i.e., pGEBV) based on solutions from the separate purebred evaluations. Second, the RA effect aims at capturing additive genetic effects that the within-breed prediction failed to capture. Among the reasons for the presence of RA variance could be, first, the different genetic background of purebred and crossbred animals, which results in different allele effects in the purebred and the crossbred animals (Christensen et al., 2014). Second, the estimated marker effects from the purebred populations are estimates, and therefore not completely accurate despite large reference groups. Third, the phenotypes of crossbred animals themselves are an information source for their genetic effects, which could improve model fit. Fourth, the RA variance could partly come from the allele frequency differences of the breeds, which should also be accounted by the LBP variance. The estimated σ^2_a from RA+LPBM(100) was indeed lower than its estimate from RAM for all 3 traits (Table 1) but the reduction was only a small proportion of the total estimated σ^2_a. The small increase in $r_{GEBV,y}$ from RAM compared with BaseM indicates that the RA effects were not very important for prediction in our data, where the number of crossbred animals was small, compared to the large reference data of purebred animals contributing to pGEBV.

Breeding Value Estimation

The proposed models here are add-ons to the GEBV calculation based on BOA and solutions from purebred evaluations (Eiríksson et al., 2021, 2022). The simplest model in our comparison for breeding value estimation, BaseM, accounted for the breed levels based on crossbred information and GBP rather than using phenotypes of purebred as in (Eiríksson et al., 2022), and required, therefore, no assumptions on the same environment for the involved purebred animals. For this data, only slight increase in predictive ability was obtained by adding either the LBP effects or the RA effects to the model. In practice, the simpler BaseM may therefore be the preferred model.

As presented above, the relevance of LBP depends on the breed composition of the crossbred animals included, and particularly the F_1 crosses do not contribute to the LBP variation. The difference in allele frequencies between the breeds in question also affects the relevance of LBP with large differences being likely to result in larger LBP effects. Therefore, the inclusion of LBP effects may have larger effect on prediction accuracy than observed in this study in the following situations: (1) groups of crossbred animals that include no or few F_1 crosses, and (2) crosses of breeds with large difference in allele frequencies of QTL. Further, the accuracy of genomic predictions depends on the number of phenotypic records included in the training set for estimation of marker effects (Meuwissen et al., 2001). In our study, phenotypic records on around 4,100 cows were included in each of the training sets for estimating LBP effects

for 3 breeds in the cross-validation. Training on larger data may result in more accurate estimates of the LBP effects, which should result in higher predictive ability.

Heterozygosity

When BOA assignment of the marker alleles is available, calculating GBH as an estimate of breed heterozygosity is simple. In principle, GBH should be able to account for realized breed heterozygosity, in contrast with estimates based on pedigree, which require the assumption that offspring receive exactly half of the breed components in their parents. The contribution of grandparent's breed in the genome of crossbreds can however vary considerably as shown in Figure 3.

The $\mathbf{k}\kappa$ term in Equations [6] and [7] was defined as the opposite of a homozygosity term in Xiang et al. (2016) and Doekes et al. (2020) (i.e., regression on genome-wide heterozygosity), except that in our study the κ parameter was estimated across breeds. Comparison of prediction ability of the BOA-HM, Ped-HM, and GT-HM models indicated that the BOA-derived heterozygosity indicator GBH was a similar or a better indicator for heterozygosity than the pedigree or genotype genome-wide heterozygosity indicators for ETGV calculation (Table 4). Breed heterozygosity indicates the proportion of the genome that has alleles from alternating breeds, including alleles of QTL. The difference in allele frequency between the breeds increases the probability of the 2 haplotypes at chromosome segments to be different compared to when the 2 haplotypes come from the same breed. That results in increased QTL heterozygosity in breed-heterozygous chromosome segments. Genotype heterozygosity indicates that the markers, which supposedly are linked to QTL, are heterozygous, regardless of breeds. However, for crossbred animals, linkage can be inconsistent based on the breed origin (Ibáñez-Escriche et al., 2009), and, thus, the marker heterozygosity might not be a good indicator for heterozygosity of QTL across breeds. Therefore, in the case of crossbred animals of breeds with low degree of shared QTL-marker LD, GBH may be a better heterozygosity indicator than genotype heterozygosity.

The results on LBH effect on FY are out of line with the other results of this study (Tables 2 and 4). First, we detected no variance of J × R LBH on the other 2 traits and no improvement of including LBH effect. Second, we detected no, or very low, variance of the other breed pair LBH effects on FY. Milk fat percentage is considerably higher for J cows than the other breeds (Årstatistik Avl, 2020). It would therefore not be surprising to detect genetic effects related to fat production that were unique for J or J crossbreeding. However, if the observed J × R LBH variance was related to J specific alleles, similar result would be expected for the H × J LBH effect. More data were available on H × J crossing than R × J (Figure 4), and yet we observed no variance explained with H × J effects (Table 2). Therefore, given the limited data we had for estimating the variance of J × R LBH effect (Figure 4), and the inconsistency with other results, further investigation is needed before any conclusions are made on the LBH effect on FY.

In purebred dairy cattle, dominance variance has been estimated for yield traits (Sun et al., 2014; Aliloo et al., 2016). However, results on increased accuracy of prediction when dominance effects are included, compared with models only including additive effects, are inconsistent (Sun et al., 2014; Aliloo et al., 2016). Further, Doekes et al. (2020) reported low dominance variance when regression on genome-wide inbreeding was included in the model for Holstein cattle, and only limited variation in inbreeding depression across the genome. The limited benefit of the dominance effects, when included in addition to the global heterozygosity indicators in this study, are consistent with these results. Additionally, low level of LD in the crossbred group and relatively few data may hamper our ability to estimate dominance effects.

For selection of crossbred heifers for milk production, providing ETGV can facilitate a good selection basis, in addition to GEBV. As an example, for crossbreeding systems where crossbred dairy cows are inseminated with beef semen (Kargo et al., 2014; Clasen et al., 2021), all selection among the crossbred cows is on their potential for production, rather than their potential to produce good offspring for dairy production. Therefore, accurate ETGV is more relevant than GEBV in that case. For predicting heterosis, GBH could be an interesting alternative to pedigree-based heterozygosity estimates. Further, GBH is an option for accounting for heterosis in phenotypes of genotyped crossbred animals for inclusion into genetic evaluation. Our results suggest that accounting for local heterozygosity is not important for production traits in crossbred dairy cows.

CONCLUSIONS

Assigned BOA can give information on breed proportions, both globally in the genome and locally in genome regions, which are useful for genomic prediction of crossbred dairy cows. We found significant variance for LBP effects on production traits in Danish crossbred dairy cows, of magnitude around 1% of phenotypic variance. The importance of LBP and the size of this variance depends on the breed composition of the crossbred animals. In our data, including LBP

or RA effects improved GEBV prediction for crossbred cows slightly, when the effects were added to prediction from solutions from separate purebred genomic evaluation. The increase in predictive ability was around a 0.5 percentage point. Assigned BOA can further give information on breed heterozygosity, which can be useful for either accounting for, or predicting, heterosis. From our results, we cannot see clear benefit from modeling heterozygosity locally in the genome rather than only globally across all loci.

ACKNOWLEDGMENTS

This study was a part of the DairyCross project supported by the Green Development and Demonstration Program (GUDP) from the Danish Ministry of Food, Agriculture and Fisheries (J. nr. 34009-18-1365; Copenhagen, Denmark). The authors have not stated any conflicts of interest.

REFERENCES

Aliloo, H., J. E. Pryce, O. González-Recio, B. G. Cocks, and B. J. Hayes. 2016. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet. Sel. Evol. 48:8. https://doi.org/10.1186/s12711-016-0186-0.

Årstatistik Avl. 2020. Husdyrinnovation Kvæg, SEGES, Skejby. Accessed Jul. 7, 2021. https://www.landbrugsinfo.dk/-/media/landbrugsinfo/public/e/0/b/aarsstat_2020.pdf.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah. 2017. Julia: A fresh approach to numerical computing. SIAM Rev. 59:65–98. https://doi.org/10.1137/141000671.

Calus, M. P. L., P. Bijma, and R. F. Veerkamp. 2015. Evaluation of genomic selection for replacement strategies using selection index theory. J. Dairy Sci. 98:6499–6509. https://doi.org/10.3168/jds.2014-9192.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su. 2015. Genetic evaluation for three-way crossbreeding. Genet. Sel. Evol. 47:98. https://doi.org/10.1186/s12711-015-0177-6.

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genet. Sel. Evol. 46:23. https://doi.org/10.1186/1297-9686-46-23.

Clasen, J. B., W. F. Fikse, M. Kargo, L. Rydhmer, E. Strandberg, and S. Østergaard. 2020. Economic consequences of dairy crossbreeding in conventional and organic herds in Sweden. J. Dairy Sci. 103:514–528. https://doi.org/10.3168/jds.2019-16958.

Clasen, J. B., M. Kargo, S. Østergaard, W. F. Fikse, L. Rydhmer, and E. Strandberg. 2021. Genetic consequences of terminal crossbreeding, genomic test, sexed semen, and beef semen in dairy herds. J. Dairy Sci. 104:8062–8075. https://doi.org/10.3168/jds.2020-20028.

Doekes, H. P., P. Bijma, R. F. Veerkamp, G. de Jong, Y. C. J. Wientjes, and J. J. Windig. 2020. Inbreeding depression across the genome of Dutch Holstein Friesian dairy cattle. Genet. Sel. Evol. 52:64. https://doi.org/10.1186/s12711-020-00583-1.

Eiríksson, J. H., K. Byskov, G. Su, J. R. Thomasen, and O. F. Christensen. 2022. Genomic predictions for crossbred dairy cows by combining solutions from purebred evaluation based on breed origin of alleles. J. Dairy Sci. 105:5178–5191. https://doi.org/10.3168/jds.2021-21644.

Eiríksson, J. H., E. Karaman, G. Su, and O. F. Christensen. 2021. Breed of origin of alleles and genomic predictions for crossbred dairy cows. Genet. Sel. Evol. 53:84. https://doi.org/10.1186/s12711-021-00678-3.

EuroGenomics. 2019. EuroGenomics genotyping microarray. Accessed Mar. 18, 2022. https://www.eurogenomics.com/eurog-md-chip.html.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman Group Ltd.

García-Cortés, L. A., and M. Á. Toro. 2006. Multibreed analysis by splitting the breeding values. Genet. Sel. Evol. 38:601–615. https://doi.org/10.1186/1297-9686-38-6-601.

Gautason, E., A. A. Schönherz, G. Sahana, and B. Guldbrandtsen. 2020. Relationship of Icelandic cattle with Northern and Western European cattle breeds, admixture and population structure. Acta Agric. Scand. A Anim. Sci. 69:25–38. https://doi.org/10.1080/09064702.2019.1699951.

Guillenea, A., G. Su, M. S. Lund, and E. Karaman. 2022. Genomic prediction in Nordic Red dairy cattle considering breed origin of alleles. J. Dairy Sci. 105:2426–2438. https://doi.org/10.3168/jds.2021-21173.

Hjortø, L., J. F. Ettema, M. Kargo, and A. C. Sørensen. 2015. Genomic testing interacts with reproductive surplus in reducing genetic lag and increasing economic net return. J. Dairy Sci. 98:646–658. https://doi.org/10.3168/jds.2014-8401.

Ibánez-Escriche, N., R. L. Fernando, A. Toosi, and J. C. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:12. https://doi.org/10.1186/1297-9686-41-12.

Karaman, E., G. Su, I. Croue, and M. S. Lund. 2021. Genomic prediction using a reference population of multiple pure breeds and admixed individuals. Genet. Sel. Evol. 53:46. https://doi.org/10.1186/s12711-021-00637-y.

Kargo, M., J. F. Ettema, M. Fjordside, and L. Hjortø. 2014. CombiCross–The use of new technologies for improving dairy crossbreeding programs. Proceedings of the 10th World Congr. Genet. Appl. Livest. Prod. Vancouver, Canada.

Khansefid, M., M. E. Goddard, M. Haile-Mariam, K. V. Konstantinov, C. Schrooten, G. de Jong, E. G. Jewell, E. O'Connor, J. E. Pryce, H. D. Daetwyler, and I. M. MacLeod. 2020. Improving genomic prediction of crossbred and purebred dairy cattle. Front. Genet. 11:598580. https://doi.org/10.3389/fgene.2020.598580.

Lidauer, M. H., E. A. Mäntysaari, I. Strandén, J. Pösö, J. Pedersen, U. S. Nielsen, K. Johansson, J.-A. A. Eriksson, P. Madsen, and G. P. Aamand. 2006. Random heterosis and recombination loss effects in a multibreed evaluation for Nordic Red Dairy Cattle. Abstract c24–02 in Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil. Brazilian Society of Animal Breeding.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations: Additive model. Theor. Appl. Genet. 87:423–430. https://doi.org/10.1007/BF00215087.

Lund, M. S., G. Su, L. Janss, B. Guldbrandtsen, and R. F. Brøndum. 2014. Genomic evaluation of cattle in a multi-breed context. Livest. Sci. 166:101–110. https://doi.org/10.1016/j.livsci.2014.05.008.

Madsen, P., and J. Jensen. 2013. DMU, A package of analyzing multivariate mixed models. Version 6, release 5.2. Accessed Jan. 28, 2021. https://dmu.ghpc.au.dk/dmu/DMU/Doc/Current/dmuv6_guide.5.2.pdf.

Makgahlela, M. L., E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen, M. J. Sillanpää, and J. Juga. 2013. Across breed multitrait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130:10–19. https://doi.org/10.1111/j.1439-0388.2012.01017.x.

Mäntysaari, E. A., Z. Liu, and P. VanRaden. 2010. Interbull validation test for genomic evaluations. Interbull Bull. 41:17–22.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Meyer, K. 2001. Estimates of direct and maternal covariance functions for growth of Australian beef calves from birth to weaning. Genet. Sel. Evol. 33:487–514. https://doi.org/10.1186/1297-9686-33-5-487.

NAV. 2021. NAV routine genetic evaluation of dairy cattle—Data and genetic models. Accessed Jan. 17, 2022. https://nordicebv.info/

wp-content/uploads/2021/10/NAV-routine-genetic-evaluation_EDITYSS-08102021.pdf.
Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.
Sevillano, C. A., J. Vandenplas, J. W. M. Bastiaansen, R. Bergsma, and M. P. L. Calus. 2017. Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. Genet. Sel. Evol. 49:75. https://doi.org/10.1186/s12711-017-0350-1.
Sørensen, M. K., E. Norberg, J. Pedersen, and L. G. Christensen. 2008. Invited review: Crossbreeding in dairy cattle: A Danish perspective. J. Dairy Sci. 91:4116–4128. https://doi.org/10.3168/jds.2008-1273.
Sun, C., P. M. VanRaden, J. B. Cole, and J. R. O'Connell. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS One 9:e103934. https://doi.org/10.1371/journal.pone.0103934.
Vandenplas, J., M. P. L. Calus, C. A. Sevillano, J. J. Windig, and J. W. M. Bastiaansen. 2016. Assigning breed origin to alleles in crossbred animals. Genet. Sel. Evol. 48:61. https://doi.org/10.1186/s12711-016-0240-y.
VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.
VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal Jr., I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. J. Dairy Sci. 103:1620–1631. https://doi.org/10.3168/jds.2019-16634.
Vitezica, Z. G., L. Varona, J.-M. Elsen, I. Misztal, W. Herring, and A. Legarra. 2016. Genomic BLUP including additive and dominant variation in purebreds and F_1 crossbreds, with an application in pigs. Genet. Sel. Evol. 48:6. https://doi.org/10.1186/s12711-016-0185-1.
Xiang, T., O. F. Christensen, Z. G. Vitezica, and A. Legarra. 2016. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet. Sel. Evol. 48:92. https://doi.org/10.1186/s12711-016-0271-4.

ORCIDS

Jón H. Eiríksson https://orcid.org/0000-0002-4655-2850
Ismo Strandén https://orcid.org/0000-0003-0161-2618
Esa A. Mäntysaari https://orcid.org/0000-0003-0044-8473
Ole F. Christensen https://orcid.org/0000-0002-8230-8062