Breed-origin-of-alleles approach using summary statistics for multi-breed genomic prediction in dairy cattle

J.B. Clasen[12*], W.F. Fikse[3], G. Su[2], and E. Karaman[2]

[1] Institution of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Ulls väg 26, 756 51 Uppsala, Sweden; [2] Centre for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, DK-8830 Tjele, Denmark; [3] Växa, Swedish University of Agricultural Sciences, Ulls väg 26, 756 51 Uppsala, Sweden; [*] julie.clasen@qgg.au.dk

Abstract

Systematic crossbreeding strategies between dairy cattle breeds in dairy herds are becoming more and more attractive to farmers, and this leads to a request for genomically enhanced breeding values for crossbred females in the dairy herds. Accurate genomic prediction of crossbred animals can be achieved if the genotypic and phenotypic data of the breeds involved in the crossbreeding are available to form reference populations, to estimate marker effects. However, sharing genotype and phenotype data between breed populations may be an issue due to privacy and competition. This study investigated genomic prediction of two-breed and three-breed rotational crossbred dairy cattle using summary statistics and a breed-origin of alleles model. The results indicate that the approach can yield almost as high prediction accuracies as having full information from the pure breeds.

Introduction

Crossbreeding between dairy cattle breeds and using genomic testing to predict genomically enhanced breeding values (GEBVs) on future replacement heifers in dairy herds are becoming more attractive to farmers (Magne and Quenon, 2021; Thomasen *et al*., 2020), and this leads to a stronger request from dairy farmers for predicting GEBVs for crossbred dairy cattle (Clasen *et al.*, 2021).

Previous studies have shown that combining genotype and phenotype data from multiple breeds and crossbreds in a joint reference population can yield higher prediction accuracies for crossbred animals than prediction based on single breed or crossbred reference populations (e.g., Vandenplas *et al*., 2018). Assuming homogeneous SNP effects across breeds and crossbreds may, in a multi-breed prediction model, not be the best approach if the structure of linkage disequilibrium (LD) and effects of quantitative trait loci (QTL) differs between the crossbred animals and their purebred ancestors (Vandenplas *et al*., 2016). Instead, assuming heterogeneous SNP effects by tracing each allele back to its breed-origin (BOA) may potentially yield higher genomic prediction accuracies for crossbred animals (Karaman *et al*., 2021).

Sharing genotype and phenotype data between breeding organisations or foreign countries is difficult due to privacy restrictions and competition (Tenopir *et al*., 2011). This issue makes genomic prediction for crossbred animals less accurate if the prediction relies on data from reference breeds outside the organisation, that cannot be obtained. A possible solution is to perform genomic prediction based on a meta-analysis of summary statistics including predicted allele substitution effects of single nucleotide polymorphism (SNP) markers and their prediction error (co)variances (Vandenplas *et al*., 2018). Such an approach is widely used in human genetics (e.g., Maier *et al*., 2018).

The aim of this study was to investigate the efficiency of using summary statistics as alternative to having full data from purebreds for genomic prediction in two- and three-breed rotational crossbreeding systems using a BOA model.

Materials & Methods

Simulated populations. Three base breed populations of 1,050 Danish Holstein (HOL), 1,050 Red Dairy Cattle (RDC), and 220 Danish Jersey (JER) were simulated based on real genotype data (Karaman *et al*., 2021). Only the first five first chromosomes (12,664 SNPs) were used in the simulation to reduce the computational demand. Among them, 500 SNPs were randomly chosen as QTL. Twelve generations were simulated by randomly mating 1,000 females to 50 males within the HOL and RDC population in each generation. Correspondingly in the JER population, 200 females were randomly mated to 20 males. The population sizes and gender distributions were kept constant in each generation. Additionally, two crossbred populations were simulated based on the purebred populations. A three-breed rotational crossbreeding system was used to produce a crossbred population consisting all three purebreds (MIX) by initially mating HOL females to JER males in the first generation, and then alternating between mating crossbred females to RDC, HOL, and JER males. A two-breed rotational crossbreeding system was used to produce a population consisting of crossbred animals between JER and HOL (JXH) starting by mating HOL females to JER males in the first generation, and then alternating between mating crossbred females to HOL and JER sires. Both crossbred populations were simulated for 12 generations resulting in 1,050 animals per generation and all animals being sired by a HOL male in the last generation. Thus, the average genomic breed proportions in generation 12 were 57.3% HOL, 28.2% RDC, and 14.5% JER for the MIX population, and 66.7% HOL and 33.3% JER for the JXH population.

Each breed population had breed-specific allele substitution at the 500 QTL. Computed from the genetic variances and co-variances in the base populations, the genetic correlations between the breeds were 0.87 between HOL and RDC, 0.62 between HOL and JER, and 0.68 between RDC and JER. Heritabilities of the simulated trait were around 0.4 for the three breeds.

The reference populations consisted of animals from generations 9, 10, and 11. Joining all pure breeds and crossbred populations in a combined reference gave a total of 13,260 animals. The average genomic breed proportion of the animals in the combined reference population were 42.2% HOL, 31.6% RDC, and 26.1% JER. The animals in the test populations were all from generation 12.

Breed-origin-of-alleles model. The SNP effects were estimated for each breed as follows:

$$\mathbf{y}=\mathbf{1}\mu+\mathbf{X}\mathbf{b}+\mathbf{M}_1\boldsymbol{\beta}_{HOL}+\mathbf{M}_2\boldsymbol{\beta}_{RDC}+\mathbf{M}_3\boldsymbol{\beta}_{JER}+\mathbf{e} \quad (1)$$

where **y** is the vector of phenotypes, **1** is a vector of 1s, μ is the overall mean, **X** is the matrix of (centred) estimated genomic breed proportions computed from genomic data, **b** is the vectors of fixed breed effects, and $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ are the column-centred matrices of breed-specific allele content of SNPs for HOL, RDC, and JER, respectively, $\boldsymbol{\beta}$s are vectors of SNP effects for HOL, RDC, and JER, respectively, and **e** is a vector of residuals. We used a Bayesian approach to estimate the dispersion and location parameters and assigned a normal distribution prior to the vectors of SNP effects for each breed: $\boldsymbol{\beta}_i|\sigma^2_{\beta i}\sim N(\mathbf{0},\mathbf{I}\sigma^2_{\beta i})$, where *i* stands for the breeds HOL, RDC and JER. Residuals were *a priori* assumed to follow a normal distribution, $\mathbf{e}\sim N(0,\mathbf{R})$ where **R** was a diagonal matrix of residual variances.

The GEBVs were predicted using SNP effects estimated from the combined reference population. In predicting GEBVs for crossbred animals, alleles of each SNP were traced back to their (known) breed of origin, and the SNP effect from the respective breeds. The GEBVs for pure breeds were predicted by multiplying SNP effects with the corresponding allele counts. Additionally, for prediction of GEBVs in crossbred animals, the fixed breed effects were multiplied by genomic breed proportions. Prediction accuracies were calculated as the correlation between the simulated genetic value and the predicted GEBVs.

Summary statistics. If only summary statistics were available for a certain breed, the prior distribution assigned to the SNP effects of the breed was reformulated using the results from a separate analysis of that individual breed. The mean and variance of the prior distribution of the SNP effects was set to the estimated SNP effect and its posterior variance from the separate analysis. The number of animals and their mean phenotypes was used to update the prior for the fixed breed effects.

Scenarios. We investigated different combinations of information sources, including full genotype and phenotype data, only summary statistics, or no data available from the pure breed populations (HOL, JER, and RDC) using the BOA approach. The crossbred populations (MIX and JXH) always had phenotype and genotype information available. All of the analyses were carried out in scripts written in Julia and were replicated 25 times. The prediction accuracies of the different scenarios were compared by a paired t-test using a significance level of $p<0.05$ within test populations.

Results

Within all test populations, using summary statistics from the pure breeds yielded slightly (0.021 – 0.046) lower prediction accuracies compared with using full genotype and phenotype data, while having no information from the pure breeds yielded much (0.086 – 0.224) lower prediction accuracies as shown in Table 1. Replacing full data with summary statistics from JER had no or little effect on the prediction accuracies for any of the test populations except JER. When only summary statistics or no information was available from either JER or RDC reference populations or both, the prediction accuracies decreased more for the RDC test population than JER.

Table 1. Prediction accuracies for Danish Holstein (HOL), Danish Jersey (JER), Swedish Red (RDC), three-breed rotational crosses (MIX) and JER x HOL rotational crosses (JXH) in different scenarios including full phenotype and genotype data (F), summary statistics (S), or no information (-).

Data from reference population					Test population				
HOL	RDC	JER	MIX	JXH	HOL	RDC	JER	MIX	JXH
F	F	F	F	F	0.798_{a}	0.761_{b}	0.743_{b}	0.753_{b}	0.789_{b}
S	S	S	F	F	0.752_{c}	0.731_{c}	0.710_{c}	0.720_{c}	0.768_{c}
F	S	S	F	F	0.795_{b}	0.554_{d}	0.673_{d}	0.718_{c}	0.782_{e}
F	S	F	F	F	0.795_{b}	0.553_{d}	0.748_{a}	0.718_{c}	0.785_{d}
F	F	S	F	F	0.799_{a}	0.762_{a}	0.713_{c}	0.755_{a}	0.791_{a}
-	-	-	F	F	0.605_{d}	0.537_{e}	0.657_{e}	0.590_{f}	0.676_{gh}
F	-	-	F	F	0.755_{c}	0.528_{f}	0.491_{g}	0.647_{e}	0.669_{h}
F	-	F	F	F	0.759_{c}	0.529_{f}	0.667_{de}	0.655_{d}	0.693_{f}
F	F	-	F	F	0.763_{c}	0.743_{c}	0.499_{f}	0.706_{c}	0.679_{g}

[a-g]Accuracies within test population without common subscripts differ significantly ($P < 0.05$)

Discussion

Our results show that all breeds represented in a crossbreeding system should be included in the reference population to obtain higher prediction accuracies. In the situation when full genotype and phenotype data is not available from all or some of the breeds, using summary statistics can yield prediction accuracies almost as high as when full data is available. Furthermore, using summary statistics to run prediction models does not require a large amount of data, which makes it more practically feasible.

The sizes of the crossbred reference populations in this study were as large as for HOL and RDC, which is probably optimistic since number of (genotyped) crossbred dairy cattle is small in most dairy cattle populations. Using summary statistics from JER (which had a smaller population size) had little or no effect on the prediction accuracies, compared with having full information from JER, because there were relatively more animals involved with JER genes in the crossbred reference populations than in the purebred JER reference. Furthermore, all animals in the MIX and JXH test populations were HOL-sired, making the information from the RDC and JER reference populations less important than HOL in predicting GEBVs for the crossbred animals. Assuming smaller crossbred reference populations with various breed combinations in the real-life situation, information from purebreds will be more important than indicated in this study.

A main reason for crossbreeding in dairy cattle is utilizing the benefits of heterosis effects (Sørensen *et al.,* 2008), which were ignored in the simulation of phenotypes and estimation of GEBVs in this study. Heterosis is caused by non-additive effects: dominance effects between alleles at the same loci and epistatic effects between alleles at different loci. Ignoring the non-additive effects may cause some biasedness of the genomic prediction (Su *et al.,* 2012). Therefore, if the current model should be applied to routine genomic breeding evaluation, non-additive effects should be considered in the model.

References

Clasen J.B., Bengtsson C., Kallström H.N., Strandberg E., Fikse W.F. *et al.* (2021) Animal 15(12):100409. https://doi.org/10.1016/j.animal.2021.100409

Clasen J.B., Fikse W.F., Kargo M., Rydhmer L., Strandberg E. *et al.* (2020) J Dairy Sci 103(1):514-528. https://doi.org/10.3168/jds.2019-16958

Karaman E., Su G., Croue I., and Lund M.S. (2021) Genet Sel Evol 53(1):46. https://doi.org/10.1186/s12711-021-00637-y

Magne M.A. and Quenon J. (2021) Agron Sustain Dev 41(2):1-15. https://doi.org/10.1007/s13593-021-00683-2

Maier R.M., Zhu Z., Lee S.H., Trzaskowski M., Ruderfer D.M *et al.* (2018) Nat Commun 9(1):989. https://doi.org/10.1038/s41467-017-02769-6

Su G., Christensen O.F., Ostersen T., Henryon M., and Lund M.S. (2012) PLoS One 7(9):e45293. https://doi.org/10.1371/journal.pone.0045293

Sørensen M.K., Norberg E., Pedersen J., and Christensen L.G. (2008) J Dairy Sci 91(11):4116-4128. https://doi.org/10.3168/jds.2008-1273

Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L. *et al.* (2011) PLoS One 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101

Thomasen J.R., Liu H., and Sørensen A.C. (2020) J Dairy Sci 103(1):597-606. https://doi.org/10.3168/jds.2019-16974

Vandenplas J., Calus M.P.L., and Gorjanc G. (2018) Genetics, 210(1):53-69. https://doi.org/10.1534/genetics.118.301109

Vandenplas, J., Calus, M.P.L., Sevillano C.A., Windig J.J., and Bastiaansen J.W.M. (2016) Genet Sel Evol 48:61. https://doi.org/10.1186/s12711-016-0240-y