Maximizing crossbred performance through purebred genomic selection

***H. Esfandyari*[*†], A. C. Sørensen[*] and P. Bijma[†]**

[*]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark, [†]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands

ABSTRACT: Genomic selection (GS) can be used to select purebreds for crossbred performance (CP). As dominance is the likely genetic basis of heterosis, explicitly including dominance in the GS model may be beneficial for selection of purebreds for CP, when estimating allelic effects from pure line data. The objective of this study was to investigate the benefits of GS of purebreds for CP based on purebred information, using simulation. Results demonstrate that when at least one of the purebreds involved in a two way crossbreeding system was selected for CP, crossbred offspring had better performance than when selection in both breeds was for purebred performance. Under the hypothesis that crossbreds differ from purebreds due to dominance, GS can be applied to select purebreds for CP without crossbred data, by using a dominance model.

Keywords: genomic selection; crossbreeding; dominance

Introduction

Numerous studies have shown encouraging results of applying genomic selection (GS) in purebred populations (Hayes et al. 2009). However, many animals used in livestock production systems are crossbreds using advantages of heterosis and breed complementarity. Different GS models have been proposed and used to select purebreds for crossbred performance (CP) (Dekkers 2007; Ibanez-Escriche et al. 2009) but in most studies, additive gene action or perfect knowledge of gene substitution effects or both have been assumed (Ibanez-Escriche et al. 2009; Toosi et al. 2010). As dominance is the likely genetic basis of heterosis, explicitly including dominance in the GS model may be beneficial for selection of purebreds for CP. With dominance, allele substitution effects and individual breeding values depend on allele frequency and thus change over time, which alters the ranking of individuals. This problem can be overcome by the use of a dominance model, which provides estimates of both additive and dominance effects and therefore enables the computation of allele substitution effects using appropriate allele frequencies (Zeng, et al. 2013). Zeng, et al. (2013) compared additive and dominance models for GS in purebred for CP and came to the conclusion that, when dominance is the key driver of heterosis, using a dominance model for GS is expected to result in greater cumulative response to selection of purebred animals for CP than the additive model.

Previous studies for selection of purebreds for CP (Ibanez-Escriche et al. 2009; Toosi et al. 2010; Zeng et al. 2013) have focused on crossbred data for estimating marker effects, which requires collection of crossbred genotypes and phenotypes. Collection of crossbred data can substantially increase the required investment in the breeding program, as crossbred animals are usually not individually identified and individual performance is not recorded. It is interesting to evaluate the potential benefit of GS within purebred lines when the objective is to improve performance of crossbreds, by using effects of markers estimated from pure line data. In other words, additive and dominance effects of alleles can be estimated from pure line data, and subsequently breeding values for CP can be estimated by using the appropriate allele frequencies. Thus, the objective of this study was to investigate the benefits of GS of purebreds for CP based on purebred information only. A second objective was to compare the use of having two separate pure line reference populations with combining both pure lines into a single reference population. Both objectives were investigated using simulation, for two cases of either low or high correlation of phase between both pure lines.

Materials and Methods

Population structure. Using QMSim (Sargolzaei and Schenkel 2009), 2000 discrete generations were simulated to establish a historical population (step 1). In the next step, to simulate the 2 purebred recent populations (breeds A and B hereafter), 2 random samples of 50 animals were drawn from the last generation of the historical population and each were randomly mated for another 100 generations. In the next step, in order to enlarge the breed A and B, eight generations were simulated with five offspring per dam (step 3). Within each breed, all animals in generation 8 of this step were considered as reference population. In the next step (step 4), for each breed 100 males and 200 females were sampled randomly from last generation of previous step and they were mated randomly to produce 1000 purebred animals (A0 and B0). In subsequent generations (step 5), a two way crossbreeding program with 5 generations of selection were simulated. The selection criterion in purebreds was the rank of the individual's genomic estimated breeding value. The SNP effects for the prediction of genomic estimated breeding value (GEBV) for each breed were estimated only once, using the purebred reference population of generation 8 of step 3. In generation 1 through 5 of step 5, 300 animals (top 100 males and top 200 females) were selected from the 1000 candidates in each parental breed based on their GEBV and were randomly mated within each breed to produce 1000 purebred replacement animals for the next generation. Meanwhile, the 100 selected males of breed A were mated randomly to the 200 selected females of breed B to produce 1000 crossbred progeny. The phenotypic mean of crossbreds was computed in each generation of selection ($AB_1 - AB_5$). The parameters for the genome and trait simulation are presented in Table 1.

Table 1 Parameters of the simulation process

Genome	
Number of chromosomes	1
Number of markers	1000
Marker distribution	Random
Number of QTL	100
QTL distribution	Random
MAF for markers and QTL	0.1
Additive allelic effects for markers	Neutral
Additive allelic effects for QTL (a_i)	Gamma
Dominance degree for QTL (h_i)	Normal (0.5,1)
Dominance effects for QTL (d_i)	$d_i = h_i.\|a_i\|$
Rate of recurrent mutation	2.5×10^{-5}
heritability	0.3
Dominance variance	0.1

True and genomic estimated breeding values Two types of true breeding values were calculated; True breeding value for purebred performance (TBVP) and true breeding value for crossbred performance (TBVC). For both cases, TBVs were calculated as the expected genotypic value of the offspring of a parent carrying a certain QTL-genotype, when this parent is mated at random to its own line (TBVP) or to the other pure line (TBVC). Thus, for animal i from breed r, the true breeding value for purebred performance was calculated as

$$\mathrm{TBVP_{ir}} = \sum_{j=1}^{100} [(x_{ij})(p_{jr}a_j + q_{jr}d_j)] + [(y_{ij})(0.5p_{jr}a_j + 0.5q_{jr}d_j + 0.5p_{jr}d_j - 0.5q_{jr}a_j)] + [(z_{ij})(-q_{jr}a_j + p_{jr}d_j)]$$

where x_{ij}, y_{ij} and z_{ij} are indicators function of the genotype of the j^{th} QTL of the i^{th} individual that $x_{ij} = 1$ when the genotype is AA otherwise is zero, $y_{ij} = 1$ when the genotype is Aa or aA otherwise is zero and $z_{ij} = 1$ when the genotype is aa otherwise is zero. Moreover, p_{jr} and q_{jr} are the allelic frequencies (A or a) for the j^{th} QTL in breed r, a_j and d_j are true additive and dominance effect of j^{th} QTL. For example, for an AA-parent at locus j, a fraction p_{jr} of its offspring will have genotype AA, while a fraction q_{jr} of its offspring will have genotype Aa. Hence, for locus j, the breeding value of this parent equals $(p_{jr}a_j + q_{jr}d_j)$, which is the first term in Equation 1. For crossbred offspring, the expected genotype frequencies of the offspring of a parent depend on the allele frequency in the other pure line. Thus, for animal i from breed r, the true breeding value for CP was calculated similar to TBVP except using allele frequency in the other pure line. Genomic estimated breeding values were calculated analogously, but now using marker genotypes rather than QTL genotypes, and estimated additive ($\hat{a}$) and dominance marker effects ($\hat{d}$) rather than true effects. For example, for animal i from breed A genomic estimated breeding value for CP (GEBVC) was calculated based on estimated marker effects obtained from breed A and using allele frequencies from breed B. The Bayesian LASSO was used to estimate marker effects using the "BLR" R package developed by Perez et al. (2010).

Table 2 Simulated scenarios

	Selection criterion		**Reference population structure**
	Breed A	Breed B	
Ref Sc.	GEBVP	GEBVP	Separate
Sc. 1	GEBVC	GEBVP	Separate
Sc. 2	GEBVC	GEBVC	Separate
Sc. 3	GEBVC	GEBVP	Common
Sc. 4	GEBVC	GEBVC	Common

GEBVP: genomic estimated breeding value for purebred performance. GEBVC: genomic estimated breeding value for crossbred performance.

Scenarios. The response to selection in CP was examined in five scenarios (Table 2). Simulated scenarios differed in structure of the reference population and also in the criterion of selection. In the reference scenario, both pure lines were selected for purebred performance, and both pure lines had their own reference population. Size of the reference population in scenarios with separate training for each breed was 1000 randomly selected animals and in scenarios with common reference population was 2000.

Finally, we compared our scenarios under two cases of low or high correlation of phase. In order to increase the correlation of phase between both breeds we increased LD in the common ancestral population of both breeds by decreasing effective population size. Sved et al. (2008) showed that if two populations diverge from a common ancestral population, their correlation of phase can be expressed as $r_0^2(1-c)^{2T}$, where r_0^2 is LD in the common ancestral population, c is the recombination rate between markers, and T is the time since breed divergence in generations.

Results

Response to selection in crossbreds. In the simulated data, the genetic correlation between purebred and crossbred performance ($r_{tbvp,tbvc}$) was 0.75 (±0.037) and 0.74 (±0.034) on average for low and high correlation of phase, respectively. Figure 1 shows mean phenotype of crossbreds in 5 generations under the five simulated scenarios in case of low and high correlation of phase between both breeds. When correlation of phase between both breeds was low, breeding for CP in at least one of both breeds leads to higher response in crossbreds. By generation 5, scenario 2 in which both breeds were selected for CP had higher mean in their crossbred offspring than other scenarios. Scenario 1 had higher response than reference scenario as in this scenario one of the breeds were selected for CP. Figure 1a

also shows that when both breeds had separate reference population (Sc. 1 and Sc. 2), their crossbred offspring had better performance than in alternative scenarios which had a common reference population (Sc. 3 and Sc. 4). Also when correlation of phase between both breeds was high, selection for CP improved the response in crossbreds. However, with high correlation of phase, using a combined reference population improved response (Sc. 3 & 4).

Discussion

Response to selection. One of the main results of this study was that selection for CP is better to improve crossbred progeny, also when EBVs for CP are based on purebred data. Selection on genotype at the purebred level reduces the need for crossbred testing that is required for combined crossbred and purebred selection, thereby saving important test resources. However, in scenarios with GEBVP as selection criterion , we still had substantial improvement in crossbreds. Obviously, correlation between GEBVP and TBVC will be the main factor defining improvement in crossbreds when selection is based on GEBVP within line. In our study this correlation ($r_{tbvc,gebvp}$) was 0.5 (±0.027) on average, suggesting that selection response for purebreds and crossbreds will have the same sign.

Non-additive effects and G×E interaction. In our simulation, we have assumed that additive and dominance effect of QTL alleles are similar in both breeds. However, this assumption is violated when there is QTL-by-environment interaction or QTL-by-genetic background interaction (epistasis). Such interactions would reduce response in CP. Also, in this study we focused on using purebred data to improve CP. However, Dekkers and Chakraborty (2004) discussed that benefit of GS for improving CP might be limited if marker effects are estimated from purebred nucleus data as resulting EBVs are strictly relevant only to the studied population and environment and may not help much to improve selection for CP if substantial G×E and genotype by genetic background interactions are present. In fact, incomplete correlation between purebreds and crossbreds can be due to both non-additive effects (dominance and epistasis), and G×E interaction. In this study we considered the G×E due to dominance, not that due to differences in the physical environment. In principle, one could use a dominance model and multitrait analysis to partition the purebred-crossbred genetic correlation into a component due to dominance and a remaining component due to G×E and epistasis. However, this would require very large data sets on CB (Bijma and Bastiaansen, these proceedings).

Conclusion

Under the hypothesis that crossbreds differ from purebreds due to dominance, GS can be applied to select purebreds for CP without crossbred data, by using a dominance model. Using simulation, we found that in a two way cross breeding system, response to selection in crossbreds was higher when selection was for GEBVC even though data were collected on purebreds. Furthermore, if correlation of phase between two breeds is high, there can be extra benefit in terms of accuracy if animals from both breeds are joined into a single reference population to estimate marker effects.

Literature Cited

Dekkers, J. C. M. (2007). J Anim. Sci, 85: 2104-2114.

Dekkers, J. C. M., R. Chakraborty (2004). *Genet Sel Evol*., 36: 297-324.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, et al. (2009). *J. Dairy Sci*., 92(3):1313

Ibanez-Escriche, N., R. L. Fernando, A. Toosi, et al. (2009). *Genet Sel Evol*., 41.

Perez, P., G. de los Campos, J. Crossa, et al. (2010). *Plant Genome* 3, 106-116.

Sargolzaei, M., F. S. Schenkel (2009). *Bioinformatics* 25, 680-681.

Sved, J. A., A. F. McRae, P.M. Visscher (2008). *A. J. Hum. Gene*. 83, 737-743.

Toosi, A., R. L. Fernando, J. C. M. Dekkers (2010). *J Anim. Sci*., 88: 32-46.

Zeng, J., A. Toosi, J. C. M. Dekkers, et al. (2013). *Genet Sel Evol*., 45(1): 11.

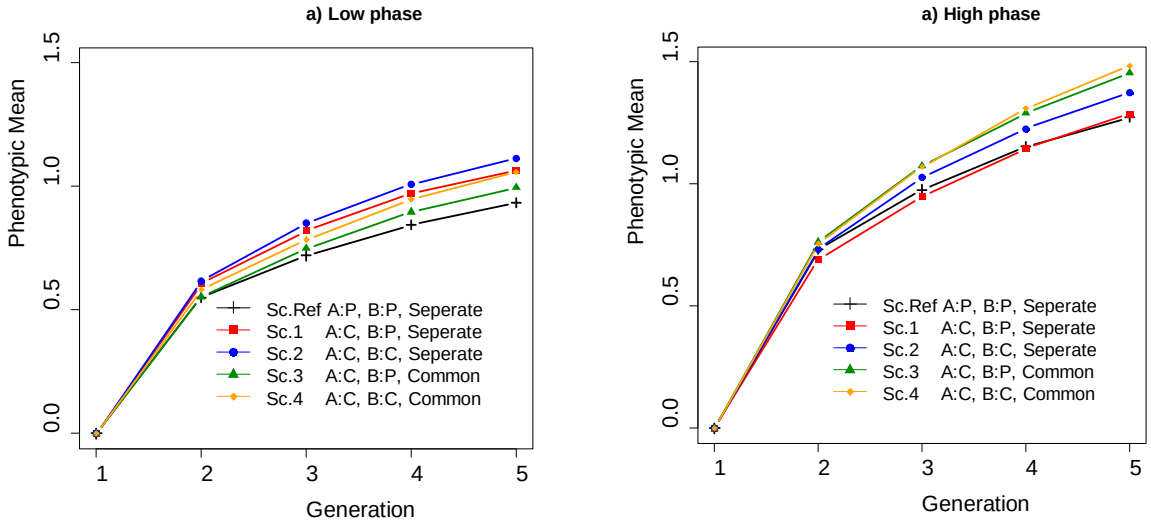

Figure 1 Mean Phenotype of crossbreds. (**a**) Results from low correlation of phase between breed A and B (r=0.2 for markers 1 cM apart) and (**b**) Results from high correlation of phase between breed A and B (r=0.7 for markers 1 cM apart). **A**: Breed A, **B**: Breed B, **P**: purebred performance, **C**: Crossbred performance. Standard Error of phenotypic mean for simulated scenarios in generation 5 ranged 0.02-0.03.