

Increasing Predictive Ability Using Dominance in Genomic Selection

C. Sun¹, P. M. VanRaden², J. B. Cole² and J. R. O'Connell³

¹National Association of Animal Breeders, Columbia, Missouri, ²Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, Maryland, ³University of Maryland School of Medicine, Baltimore, Maryland, USA

ABSTRACT: This study investigated the dominance for Holstein and Jersey milk, protein, and fat yields; somatic cell score, productive life, and daughter pregnancy rate. Additive and dominance variance components were estimated. Predictive abilities between three models with both additive and dominance effects (M_{AD1} , M_{AD2} and M_{AD3}) and a model with additive effect only (M_A) were assessed using 10-fold cross-validation. The M_{AD1} model estimated dominance values; M_{AD2} estimated dominance deviations with a different dominance relationship matrix. M_{AD3} enlarges dataset by including cows whose genotype probabilities were derived using genotyped ancestors. Dominance from M_{AD1} accounted for 5 and 7% of total variance for Holstein and Jersey yield traits, respectively. Heritability estimates were lower for dominance and higher for additive effects with M_{AD2} than with M_{AD1} . M_{AD1} and M_{AD2} increased prediction accuracy relative to the M_A model for yield traits. M_{AD3} model did not further improve prediction.

Keywords: prediction; dominance; genomic selection

Introduction

Genomic selection has become a standard tool in dairy cattle breeding. However, few studies have attempted to generalize and apply genomic selection models to include nonadditive genetic effects with large data sets (Calus, 2010; Da et al., 2014). Dominance variance usually is much less than additive variance, but it is also an important non-additive genetic effect. Inclusion of dominance effects in models to predict genomic breeding values could increase prediction accuracy (Su et al., 2012). However, genotypes and phenotypes for the same individuals are needed to detect allelic interaction. The increased availability of cows with phenotypes and genotypes in the United States now provides an opportunity to investigate models that include dominance effects.

Many cows with phenotypes do not have genotypes, but their sire and dam or their sire and maternal-grandsire have genotypes. Thus, genotype probabilities for a cow can be derived from her ancestors' genotypes and allele frequencies in the population. Boysen et al. (2013) discovered significant dominance effects for milk production traits of dairy cattle by regression of phenotypes on such derived genotype probabilities but did not test to determine if model prediction was improved if cows with derived genotype probabilities were included.

The objectives of this study were to 1) estimate additive and dominance variance components for 6 traits of Holsteins and Jerseys, 2) compare predictive ability be-

tween models including additive and dominance effects and a model including only additive effects, 3) compare predictions obtained using different dominance coefficients, and 4) test model prediction by expanding the data set to include cows with genotype probabilities that were based on ancestors' genotypes.

Materials and Methods

Data. Table 1 shows numbers of first-parity Holstein and Jersey records for yield traits (milk, fat, and protein), somatic cell score (SCS), productive life (PL), and daughter pregnancy rate (DPR), which were from the U.S. national database maintained by the Council on Dairy Cattle Breeding (CDCB; Reynoldsburg, Ohio). Phenotypes were deviations of a cow's record from population mean adjusted for fixed effects of age and parity group, herd management group, inbreeding, and heterosis. Fixed effects were estimated using a linear mixed model from the full national data set of phenotype and pedigree information.

Table 1. Numbers of first-parity records by data group and trait² for Holsteins and Jerseys

Breed	Data group				
	Yield	SCS	PL	DPR	
Holstein	D _C	30,482	30,352	14,780	23,811
	D _{SD}	25,926			
	D _{SMGS}	33,897			
		(2,278,652)			
Jersey	D _C	8,321	8,292	5,492	7,422
	D _{SD}	4,896			
	D _{SMGS}	11,823			
		(379,713)			

¹D_C = cows with known genotypes; D_{SD} = cows with genotypes imputed from sire and dam genotypes; D_{SMGS} = cows with genotypes imputed from sire and maternal grandsire genotypes (total number of daughters in D_{SMGS} groups in parentheses).

²Yield = milk, fat, and protein yields; SCS = somatic cell score; PL = productive life; DPR = daughter pregnancy rate.

Genotypes for Holstein and Jersey cows and their dams, sires, and maternal grandsires (MGS) were also available from the CDCB database. All genotypes were imputed to a BovineSNP50 basis using findhap.f90 software (VanRaden, 2011) before estimating genomic breeding values and dominance effects.

The data set was divided into 3 groups: 1) cows with known genotypes and phenotypes (D_C); 2) cows with phenotypes but genotype probabilities derived from genotyped sires and dams (D_{SD}); and 3) cows with phenotypes

but genotype probabilities derived from genotyped sires and MGS (D_{SMGS} ; Boysen et al., 2013).

Variance components. Variance components were estimated with a linear mixed model method and a marker-based relationship matrix among animals (GBLUP). The genomic relationship matrix (G) and dominance relationship matrix (D_1 or D_2) were calculated as (Da et al., 2014; Vitezica et al., 2013):

$$G = \frac{ZZ'}{\sum 2p_i q_i} \quad D_1 = \frac{HH'}{\sum 2p_i q_i (1 - 2p_i q_i)} \quad D_2 = \frac{MM'}{\sum (2p_i q_i)^2}$$

where Z is a centered genotype matrix with Z equal to a genotype code (0, 1, or 2) minus $2p_i$ (the frequency of the second of 2 alleles at locus i) and q_i is the frequency of the first of 2 alleles at locus i ; $H = -2p_i q_i$ for homozygous alleles or $1 - 2p_i q_i$ for heterozygous alleles; and $M = -2p_i^2$, $2p_i q_i$, or $-2q_i^2$ for genotype code 0, 1, or 2, respectively.

Three different models that included additive and dominance effects (M_{AD1} , M_{AD2} and M_{AD3}) and one model with only additive genetic effects (M_A) were used for analysis. The M_{AD1} model used G and D_1 , and the M_{AD2} model used G and D_2 . Data group D_C was analyzed with both the M_{AD1} and M_{AD2} models for all 6 traits; the complete data set (data groups $D_C + D_{SD} + D_{SMGS}$) was analyzed only with M_{AD3} and only for yield traits. The additive and dominance relationship matrix for M_{AD3} were calculated in the same way used in G and D_1 .

SNP effects. The additive or dominance effects for each SNP were estimated using random regression on marker genotypes, and genomic breeding values were calculated as the sum of estimated marker effects (SNP-BLUP; VanRaden, 2008). Solutions for small populations can be obtained directly by building mixed model equations and inverting the left-hand side. For data group D_C , the SNP equations were solved by the inversion method. However,

for the complete data set, genotype probabilities for some cows required >1 character for storage; therefore, calculations for the left-hand side of the mixed model equations required much more computing time, memory, or disk space. An iteration-based program was developed to solve for SNP additive and dominance effects for the complete data set. A blend of first- and second-order Jacobi iteration was implemented with 2 relaxation factors (VanRaden, 2008).

Model validation. Cross-validation was used to measure prediction accuracy with the data set randomly divided into 10 approximately equal subsets. Nine of the subsets were used in turn as training sets to estimate SNP effects; the remaining subset was used for testing prediction accuracy. The predictive ability of the model was evaluated by comparing predictions and phenotypes of animals in the testing data set and was measured as the correlation between predicted genetic values and phenotypes. Predictions of additive genetic effect (breeding value) and total genetic value (defined as the sum of additive and dominance effects in the model) were both evaluated.

Results and Discussion

Variations. The h^2 estimates are in Table 2. For both Holsteins and Jerseys, M_{AD1} has lower additive heritabilities and higher dominance heritabilities compared to M_{AD2} , but the sum of additive and dominance variances were almost the same for M_{AD1} and M_{AD2} . The additive heritabilities from M_{AD2} were much closer to those from M_A . Based on M_{AD1} and M_{AD2} , the dominance variance accounted for 5% and approximately 4% of the phenotypic variance for Holstein production traits, and 7 and 5% of Jersey production traits. The additive heritability estimates from M_{AD3} were lower than those from M_{AD1} and M_{AD2} . M_{AD3} gave similar dominance heritabilities with M_{AD2} for Jersey, but smaller estimates for Holstein. Ertl et al (2013) reported higher dominance variances using 2000 Jersey cows. Dominance variances were very small for DPR and SCS regardless of breed, especially for DPR. The percent of dominance variance for PL was higher for Jersey than Holstein. Few other

Table 2. Heritability (h^2) estimates¹ for additive (a) and dominance (d) effects by model,² data group,³ and trait for Holsteins and Jerseys

Breed	Model	Data groups	h^2	Milk yield	Fat yield	Protein yield	Somatic cell score	Productive life	Daughter pregnancy rate
Holstein	M_A	D_C	a	28.8	25.3	22.1	8.7	4.3	5.6
	M_{AD1}	D_C	a	27.0	23.3	20.2	8.4	4.2	5.7
			d	5.1	5.1	5.3	1.0	0.0	0.0
	M_{AD2}	D_C	a	28.5	25.0	21.7	8.7	4.2	5.6
			d	3.7	3.4	3.9	1.0	0.5	0.0
	M_{AD3}	$D_C + D_{SD} + D_{SMGS}$	a	21.5	20.2	18.6			
d			2.4	2.4	2.5				
Jersey	M_A	D_C	a	35.2	22.2	25.8	10.2	7.1	3.4
	M_{AD1}	D_C	a	32.2	19.2	23.0	9.8	5.7	3.0
			d	7.0	7.2	7.0	1.2	3.8	1.2
	M_{AD2}	D_C	a	34.4	21.4	25.1	10.2	7.0	3.4
			d	5.4	5.5	5.6	1.0	2.4	0.04
	M_{AD3}	$D_C + D_{SD} + D_{SMGS}$	a	27.1	18.2	20.6			
d			5.2	5.8	5.4				

¹Variance of effect expressed as percentage of total phenotypic variance. ² M_A = only additive genetic effect included; M_{AD1} = additive and dominance effects included with dominance estimated as a dominance value; M_{AD2} = additive and dominance effects included with dominance estimated as a dominance deviation. ³ D_C = cows with known genotypes; D_{SD} = cows with genotypes imputed from sire and dam genotypes; D_{SMGS} = cows with genotypes

studies have estimated dominance variance using genomic data in Holstein. Additive and non-additive variances usually were estimated using models with pedigree-based relationship matrices. Van Tassell et al. (2000) estimated additive and dominance variance using Method R and reported 5 and 1% dominance variance for yield traits and SCS, respectively, which is consistent with the findings of this study.

Prediction accuracy. The M_{AD1} and M_{AD2} models had better predictive ability than the M_A model based on correlations from the 10-fold cross-validation (Table 3). The correlations between phenotype and total genetic effects (additive plus dominance) from M_{AD1} and M_{AD2} were higher than those between phenotype and additive-only effects from M_{AD1} and M_{AD2} for three production traits, and both were higher than the correlations between phenotype and additive effect from M_A . For Holstein SCS, M_{AD1} and M_{AD2} also were better predictors than the M_A model. However, for Jersey SCS, PL, and DPR and as well as Holstein PL and DPR, correlations were almost the same regardless of model. Jersey correlations were lower than those for Holsteins except for PL. By enlarging the dataset, M_{AD3} did not provide better prediction for either Holsteins or Jerseys primarily because additive heritability was underestimated, perhaps because of the more complex model needed to deal with the combined data. A better model might treat the 3 groups as correlated phenotypes. Because cows with imputed genotype probabilities were progeny of elite animals, preselection may have caused bias.

In addition to increased prediction accuracy, an additive and nonadditive genetic model could be beneficial for exploiting specific combining ability. Breeders should continue selecting for additive merit but can also improve nonadditive merit by considering interactions in mating programs (VanRaden, 2006). Sun et al. (2013) compared mating programs and found that expected progeny milk from linear programming using genomic relationship matrices increased by including dominance effects.

Conclusion

Dominance variances accounted for about 5 and 7% of total variance for Holstein and Jersey yield traits, respectively, based on the M_{AD1} model. For SCS, PL, and DPR, dominance variance was very low, especially for Holsteins. The M_{AD1} model had lower additive and higher dominance variance estimates than the M_{AD2} model. The M_{AD1} and M_{AD2} models can increase prediction ability for Holstein and Jersey yield traits based on 10-fold cross-validation; improvement from the 2 models was similar. Prediction accuracy did not improve by including 2.3 million cows with genotype probabilities derived from ancestors.

Acknowledgement

The authors thank the Council on Dairy Cattle Breeding (CDCB) of United States for providing genotype and phenotype data.

Literature Cited

- Boysen, T.-J., Heuer, C., Tetens, J. et al. (2013). *Genetics* 193:431–442.
- Calus, M. P. L. (2010). *Animal* 4:157–164.
- Da, Y., Wang, C., Wang, S. et al. (2014). *PLoS ONE* 9: e87666.
- Ertl, J., Legarra, A., Vitezica, Z. G. et al. (2013). *INTERBULL BULLETIN* NO. 47
- Su, G., Christensen, O. F., Ostensen, T. et al. (2012). *PLoS ONE* 7:e45293.
- Sun, C., VanRaden, P. M., O’Connell, J. R. et al. (2013). *J. Dairy Sci.* 96:8014–8023.
- Van Tassell, C. P., Misztal, I., and Varona, L. (2000). *J. Dairy Sci.* 83:1873–1877.
- VanRaden, P. (2011). <http://aipl.arsusda.gov/software/findhap/> Accessed Feb. 14, 2014.
- VanRaden, P. M. (2006). *Proc. 8th World Congr. Genet. Appl. Livest. Prod., Commun.* 01-39.
- VanRaden, P. M. (2008). *J. Dairy Sci.* 91:4414–4423.
- VanRaden, P. M., O’Connell, J. R., Wiggans, G. R. et al. (2011). *Genet. Sel. Evol.* 43:10.
- Vitezica, Z. G., Varona, L., and Legarra, A. (2013). *Genetics*. 195:1223-1230

Table 3. Mean correlations of estimated additive genetic (a) and dominance (d) effects with phenotype from 10-fold cross-validation by model¹, data group,² and trait

Breed	Trait	D_C					$D_C + D_{SD} + D_{SMGS}$	
		M_A	M_{AD1}		M_{AD2}		M_{AD3}	
		a	a	a + d	a	a + d	a	a + d
Holstein	Milk yield	0.440	0.452	0.458	0.451	0.460	0.433	0.441
	Fat yield	0.409	0.419	0.425	0.419	0.426	0.388	0.396
	Protein yield	0.399	0.405	0.412	0.405	0.415	0.375	0.385
	Somatic cell score	0.198	0.202	0.203	0.202	0.203		
	Productive life	0.108	0.108	0.108	0.108	0.108		
	Daughter pregnancy rate	0.158	0.159	0.159	0.159	0.159		
Jersey	Milk yield	0.419	0.428	0.441	0.428	0.434	0.392	0.427
	Fat yield	0.356	0.361	0.371	0.362	0.368	0.340	0.358
	Protein yield	0.356	0.361	0.373	0.361	0.369	0.324	0.350
	Somatic cell score	0.170	0.169	0.169	0.169	0.169		
	Productive life	0.107	0.109	0.115	0.110	0.109		
	Daughter pregnancy rate	0.092	0.091	0.092	0.092	0.092		

¹ M_A = only additive genetic effect included; M_{AD1} = additive and dominance effects included, with dominance estimated as a dominance value; M_{AD2} = additive and dominance effects included, with dominance estimated as a dominance deviation. ² D_C = cows with known genotypes; D_{SD} = cows with genotypes imputed from sire and dam genotypes; D_{SMGS} = cows with genotypes imputed from sire and maternal grandsire genotypes.