2017

The impact of training strategies on the accuracy of genomic predictors in United States Red Angus cattle

J. Lee
University of Nebraska-Lincoln

Stephen D. Kachman
University of Nebraska-Lincoln, steve.kachman@unl.edu

Matthew L. Spangler
University of Nebraska-Lincoln, mspangler2@unl.edu

The impact of training strategies on the accuracy of genomic predictors in United States Red Angus cattle[1]

J. Lee,* S. D. Kachman,† and M. L. Spangler*[2]

*Department of Animal Science, University of Nebraska, Lincoln 68583; and
†Department of Statistics, University of Nebraska, Lincoln 68583

ABSTRACT: Genomic selection (GS) has become an integral part of genetic evaluation methodology and has been applied to all major livestock species, including beef and dairy cattle, pigs, and chickens. Significant contributions in increased accuracy of selection decisions have been clearly illustrated in dairy cattle after practical application of GS. In the majority of U.S. beef cattle breeds, similar efforts have also been made to increase the accuracy of genetic merit estimates through the inclusion of genomic information into routine genetic evaluations using a variety of methods. However, prediction accuracies can vary relative to panel density, the number of folds used for folds cross-validation, and the choice of dependent variables (e.g., EBV, deregressed EBV, adjusted phenotypes). The aim of this study was to evaluate the accuracy of genomic predictors for Red Angus beef cattle with different strategies used in training and evaluation. The reference population consisted of 9,776 Red Angus animals whose genotypes were imputed to 2 medium-density panels consisting of over 50,000 (50K) and approximately 80,000 (80K) SNP. Using the imputed panels, we determined the influence of marker density, exclusion (deregressed EPD adjusting for parental information [DEPD-PA]) or inclusion (deregressed EPD without adjusting for parental information [DEPD]) of parental information in the deregressed EPD used as the dependent variable, and the number of clusters used to partition training animals (3, 5, or 10). A BayesC model with π set to 0.99 was used to predict molecular breeding values (MBV) for 13 traits for which EPD existed. The prediction accuracies were measured as genetic correlations between MBV and weighted deregressed EPD. The average accuracies across all traits were 0.540 and 0.552 when using the 50K and 80K SNP panels, respectively, and 0.538, 0.541, and 0.561 when using 3, 5, and 10 folds, respectively, for cross-validation. Using DEPD-PA as the response variable resulted in higher accuracies of MBV than those obtained by DEPD for growth and carcass traits. When DEPD were used as the response variable, accuracies were greater for threshold traits and those that are sex limited, likely due to the fact that these traits suffer from a lack of information content and excluding animals in training with only parental information substantially decreases the training population size. It is recommended that the contribution of parental average to deregressed EPD should be removed in the construction of genomic prediction equations. The difference in terms of prediction accuracies between the 2 SNP panels or the number of folds compared herein was negligible.

Key words: beef cattle, cross-validation, deregressed estimated progeny difference, genomic prediction

 J. Anim. Sci. 2017.95:3406–3414
doi:10.2527/jas2017.1604

INTRODUCTION

Genomic prediction using thousands of SNP has been applied in various cattle breeds after genomic selection (**GS**) was first introduced by Meuwissen et al. (2001). In practice, GS has been shown to increase the accuracy of EPD and has become widespread in both the dairy and beef cattle industries. Since its inception, numerous GS models, validation approaches, and SNP panels have been proposed to further improve the accuracy of genomic prediction.

The benefit of implementing GS depends, in part, on the accuracy of molecular breeding values (**MBV**). The accuracy of MBV can be affected by several fac-

[1]This work was supported by the Red Angus Association of America and GeneSeek, a Neogen company.

[2]Corresponding author: mspangler2@unl.edu

Received April 1, 2017.
Accepted June 8, 2017.

tors related to information content and model choice, including size of the training population and marker density (Daetwyler et al., 2008; Goddard, 2009; Su et al., 2012; Haiber, 2013), method of clustering animals for cross-validation and the relationship between the reference and validation populations (Habier et al., 2007, 2010; Saatchi et al., 2011), the choice of response variables (e.g., EPD or deregressed EPD without adjusting for parental information [**DEPD**]; Ostersen et al., 2011; Gunia et al., 2014), and the use of a weighting factor (or residual polygenic component) in the model (Calus and Veerkamp, 2007; Garrick et al., 2009). The impact of GS is also dependent on the genetic architecture of the trait (Hayes et al., 2010; Gunia et al., 2014; Gao et al., 2015) and the structure of the population (Goddard and Hayes, 2009).

The majority of GS strategies and statistical method comparisons relative to cattle have been performed using medium- (e.g., over 50,000 SNP [50K]) and high-density (e.g., over 770,000 SNP [**777K**]) SNP panels. Currently, a multitude of SNP panels are available. Consequently, the objectives of the current study were to evaluate the effect of 1) DEPD with and without (deregressed EPD adjusting for parental information [**DEPD-PA**]) mid-parent average information as a response variable in training, 2) the impact of 2 marker densities commonly used in beef cattle populations in the United States (50K and approximately 80,000 SNP [**80K**]), and 3) the number of clusters for cross-validation on the accuracy of genomic prediction in U.S. Red Angus beef cattle.

MATERIALS AND METHODS

Animal Care and Use Committee approval for this study was not obtained given that the data were obtained from existing databases.

Imputation and Data Editing

A total of 9,776 Red Angus beef cattle were genotyped using 5 different SNP panels (Table 1): BovineSNP50 version 2 (Illumina, Inc., San Diego, CA) and GGPLD version 1, GGPLD version 2, GGPHD, and GGPHD version 2 (Neogen Agrigenomics, Lincoln, NE). These SNP panels were imputed to 2 medium-density SNP panels (BovineSNP50 version 2 and GGPHD) consisting of 50K and 80K SNP markers, respectively, with the following 2 steps. First, Beagle version 3.3.2 (Browning and Browning, 2007) was used for phasing of 2 SNP panels (50K and 80K) to be used for the reference panel. Then, FImpute version 2.2 (Sargolzaei et al., 2014) was used to impute from the various SNP panels to both of the reference panels (50K or 80K). After excluding unmapped SNP and SNP on sex chromosomes, the available number of SNP markers for the 50K and 80K panels were 52,184 and 76,681, respectively. Duplicate animals (n = 396) and animals with a recorded sex in the pedigree file that did not match the sex determined by genotypes on the X chromosome (n = 230) were removed. Duplicate animals occurred due to regenotyping to generate acceptable marker call rates. Consequently, no animals were removed based on marker call rates. Furthermore, genotype identification that could not be matched to a corresponding animal in the pedigree file was removed, leaving 7,652 animals for further analysis.

Table 1. Number of animals genotyped by SNP panel and number of SNP by panel

SNP panel version	SNP	No. of animals
GGPLD version 1[1]	32,185	5,759
GGPLD version 2[1]	32,531	707
GGPHD[1]	77,376	2,671
GGPHD version 2[1] (UHD[2])	140,113	322
BovineSNP50 version 2[3]	54,069	317
All animals		9,776

[1]Neogen Agrigenomics, Lincoln, NE.

[2]UHD = Ultra high density.

[3]Illumina, Inc., San Diego, CA.

Response Variables

Expected progeny differences and corresponding Beef Improvement Federation accuracies (BIF, 2010) for the genotyped animals and their sires and dams were obtained from the Red Angus Association of America for 13 traits including back fat thickness (**BFAT**), birth weight (**BWT**), calving ease direct (**CE**), calving ease maternal (**CETM**), carcass weight (**CWT**), marbling score (**MARB**), maternal milk ability (**MILK**), rib eye muscle area (**REA**), stayability (**STAY**), weaning weight (**WWT**), yearling weight (**YWT**), heifer pregnancy (**HPG**), and maintenance energy (**MEN**). Beef Improvement Federation accuracies were transformed to reliabilities and EPD were deregressed following the methods of Garrick et al. (2009). The proportion of genetic variance not accounted for by the markers, c, was assumed to be equal to 0.40 (Saatchi et al., 2012). Deregressed EPD adjusting for parental information were formed by a combination of deregression (dividing by the reliability of EPD) after adjusting for ancestral information (i.e., parental average) such that the information content of the EPD contained only their own phenotypic information and that of their descendants. Deregressed EPD without adjusting for parental information were formed by deregression (dividing by the reliability of EPD) without adjusting for parental information. To

Table 2. Summary statistics of expected progeny differences and the response variables used for analysis[1]

Type of record	Parameter	Trait[2]												
		BFAT	BWT	CE	CETM	CWT	HPG	MARB	MEN	MILK	REA	STAY	WWT	YWT
EPD	No.	7,378	7,547	7,597	7,591	7,497	3,579	7,292	2,086	7,530	7,228	2,671	7,314	6,905
	Mean (SD)	0.01 (0.03)	−1.61 (2.09)	0.02 (0.10)	−0.01 (0.07)	25.76 (12.70)	0.03 (0.08)	0.59 (0.23)	1.60 (5.29)	21.27 (4.49)	0.20 (0.22)	0.24 (0.09)	62.35 (10.81)	97.88 (20.52)
	Minimum	−0.10	−9.90	−0.36	−0.35	−42.10	−0.29	−0.38	−18.48	−2.70	−0.58	−0.03	11.80	−5.30
	Maximum	0.16	8.60	0.43	0.31	72.80	0.35	1.60	24.45	49.60	1.24	0.47	100.20	166.70
	Mean BIF[3] accuracy	0.24	0.33	0.21	0.19	0.26	0.13	0.19	0.25	0.22	0.17	0.16	0.30	0.30
DEPD-PA[4]	No.	4,141	5,706	1,142	1,240	5,702	227	4,452	505	3,653	4,420	755	5,639	5,312
	Mean (SD)	0.01 (0.09)	−1.65 (4.86)	0.02 (0.29)	−0.01 (0.18)	29.26 (37.81)	0.03 (0.23)	0.66 (0.93)	1.44 (9.78)	20.38 (12.66)	0.27 (0.99)	0.38 (0.41)	65.09 (27.35)	100.70 (43.49)
	Minimum	−0.65	−45.59	−2.35	−1.10	−202.71	−1.30	−6.02	−108.16	−69.45	−3.68	−0.60	−210.44	−289.35
	Maximum	0.41	32.38	3.01	1.19	210.70	0.66	5.99	29.50	123.94	5.97	1.78	175.27	346.21
	Mean reliability	0.52	0.58	0.59	0.55	0.47	0.61	0.39	0.73	0.41	0.35	0.44	0.53	0.54
DEPD[5]	No.	7,378	7,547	7,597	7,591	7,479	3,579	7,292	2,086	7,530	7,228	2,671	7,314	6,905
	Mean (SD)	0.02 (0.07)	−3.01 (3.94)	0.06 (0.32)	−0.03 (0.23)	61.37 (33.16)	0.14 (0.38)	1.92 (0.89)	3.83 (17.05)	59.33 (20.33)	0.71 (0.79)	1.07 (0.64)	132.42 (40.12)	205.55 (66.20)
	Minimum	−0.33	−17.96	−1.50	−168.42	−1.33	−121.65	−0.67	−121.65	−2.79	−3.26	−0.06	18.50	−11.57
	Maximum	0.37	17.58	1.62	203.61	1.22	65.53	6.62	65.53	221.65	4.52	3.32	502.58	674.40
	Mean reliability	0.41	0.54	0.36	0.34	0.44	0.24	0.33	0.41	0.38	0.30	0.28	0.49	0.50

[1]Summary statics are after the removal of animals with reliability less than 0.1.

[2]BFAT = back fat thickness; BWT = birth weight; CE = calving ease direct; CETM = calving ease maternal; CWT = carcass weight; HPG = heifer pregnancy; MARB = marbling score; MEN = maintenance energy; MILK = maternal milk ability; REA = rib eye muscle area; STAY = stayability; WWT = weaning weight; YWT = yearling weight.

[3]Beef Improvement Federation (2010) accuracy.

[4]DEPD-PA = deregressed EPD adjusting for parental information.

[5]DEPD = deregressed EPD without adjusting for parental information.

ensure the quality of deregressed EPD (DEPD-PA and DEPD), animals with a reliability less than 0.10 were removed. After filtering, 7,652 registered Red Angus animals with deregressed EPD and genotypes were available for further analysis. Descriptive statistics of EPD and the 2 response variables (DEPD-PA and DEPD) for these animals after removing animals with a reliability less than 0.1 are reported in Table 2.

K-*Means Clustering*

K-means cross-validation (Hartigan and Wong, 1979) using 3, 5, or 10 folds was used to partition animals for the purpose of reducing the relationships between training and evaluation populations. Animals were assigned to folds based on the numerator relationship matrix such that relationships within a fold were maximized and relationships between folds were minimized. The distance matrix between genotyped animals was computed from elements of the relationship matrix as $d_{ij} = 1 - \{a_{ij}/[(a_{ii}a_{jj})^{1/2}]\}$ as described by Saatchi et al. (2011), in which d_{ij} is a measure of pedigree distance between individual i and individual j, a_{ij} is the additive genetic relationship between individual i and individual j, and a_{ii} and a_{jj} are diagonal elements of the **A** matrix. We computed the relationship matrix for 7,652 genotyped animals using a pedigree of 44,570 animals and then computed a distance matrix between the genotyped animals as described above. Given that *K*-means clustering is defined by a data matrix, the distance matrix was used and treated as a data matrix for the purposes of partitioning animals using *K*-means clustering following Saatchi et al. (2011). The number of individuals within each fold and within and between fold averages of $a_{\max}$ and a_{ij} and their SD are presented in Table 3.

Estimation of SNP Effects

A BayesC model (Kizilkaya et al., 2010) with π set to 0.99 was used to estimate SNP effects using GenSel4R software (Garrick and Fernando, 2013). BayesC was chosen because Habier et al. (2011) reported that BayesC is less sensitive to prior assumptions than BayesB (Meuwissen et al., 2001) and the majority of U.S. beef associations have used prediction equations constructed using BayesC with fixed π set to be 0.99 (Kachman et al., 2013). For each trait, the following model was fitted to estimate marker effects:

Table 3. Comparison of relationship among animals within and across clusters in 3-, 5-, and 10-fold cross-validations

Clusters	No. of clusters	No. of animals	a_{max_within}[1]	$a_{max\_between}$[2]	a_{max_ratio}[3]	a_{ij_within}[4]	$a_{ij\_between}$[5]	a_{ij_ratio}[6]
3 fold	1	2,615	0.371	0.279	1.33	0.037	0.027	1.37
	2	2,459	0.414	0.306	1.35	0.061	0.032	1.91
	3	2,578	0.375	0.290	1.29	0.029	0.027	1.07
Average		–	0.387	0.292	1.33	0.043	0.029	1.48
5 fold	1	1,574	0.382	0.334	1.14	0.047	0.032	1.47
	2	1,487	0.351	0.315	1.11	0.038	0.026	1.46
	3	1,582	0.349	0.219	1.59	0.022	0.013	1.69
	4	1,549	0.399	0.307	1.30	0.068	0.035	1.94
	5	1,460	0.444	0.290	1.53	0.090	0.035	2.57
Average		–	0.385	0.293	1.31	0.053	0.029	1.83
10 fold	1	594	0.423	0.285	1.48	0.096	0.030	3.20
	2	655	0.454	0.322	1.41	0.137	0.044	3.11
	3	779	0.359	0.289	1.24	0.087	0.033	2.64
	4	839	0.423	0.303	1.40	0.124	0.035	3.54
	5	591	0.358	0.391	0.92	0.085	0.041	2.07
	6	617	0.442	0.293	1.51	0.127	0.037	3.43
	7	653	0.397	0.236	1.68	0.085	0.016	5.31
	8	929	0.306	0.235	1.30	0.011	0.010	1.10
	9	846	0.357	0.352	1.01	0.048	0.031	1.55
	10	1,149	0.335	0.326	1.03	0.026	0.020	1.30
Average		–	0.385	0.303	1.27	0.084	0.030	2.80

[1]a_{max_within} = the average of a_{max} (the maximum value of relationships [a_{ij}] for each animal) values within cluster 1.
[2]$a_{max\_between}$ = the average of a_{max} values between the clustered (training and testing) groups.
[3]a_{max_ratio} = the ratio between a_{max_within} and $a_{max\_between}$.
[4]a_{ij_within} = the average of a_{ij} (relationships) values within cluster 1.
[5]$a_{ij\_between}$ = the average of a_{ij} values between the clustered groups.
[6]a_{ij_ratio} = the ratio between a_{ij_within} and $a_{ij\_between}$.

$$y_i = \mu + \sum_{j=1}^{k} Z_{ij} u_j \delta_j + e_i,$$

in which y_i is DEPD-PA or DEPD for animal i for the respective trait; μ is the population mean; k is the number of markers; Z_{ij} is allelic state at SNP j in individual i; and u_j is the random substitution effect for marker j, which is conditional on σ_u^2, which was assumed to be normally distributed $N(0, \sigma_u^2)$. A mixture of 2 distributions for the random substitution effect was assumed according to the indicator variable (δ_j), in which δ_j is a Bernoulli variable indicating the absence or presence of marker j in the model, and e_i is a random residual effect assumed normally distributed $N(0, \sigma_e^2)$. The total number of Markov chain Monte Carlo samples was 41,000, of which the first 1,000 were discarded as burn-in. Molecular breeding values were calculated as the sum of marker effects weighted by the SNP content.

The Accuracy of Molecular Breeding Values

A bivariate animal model was used that included MBV of genotyped animals estimated using genotypes from animals not in that animal's fold and weighted deregressed EPD to estimate genetic correlations using ASReml version 4.1 software (Gilmour et al., 2015). The 2-generation pedigree included 14,329 animals, and the same pedigree was used for each bivariate analysis. The model for MBV included fixed effects for the intercept and fold, random common and fold-specific additive genetic effects, and a residual with variance fixed at 0.0001% of the unweighted phenotypic variance of the deregressed EPD. The model for the deregressed EPD included a fixed effect for the intercept, a random additive genetic effect, and a weighted random residual with var(e) = $W\sigma_e^2$, in which W is the r-inverse weights according to the reliabilities of animal's DEPD, which were the same as used in training for the estimation of SNP effects. The additive genetic and unweighted residual variances were fixed at 0.4 and 0.6, respectively, of the deregressed unweighted phenotypic variance of the EPD. Resulting genetic correlations were the genetic correlations between the

Table 4. Genetic correlations between molecular breeding values and deregressed EPD (deregressed EPD adjusting for parental information [DEPD-PA] and deregressed EPD without adjusting for parental information [DEPD]) and their SE in U.S. Red Angus beef cattle across the studied traits for the panel consisting of over 50,000 SNP

	DEPD-PA								DEPD							
		3 fold		5 fold		10 fold				3 fold		5 fold		10 fold		
Trait[1]	No.	r_g[2]	SE	r_g	SE	r_g	SE	Average	No.	r_g	SE	r_g	SE	r_g	SE	Average
BFAT	4,141	0.420	0.027	0.455	0.027	0.453	0.027	0.443	7,378	0.438	0.027	0.424	0.027	0.429	0.027	0.430
BWT	5,706	0.683	0.018	0.693	0.018	0.706	0.018	0.694	7,547	0.648	0.019	0.667	0.019	0.697	0.019	0.671
CE	1,142	0.619	0.045	0.663	0.044	0.633	0.047	0.638	7,597	0.701	0.040	0.766	0.037	0.784	0.038	0.750
CETM	1,240	0.601	0.045	0.632	0.045	0.615	0.046	0.616	7,591	0.602	0.023	0.618	0.023	0.649	0.035	0.623
CWT	5,702	0.773	0.019	0.830	0.017	0.842	0.017	0.815	7,479	0.798	0.036	0.818	0.035	0.824	0.023	0.813
HPG	227	0.300	0.107	0.246	0.110	0.353	0.110	0.300	3,579	0.692	0.080	0.643	0.082	0.671	0.083	0.669
MEN	505	0.515	0.056	0.445	0.060	0.508	0.059	0.489	2,086	0.537	0.053	0.514	0.057	0.530	0.057	0.527
MARB	4,452	0.503	0.032	0.472	0.033	0.494	0.033	0.489	7,292	0.425	0.033	0.429	0.033	0.437	0.033	0.430
MILK	3,653	0.402	0.034	0.393	0.034	0.403	0.034	0.399	7,530	0.261	0.032	0.254	0.032	0.271	0.032	0.262
REA	4,420	0.601	0.033	0.589	0.033	0.625	0.033	0.605	7,228	0.518	0.034	0.510	0.034	0.536	0.034	0.521
STAY	755	0.348	0.065	0.396	0.067	0.462	0.066	0.402	2,671	0.378	0.063	0.259	0.070	0.307	0.071	0.315
WWT	5,639	0.664	0.021	0.711	0.020	0.701	0.021	0.692	7,314	0.346	0.025	0.358	0.025	0.368	0.025	0.357
YWT	5,312	0.664	0.021	0.682	0.021	0.691	0.021	0.679	6,905	0.410	0.025	0.412	0.025	0.427	0.025	0.416
Average		0.546		0.554		0.576		0.559		0.520		0.513		0.533		0.522

[1]BFAT = back fat thickness; BWT = birth weight; CE = calving ease direct; CETM = calving ease maternal; CWT = carcass weight; HPG = heifer pregnancy; MARB = marbling score; MEN = maintenance energy; MILK = maternal milk ability; REA = rib eye muscle area; STAY = stayability; WWT = weaning weight; YWT = yearling weight.

[2]r_g = genetic correlation.

common additive genetic effect of the MBV and the additive genetic effect of the deregressed EPD.

RESULTS

Single Nucleotide Polymorphism Density

The accuracies of MBV for the 50K and 80K are presented in Tables 4 and 5, respectively. Comparing SNP panels, the accuracies of MBV using the 50K ranged from 0.246 (HPG) to 0.842 (CWT) and from 0.261 (MILK) to 0.824 (CWT) for DEPD-PA and DEPD, respectively, across all variations in the number of folds used for cross-validation. A similar trend was observed when using the 80K, with a range of accuracies from 0.270 (HPG) to 0.836 (CWT) and from 0.264 (MILK) and 0.836 (CWT) for DEPD-PA and DEPD, respectively. The mean accuracy of MBV across all traits, response variables, and number of folds for cross-validation were 0.540 using a 50K genotype panel and 0.552 using a 80K genotype panel. The largest difference in mean accuracy of MBV between the 2 genotype panels was found for MARB (0.036) followed by ME (0.027) and REA (0.025). These results indicate very similar levels of accuracy of MBV between using Bovine SNP50K and GGPHD genotype panels.

Number of Folds for Cross-Validation

To verify the effect of the number of folds used to partition animals in training on accuracies of MBV, the predictive abilities of MBV were further assessed by comparing 3-, 5-, and 10-fold cross-validation. Tables 4 and 5 show the accuracies of MBV by the number of folds using the same response variables and SNP density for all traits. The mean accuracies of MBV across all traits were 0.549, 0.559, and 0.581 using 3, 5, and 10 folds, respectively, with DEPD-PA. With DEPD, the mean accuracies of MBV were 0.526, 0.523, and 0.540 for 3, 5, and 10 folds, respectively. These results indicate that the mean accuracies of MBV were similar or slightly greater as the number of folds increased. However, higher accuracies were observed when the number of folds was less than 10 (e.g., 3 or 5 folds) for CE, CETM, and ME when DEPD-PA was used as the response variable or for HPG, ME, and STAY when DEPD was used as the response variable for both SNP densities.

Removal of Parental Information

The average accuracies across the studied traits were 0.546 and 0.553 for 50K and 80K, respectively, based on DEPD-PA and 0.520 and 0.532 for 50K and 80K, respectively, based on DEPD with 3-fold cross-validation (Tables 4 and 5). The use of different response variables (DEPD-PA and DEPD) produced noticeable differences in the accuracy of MBV.

Table 5. Genetic correlations between molecular breeding values and deregressed EPD (deregressed EPD adjusting for parental information [DEPD-PA] and deregressed EPD without adjusting for parental information [DEPD]) and their SE in U.S. Red Angus beef cattle across the studied traits for the panel consisting of approximately 80,000 SNP

	DEPD-PA								DEPD							
		3 fold		5 fold		10 fold				3 fold		5 fold		10 fold		
Trait[1]	No.	r_g[2]	SE	r_g	SE	r_g	SE	Average	No.	r_g	SE	r_g	SE	r_g	SE	Average
BFAT	4,141	0.410	0.027	0.450	0.027	0.438	0.027	0.429	7,378	0.437	0.027	0.437	0.027	0.444	0.027	0.439
BWT	5,706	0.687	0.018	0.697	0.018	0.719	0.018	0.701	7,547	0.638	0.019	0.665	0.019	0.687	0.019	0.663
CE	1,142	0.604	0.046	0.663	0.044	0.626	0.047	0.631	7,597	0.703	0.040	0.773	0.037	0.773	0.038	0.749
CETM	1,240	0.618	0.044	0.645	0.045	0.620	0.046	0.628	7,591	0.611	0.023	0.631	0.023	0.664	0.034	0.635
CWT	5,702	0.775	0.019	0.830	0.017	0.836	0.017	0.813	7,479	0.799	0.036	0.833	0.035	0.836	0.023	0.823
HPG	227	0.302	0.107	0.270	0.110	0.378	0.110	0.317	3,579	0.714	0.080	0.667	0.082	0.681	0.083	0.687
MEN	505	0.549	0.054	0.465	0.059	0.535	0.059	0.517	2,086	0.563	0.053	0.541	0.057	0.555	0.056	0.553
MARB	4,452	0.526	0.032	0.495	0.032	0.521	0.033	0.514	7,292	0.468	0.033	0.489	0.033	0.478	0.033	0.478
MILK	3,653	0.417	0.034	0.397	0.034	0.428	0.034	0.414	7,530	0.264	0.032	0.266	0.032	0.271	0.031	0.267
REA	4,420	0.616	0.032	0.621	0.033	0.658	0.033	0.632	7,228	0.548	0.034	0.529	0.034	0.556	0.034	0.544
STAY	755	0.353	0.065	0.384	0.067	0.461	0.066	0.399	2,671	0.392	0.063	0.296	0.070	0.340	0.069	0.343
WWT	5,639	0.676	0.021	0.721	0.020	0.710	0.021	0.702	7,314	0.368	0.025	0.381	0.025	0.378	0.025	0.375
YWT	5,312	0.655	0.021	0.689	0.021	0.687	0.021	0.677	6,905	0.417	0.025	0.430	0.025	0.436	0.025	0.428
Average		0.553		0.563		0.586		0.567		0.532		0.534		0.546		0.537

[1]BFAT = back fat thickness; BWT = birth weight; CE = calving ease direct; CETM = calving ease maternal; CWT = carcass weight; HPG = heifer pregnancy; MARB = marbling score; MEN = maintenance energy; MILK = maternal milk ability; REA = rib eye muscle area; STAY = stayability; WWT = weaning weight; YWT = yearling weight.

[2]r_g = genetic correlation.

Using either the 50K or the 80K, the accuracy of MBV based on the DEPD-PA was greater than that based on DEPD for WWT (33.1%), YWT (25.6%), MILK (14.2%), REA (8.5%), STAY (7.2%), MARB (4.7%), and BWT (3.0%). However, increased accuracy of MBV based on DEPD was found for HPG (37%), CE (11.5%), and ME (3.7%). Differences in accuracy of MBV between DEPD-PA and DEPD were not found for BFAT, CETM, and CWT. Across both SNP densities and all numbers of folds in cross-validation, average accuracies of MBV across all traits were 3.3% greater for DEPD-PA than for DEPD.

DISCUSSION

In the current study, we compared the accuracy of MBV across 13 traits with 2 different marker density genotype panels (50K vs. 80K). The results showed no significant improvement in the accuracy of MBV based on increasing the SNP density. These results are consistent with those of previous studies (Erbe et al., 2012; Su et al., 2012; Gunia et al., 2014; Lu et al., 2016), which compared the accuracy between the 777K and 50K genotype panels. Su et al. (2012) observed no difference in prediction accuracy when the imputed 777K panel was used rather than 50K panel for the traits of protein, fertility, and udder health in Nordic Holstein and Red Dairy cattle. Erbe et al. (2012) also reported no gain in accuracy when using the 777K panel vs. the 50K panel within breed. In French Charolais beef cattle, Gunia et al. (2014) reported that the mean accuracy for 5 traits (birth and weaning weight, calving ease, and muscular and skeletal developments) using a 777K panel was slightly reduced (0.03) compared with using a 50K panel when using genomic BLUP (**GBLUP**), and similar accuracy was observed using a BayesC model. In other beef cattle breeds, for feed efficiency traits, Lu et al. (2016) compared the accuracies between 50K and high-density (e.g., 777K) panels and found that using the 50K panel resulted in slightly higher accuracy in pure breeds (Angus and Charolais) and similar accuracy in crossbreds (Angus–Hereford cross and Beefbooster composite). Taken together, our current results (50K vs. 80K) and the results of other genomic evaluation studies (777K vs. 50K) suggest that there are no significant differences in prediction accuracies by simply increasing SNP density. These are contrary to the expectation that the accuracy of genomic prediction could improve as a result of an increased degree of linkage disequilibrium (**LD**) between SNP markers and QTL (Meuwissen and Goddard, 2010) and the result that the extent of LD had major impact on the prediction accuracy in simulation study (Brito et al., 2011). The reason why this expectation was not realized could be due to a general lack of strong LD between markers and QTL and the fact that adding additional noninformative SNP simply adds an additional source of noise to prediction equations.

Cross-validation using the *K*-folds clustering method is often used to evaluate the performance of genomic predictions (Saatchi et al., 2011; Boddhireddy et al., 2014). The genotyped animals in this study were clustered by various numbers of folds (3, 5, and 10 folds) for cross-validation using *K*-means clustering methodology to maximize the differences in relatedness between training and evaluation data sets. As we applied the same clustering methodology, we observed the ratios of average relatedness between training and evaluation data sets (a_{ij_ratio}; 1.48 vs. 1.83 vs. 2.80) were larger by increasing the number of folds for cross-validation (Table 3). According to previous studies (Boddhireddy et al., 2014; Lu et al., 2016), increasing the number of animals in training has been shown to increase the accuracy of genomic predictors. The results of the current study showed similar or very slight improvements in mean accuracy of MBV with a larger number of folds for cross-validation for growth (BWT, WWT, and YWT) and carcass (BFAT, CWT, MARB, and REA) traits. These traits are at least moderately heritable and are not sex limited, meaning that the information content of individual animal EPD is greater and more animals were available for training. It is possible that the advantage of increasing the training data size by increasing the number of folds was offset by increasing the average relatedness between training and evaluation data sets. However, this similar or very slight increase in accuracy of MBV was not observed for CE, CETM, and ME based on DEPD-PA and for HPG, ME, and STAY based on DEPD due to the decreased number of animals contained in the training data set because several animal records were eliminated based on a reliability cutoff in the deregression process. More distant relationships compared with other traits in terms of the ratios of average relatedness (a_{ij_ratio}) between training and evaluation data sets were observed for CE, CETM, and ME based on DEPD-PA and for HPG, ME, and STAY based on DEPD in 10-fold cross-validation. This likely resulted in decreased accuracies of MBV for 10-fold cross-validation compared with either 3 or 5 folds.

When comparing results using either DEPD-PA or DEPD, we found increased prediction accuracies for growth and carcass traits based on DEPD-PA. As opposed to sex-limited traits, it is possible for all individuals to have a growth or carcass (using ultrasound proxies) record relatively early in life (before or shortly after 1 yr of age). Given the added information content available for these traits, the removal of parental average removed a source of "noise" in the analysis. The comparison of results from differing response variables were also reported in previous studies (Guo et al., 2010; Ostersen et al., 2011; Boddhireddy et al., 2014; Gunia et al., 2014), which compared the prediction accuracies between daughter yield variation and EBV or DEBV and EBV. Ostersen et al. (2011) reported that the improvement in accuracy was approximately 39 and 18% for daily gain and feed conversion ratio when using DEBV compared with EBV as response variables, respectively, in a swine population. Gunia et al. (2014) also reported a slightly greater accuracy by using DEBV rather than EBV for growth and development traits in Charolais beef cattle. On the other hand, Guo et al. (2010) reported that using EBV performed slightly better than using daughter yield variation in terms of prediction accuracies across simulation scenarios. Using Angus cattle, Boddhireddy et al. (2014) observed that the accuracies obtained using EBV as response variables were higher than those obtained using DEBV in both cross- and external validations. Interestingly, in the current study we found decreased prediction accuracies for recently introduced and sex-limited traits (e.g., CE, CETM, HPG, and ME) when using DEPD-PA as the response variable. The EPD of these traits were associated with lower reliabilities than growth and carcass traits and thus inherently reduce the size of the training population compared with these other traits after removing low-reliability animals from the analysis. Removing parental average as part of the deregression process has been reported as theoretically more appropriate to address the issue of double counting (Garrick et al., 2009; Ostersen et al., 2011). We contend that the inclusion of animals in training without information content beyond parental average contributes to the issue of double counting as evidenced by the increased, perhaps inflated, genetic correlations when DEPD was used for traits that suffer from a general lack of phenotypic information. These results are consistent with those reported by Boddhireddy et al. (2014), in which using EBV as the response variable without removing parental contribution yielded greater prediction accuracies compared with DEBV with the parental contribution removed for the traits with low heritability.

Comparison of Genomic Prediction Accuracy in Other Beef Cattle Breeds

A direct comparison of the accuracy of MBV across studies and populations is challenging given differences in clustering methodologies (e.g., random vs. *K*-means vs. identity by state IBS clustering), the metric used to assess accuracy (e.g., simple vs. genetic correlation), and the models used (e.g., GBLUP vs. Bayesian mixture models). The mean accuracies of MBV across the studied traits obtained for Red Angus beef cattle in this study were greater than those reported by Gunia et al. (2014) in Charolais beef cattle (0.40) using a BayesC model and the Bovine SNP50K

genotype panel and by Fernandes Júnior et al. (2016) in Nellore cattle when applying adjusted phenotype (0.35) or EBV (0.31) as response variables using a BayesC model. Similar mean accuracy of MBV across the studied traits was reported by Saatchi et al. (2012) in Limousin (0.55) and Simmental (0.50) beef cattle breeds using *K*-means clustering. Boddhireddy et al. (2014) reported mean accuracies of 0.51 in Angus beef cattle with *K*-means cross-validation and using DEBV as the response variable.

Conclusion

This study compared 2 response variables, 2 panel densities, and various cluster sizes for *K*-mean cross-validation and reported the corresponding impacts on the accuracy of MBV. The differences between panel densities (50K vs. 80K) were negligible. The number of clusters used for cross-validation is likely population specific and is defined a priori. The current study showed little impact on the number of clusters assumed and further illustrated that there is a trade-off between the number of clusters, the training population size, and the relationship between clusters. Of the 3 considerations contemplated herein (panel density, number of clusters for cross-validation, and the choice of response variables), resulting genetic correlations seem to be the most sensitive to the choice of response variables. The DEPD-PA response variable resulted in higher MBV accuracies for growth and carcass traits and much more conservative values for recent sex-limited traits where information content, and therefore, the number of animals in training, is the lowest. Genomic predictions built for traits with limited training data, as was the case for some sex-limited traits in the current study, should be viewed with caution. Consequently, for U.S. Red Angus beef cattle, the recommendation would be to use genomic prediction equations derived from the use of DEPD-PA, and that the choice of the 2 panels considered herein is not consequential and should be based on factors other than prediction accuracy.

LITERATURE CITED

Beef Improvement Federation (BIF). 2010. Guidelines for uniform beef improvement programs. 9th ed. BIF, Raleigh, NC. http://beefimprovement.org/content/uploads/2013/07/BIFGuidelinesFinal_updated0916.pdf (Accessed 1 January 2016.)

Boddhireddy, P., M. J. Kelly, S. Northcutt, K. C. Prayaga, J. Rumph, and S. Denise. 2014. Genomic predictions in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. J. Anim. Sci. 92:485–497. doi:10.2527/jas.2013-6757

Brito, F. V., J. B. Neto, M. Sargikzaei, J. A. Cobuci, and F. S. Schenkel. 2011. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. BMC Genet. 12:80. doi:10.1186/1471-2156-12-80

Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

Calus, M. P. L., and R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with marker density of one SNP per cM. J. Anim. Breed. Genet. 124:362–368. doi:10.1111/j.1439-0388.2007.00691.x

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. B. Robinson. 2008. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. J. Dairy Sci. 91:3225–3236. doi:10.3168/jds.2007-0333

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–4129. doi:10.3168/jds.2011-5019

Fernandes Júnior, G. A., G. J. M. Rosa, B. D. Valente, R. Carvalheiro, F. Baldi, D. A. Garcia, D. G. M. Gordo, R. Espigolan, L. Takada, R. L. Tonussi, W. B. F. de Andrade, A. F. B. Magalhaes, L. A. L. Chardulo, H. Tonhati, and L. G. de Albuquerque. 2016. Genomic prediction of breeding values for carcass traits in Nellore cattle. Genet. Sel. Evol. 48:7. doi:10.1186/s12711-016-0188-y

Gao, N., J. Li, J. He, G. Xiao, Y. Luo, H. Zhang, Z. Chen, and Z. Zhang. 2015. Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. BMC Genet. 16:120. doi:10.1186/s12863-015-0278-9

Garrick, D. J., and R. L. Fernando. 2013. Implementing a QTL detection study (GWAS) using genomic prediction methodology. In: C. Gondro, J. H. J. van der Werf, and B. Hayes, editors, Genome-wide association studies and genomic prediction. Springer Series, Berlin, Germany. p. 275–298. doi:10.1007/978-1-62703-447-0_11

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55. doi:10.1186/1297-9686-41-55

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2015. ASReml User Guide Release 4.1 Structural Specification, VSN International Ltd, Hemel Hempstead, HP1 1ES, UK

Goddard, M. E. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica (The Hague) 136:245–257.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381–391. doi:10.1038/nrg2575

Gunia, M., R. Saintilan, E. Venot, C. Hoze, M. N. Fouilloux, and F. Phocas. 2014. Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. J. Anim. Sci. 92(8):3258–3269. doi:10.2527/jas.2013-7478

Guo, G., M. S. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. J. Anim. Breed. Genet. 127:423–432. doi:10.1111/j.1439-0388.2010.00878.x

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607. doi:10.1534/genetics.113.152207

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinf. 12:186. doi:10.1186/1471-2105-12-186

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42:5–10. doi:10.1186/1297-9686-42-5

Hartigan, J. A., and M. A. Wong. 1979. A *k*-means clustering algorithm. Appl. Stat. 28:100–108. doi:10.2307/2346830

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat color, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6(9):e1001139. doi:10.1371/journal.pgen.1001139

Kachman, S. D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, J. F. Taylor, and E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45:30. doi:10.1186/1297-9686-45-30

Kizilkaya, K., R. L. Fernando, and D. J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotype. J. Anim. Sci. 88:544–551. doi:10.2527/jas.2009-2064

Lu, D., E. C. Akanno, J. J. Crowley, F. Schenkel, H. Li, M. De Pauw, S. S. Moore, Z. Wang, C. Li, P. Stothard, G. Plastow, S. P. Miller, and J. A. Basarab. 2016. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. J. Anim. Sci. 94(4):1342–1353. doi:10.2527/jas.2015-0126

Meuwissen, T. H. E., and M. E. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole genome resequencing. Genetics 185:623–631. doi:10.1534/genetics.110.116590

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

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet. Sel. Evol. 43:38. doi:10.1186/1297-9686-43-38

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. Dekkers, R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using means clustering for cross-validation. Genet. Sel. Evol. 43:40. doi:10.1186/1297-9686-43-40

Saatchi, M., R. D. Schnabel, M. M. Rolf, J. F. Taylor, and D. J. Garrick. 2012. Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. Genet. Sel. Evol. 44:38. doi:10.1186/1297-9686-44-38

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. doi:10.1186/1471-2164-15-478

Su, G., R. F. Brondum, P. Ma, B. Guldbrandtsen, G. R. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (similar to 54,000) and high-density (similar to 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657–4665. doi:10.3168/jds.2012-5379