The impact of clustering methods for cross-validation, choice of phenotypes, and genotyping strategies on the accuracy of genomic predictions

Johnna L. Baller,^{*,1} Jeremy T. Howard,^{*} Stephen D. Kachman,[†] and Matthew L. Spangler^{*}

^{*}Department of Animal Science, University of Nebraska, Lincoln, NE 68583; and [†]Department of Statistics, University of Nebraska, Lincoln, NE 68583

ABSTRACT: For genomic predictors to be of use in genetic evaluation, their predicted accuracy must be a reliable indicator of their utility, and thus unbiased. The objective of this paper was to evaluate the accuracy of prediction of genomic breeding values (GBV) using different clustering strategies and response variables. Red Angus genotypes ($n = 9,763$) were imputed to a reference 50K panel. The influence of clustering method [k-means, k-medoids, principal component (PC) analysis on the numerator relationship matrix (**A**) and the identical-by-state genomic relationship matrix (**G**) as both data and covariance matrices, and random] and response variables [deregressed estimated breeding values (DEBV) and adjusted phenotypes] were evaluated for cross-validation. The GBV were estimated using a Bayes C model for all traits. Traits for DEBV included birth weight (BWT), marbling (MARB), rib-eye area (REA), and yearling weight (YWT). Adjusted phenotypes included BWT, YWT, and ultrasonically measured intramuscular fat percentage and REA. Prediction accuracies were estimated using the genetic correlation between GBV and associated response variable using a bivariate animal model. A simulation mimicking a cattle

population, replicated 5 times, was conducted to quantify differences between true and estimated accuracies. The simulation used the same clustering methods and response variables, with the addition of 2 genotyping strategies (random and top 25% of individuals), and forward validation. The prediction accuracies were estimated similarly, and true accuracies were estimated as the correlation between the residuals of a bivariate model including true breeding value (TBV) and GBV. Using the adjusted Rand index, random clusters were clearly different from relationship-based clustering methods. In both real and simulated data, random clustering consistently led to the largest estimates of accuracy, while no method was consistently associated with more or less bias than other methods. In simulation, random genotyping led to higher estimated accuracies than selection of the top 25% of individuals. Interestingly, random genotyping seemed to overpredict true accuracy while selective genotyping tended to underpredict accuracy. When forward in time validation was used, DEBV led to less biased estimates of GBV accuracy. Results suggest the highest, least biased GBV accuracies are associated with random genotyping and DEBV.

Key words: beef cattle, bias, genomic prediction, simulation

© The Author(s) 2019. Published by Oxford University Press on behalf of the American Society of Animal Science. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

J. Anim. Sci. 2019.97:1534–1549

doi: 10.1093/jas/skz055

INTRODUCTION

Many clustering methods have been proposed for cross-validation to assess the accuracy of genomic breeding values (GBV). Legarra et al.

¹Corresponding author: johnna.baller@huskers.unl.edu

Received November 23, 2018.

Accepted February 4, 2019.

(2014) used birth year within dairy sheep, Luan et al. (2009) used random assignment and year of progeny testing in Norwegian Red Cattle, and Liu et al. (2014) used random assignment and sets of half-sib families within Chinese triple-yellow chickens to determine training and validation sets. K-means clustering has been used to assess the accuracy of GBV in a variety of beef cattle breeds (Saatchi et al., 2011, 2012, 2013; Boddhireddy et al., 2014). However, Boddhireddy et al. (2014) showed that principal component (PC) clustering based on an identical-by-state (IBS) genomic relationship matrix (G) led to higher estimated accuracies than k-means clustering within an Angus population. The response variables used to estimate marker effects have also differed. Adjusted phenotypes led to higher estimated accuracies than nonadjusted phenotypes in sheep (Daetwyler et al., 2012). Deregressed Expected Progeny Differences (DEPD) have been used in the past to develop genomic predictors in U.S. beef breeds given genotyped animals were limited and DEPDP have greater information content than phenotypes alone. Moreover, genotyping strategy has also been shown to have an impact on estimated GBV accuracies as demonstrated by Ehsani et al. (2010).

While partial solutions exist relative to clustering method and choice of dependent variables, a direct comparison of multiple clustering methods with the use of adjusted phenotypes or deregressed estimated breeding values (DEBV) does not currently exist in the literature. Consequently, the current study aims to evaluate the effect of k-means, k-medoids, PC clustering based on the numerator relationship matrix and IBS genomic relationship matrix when these relationship matrices were treated as both a data matrix and covariance matrix, and random clustering on the estimates of accuracy of GBV using adjusted phenotypes or DEBV.

MATERIALS AND METHODS

Animal care and use committee approval was not required for this study as all data were either obtained from existing databases or simulated.

Red Angus

Red Angus animals ($n = 11,972$) were genotyped with multiple SNP panels ranging from 25,259 to 139,376 SNP. Phenotypic data could be matched to 9,763 of these animals. Unmapped SNP as well as SNP from different panels with the same name but different positions were discarded. Animals with a

call rate less than 80% were removed from the analysis. Using FImpute v2.2 (Sargolzaei et al., 2014), the SNP panels were imputed to a 50K reference panel. After SNP located on the sex chromosomes were removed, 48,677 SNP were left for analysis.

Expected progeny differences (EPD) and their associated Beef Improvement Federation (BIF) accuracies (Beef Improvement Federation, 2010) were obtained from the Red Angus Association of America (RAAA) for the animals with genotypes as well as for their sires and dams. The EPD used for this analysis were birth weight (BWT), marbling (MARB), rib-eye area (REA), and yearling weight (YWT). The EPD were multiplied by 2 to form estimated breeding values (EBV) for consistency of scale with phenotypes and simulated data. Beef Improvement Federation accuracies were transformed into reliabilities and deregressed estimated breeding values (DEBV) that removed information from parental average contributions were computed following Garrick et al. (2009). The assumptions underlying the DEBV were that the proportion of genetic variance not accounted for by markers, c , was 0.4 (Saatchi et al., 2012), and heritability, h^2 , was also assumed to be 0.4. Animals with a reliability less than 0.1 were excluded from further analysis.

Phenotypes including BWT, ultrasonically measured intramuscular fat percentage, ultrasonically measured REA, and YWT were also obtained from RAAA. These phenotypes were preadjusted for sex, age, and breed composition. The final response variable used for analysis was the contemporary group deviation from these preadjusted phenotypes. Contemporary group included herd-year-season for birth and yearling weight, and the addition of date of measurement for ultrasound traits. Animals from a contemporary group less than 5 were excluded from further analysis. The numbers of contemporary groups (mean number of animals per group) were: 982 for BWT (105), 594 for YWT (53), and 487 for ultrasonic measurements (54). Of the animals used for analysis, 5,938 were male and 3,825 were female.

Simulation

A simulation was carried out using Geno-Diver (Howard et al., 2017) to mimic a purebred beef cattle population. Five replicates, each with a different founder genome, were simulated. The replicates contained 29 chromosomes of length 87 Mb, the average chromosome length as determined with the NCBI *Bos taurus* 2009 assembly. Markers

representing a 50K SNP panel were randomly distributed across the genome; locations were randomly drawn from a uniform distribution, with 1,724 markers per chromosome. Quantitative trait loci (QTL) were assumed to occur once per 3 Mb, resulting in 29 QTL per chromosome. Locations of the QTL, placed randomly across the whole chromosomal range, were drawn from a uniform distribution. The phenotypic variance was set to 1 and the additive and dominance variances were set to 0.4 and 0.0, respectively, resulting in a phenotype with heritability of 0.4. The founder genome, generated by Markovian Coalescence Simulator (MaCS) program (Chen et al., 2009), employed a scenario in which a large amount of short-range linkage disequilibrium (LD) was generated. To generate the sequence data for the founder population, the “Ne70” option was specified within GenoDiver, which sets the effective population size of the founder population to 70. de Roos et al. (2008) found that cattle have a small effective population size, approximately 100 or less, and large amounts of LD at short distances. To establish a pedigree, founder animals consisting of 100 sires and 2,000 dams were randomly selected and mated for 5 generations. Selection continued for an additional 10 generations, where animals were mated randomly with the caveat that animals with additive relationships greater than 0.125 were not mated together in order to reduce inbreeding. Replacement animals were chosen based on the highest EBV determined by pedigree-based BLUP with a replacement rate of 0.4 for sires and 0.2 for dams. Animals were culled based on EBV or when they were in the population as a parent for 12 generations. Figure 1 provides a schematic of the simulation process.

All individuals ($n = 32,100$) from the 15 generations had a genotype retained. However, in current beef cattle populations, not all individuals are genotyped. Thus, approximately 25% of the animals from generations 6 to 15 were chosen as animals to have genotypes available. The genotyped animals were chosen using 2 scenarios: 25% of the animals born in generations 6 to 15 were chosen at random, or the top 25% of animals born in generations 6 to 15 were chosen based on EBV. The EBV were calculated using all information through generation 15 meaning that candidates for genotyping were selected based on all available information. These top animals were distributed across the 10 generations where selection occurred to account for genetic trend, so as not to include animals from only the last generations. Approximately the same number of animals came from generations 6 to

Figure 1. Schematic of simulation process. ¹Replacement rates: 0.4 for sires; 0.2 for dams. ²Animals culled randomly or based on EBV or when they were in the population as a parent for 12 generations. ³Sires and dams mated randomly with the caveat A_{ij} (additive relationship between animals i and j) was less than 0.125.

15 for the randomly chosen scenario. Phenotypes as well as EBV and associated accuracies were obtained for each replicate. Estimated breeding values were transformed into DEBV using the same assumptions as for the Red Angus data.

To assess the accuracy of GBV using available animals to predict the genetic merit of young selection candidates, forward selection was also performed with the simulated data. Breeding values using information through a specified generation were estimated using ASReml v3.0 software (Gilmour et al., 2008). All pedigree, genotype, and phenotype information were truncated at generations 11, 12, 13, and 14 in order to assess the impact of the addition of animals generationally closer to the youngest selection candidates. Data were truncated at the specified generations so that data in subsequent generations were not used in the estimation of the EBV for an animal in the training set. The model to estimate breeding values included phenotype as the response variable, intercept as the fixed effect, and animal as the random effect. Animals chosen to have available genotypes

were again picked randomly, or based on highest EBV distributed equally across the available generations. Animals genotyped in 1 scenario or truncation point were not guaranteed to be genotyped in other scenarios or truncation points. The new EBV were then transformed into DEBV using the same assumptions as the Red Angus data.

Cross-Validation Methods

Seven different clustering methods were employed for cross-validation: k-means, k-medoids, PC analysis of the numerator relationship matrix (**A**) and the **IBS** genomic matrix (**G**) assuming the matrices were either a data matrix or a covariance matrix, and random clustering. Each method used 5 folds for the Red Angus data. Lee et al. (2017) found that differences in the number of folds led to negligible differences in terms of prediction accuracies. Consequently, 3 folds were used for the simulated data given the reduced number of animals in the simulated data compared to the real cattle data set. For both data sets, training and evaluation sets were arranged using the evaluation set as 1 fold and the remaining folds as the training set. This was repeated so that each fold was used once as the evaluation set.

The **A** matrix was used to create the folds based on k-means and k-medoids. A distance matrix, **D**, was calculated as described by Saatchi et al. (2011). The elements of **D** were $d_{ij} = 1 - \frac{a_{ij}}{\sqrt{a_{ii} \times a_{jj}}}$ where d_{ij} is the measure of pedigree distance between animals i and j , a_{ij} is the additive genetic relationship between animals i and j , and a_{ii} and a_{jj} are the diagonal elements of the **A** matrix. A pedigree matrix was computed using the pedigree package (Coster, 2012) in R (R Core Team, 2017) for the genotyped animals. The Red Angus data made use of a 6-generation pedigree that consisted of 45,738 animals. The simulated data made use of the full pedigree of all 15 generations. K-means clusters were determined using the **D** matrix within the kmeans() function and specifying the Hartigan and Wong algorithm in the stats package (R Core Team, 2017) of R. K-medoids used the **D** matrix as a dissimilarity matrix in the pam() function within the cluster package (Maechler et al., 2018) of R.

The **G** matrix was computed as $\frac{\mathbf{M}\mathbf{M}'}{2 \sum p_i(1-p_i)}$, where **M** is the centered genotype incidence matrix and p_i is the allelic frequency of the second allele of the i th SNP (VanRaden, 2008). The correlation matrix of **A** or **G** was used in the princomp() function in the stats package (R Core Team, 2017) of R

in order to create the folds for the PC analysis using the **A** matrix (**PCN**) and the **G** matrix (**PCG**). The **A** or **G** matrix was considered as a data matrix to form the folds of the PCN (Data) or PCG (Data) methods, respectively, and considered as a covariance matrix to form the folds of the PCN (Cov) and PCG (Cov) methods, respectively. When the **A** or **G** matrix was considered as a data matrix, a covariance matrix was first formed from **A** or **G** and this resulting covariance matrix was used for PC analysis. If **A** or **G** was considered as a covariance matrix, the **A** or **G** matrix itself was subjected directly to PC analysis. The coefficients of the first PC were ordered and then divided evenly into fifths for the Red Angus data or thirds for the simulated data. This led to animals with the highest coefficients being in 1 fold and animals with the lowest coefficients being in another.

Random clusters were determined by randomly assigning animals to 1 of 5 clusters for the Red Angus cattle or to 1 of 3 clusters for the simulated individuals.

The adjusted Rand index (Hubert and Arabie, 1985) measures the degree of agreeance between different partitions of a data set. The adjusted Rand index is corrected for chance using a generalized hypergeometric distribution to model randomness. Thus, the index has an expectation of 0 when partitions are random and has an upper bound of 1 in the case of complete agreeance between partitions. The higher the adjusted Rand index, the more agreement between the clustering methods. The adjusted Rand index was calculated between the 7 clustering methods for the Red Angus and simulated data to test the agreement between the clustering methods using the adjustedRandIndex() function within the mclust package (Fraley et al., 2012) of R.

For forward validation, training and evaluation sets were assigned based on generations. Training sets consisted of 5,000 animals included in generations 6 to 11, 6 to 12, 6 to 13, or 6 to 14. The evaluation set consisted of all 2,000 selection candidates in generation 15.

SNP Effect Estimation

SNP effects were estimated using a Bayes C model (Kizilkaya et al., 2010) implemented in GenSel4R (Garrick and Fernando, 2013). The model used for both the Red Angus and simulated data was:

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

where y_i is the DEBV or the adjusted phenotype for animal i for each of the 4 traits, μ is the overall mean, Z_{ij} is the covariate matrix for SNP j for animal i and k is the number of SNP, u_j is the random effect of SNP j , δ_j is a Bernoulli indicator variable indicating whether SNP j is included in the model, and e_i is the random residual of animal i . The random SNP effects and random residuals were both assumed to be identically and independently distributed with Gaussian distributions of $N(0, \sigma_u^2)$ and $N(0, \sigma_e^2)$, respectively. Independent inverse scaled chi-square priors were placed on the variance estimates for the random SNP effects and random residuals, σ_u^2 and σ_e^2 . The probability of a SNP not having an effect, π , was set to 0.99, as indicated by the Bernoulli indicator variable. Each model was run with 42,000 iterations, discarding the first 2,000 as the burn-in period.

Genetic Correlation and Regression Coefficients

Estimates of the genetic correlations ($r_{g,y}$) between the GBV and the DEBV or adjusted phenotypes were used as an estimate of the GBV accuracy. The square of the genetic correlations estimate the proportion of genetic variance explained by the GBV (Thallman et al., 2009). A bivariate animal model for each fold within each clustering method was fit using ASReml v3.0 software (Gilmour et al., 2008) in order to estimate genetic variances and covariances. Similar studies have also used a bivariate model approach to estimate GBV accuracy (e.g., Saatchi et al., 2012; Weber et al., 2012; Kachman et al., 2013; Lee et al., 2017). The model for the GBV and DEBV for the Red Angus data consisted of a fixed effect for the intercept and an unweighted residual for GBV and r -inverse for DEBV, where r -inverse is the weight according to the reliability of the DEBV. The model for the simulated data was the same except for the addition of the fixed effect of generation to account for the rapid genetic improvement across generations. For Red Angus animals, the model for GBV and adjusted phenotype consisted of a fixed effect for the intercept. Again, the model for the simulated data was similar except the response variable was phenotype and the model contained a fixed effect for generation. Regression coefficients of the response variable on GBV for Red Angus and simulation were calculated as the genetic covariance between the GBV and the associated response variable divided by the genetic variance of the GBV. An ideal regression coefficient would be 1, as the DEBV or adjusted phenotype

would not over- or underpredict the GBV. The Red Angus estimated genetic correlations and regression coefficients are presented as the average across the 5 folds for each trait. Estimated genetic correlations and regression coefficients from the simulated data are presented as the average of 3 folds averaged over the 5 replicates for cross-validation methods. For forward validation, estimated genetic correlations and regression coefficients were averaged over the 5 replicates.

The advantage of simulated data is that true breeding values (TBV) are known. A bivariate model including GBV and TBV as response variables and fixed effects of overall mean, generation, fold, and interaction of generation and fold was used to obtain residuals. The correlation between the residuals was used as the true accuracy of the genomic predictor. The regression coefficient of TBV on GBV was computed as the covariance between residuals of GBV and TBV divided by the variance of the residuals of GBV and considered as the true regression coefficient.

RESULTS AND DISCUSSION

Simulation

After the first round of selective replacement, generation 6, the 5 replicates had a mean (variance) of 0.267 (0.003), 0.256 (0.003), and 0.254 (0.003) for the phenotype, TBV, and EBV, respectively. Across the 5 replicates the mean (variance) were 2.94 (0.003), 2.94 (0.003), and 2.937 (0.003) for the phenotype, TBV, and EBV of animals at generation 15, which occurred after a total of 10 generations of selective replacement. The average correlation (r^2) between 2 SNP across a range of distances at generation 15 was consistent with having generated a large amount of short-range LD (results not shown).

Clustering Method

The purpose of clustering is to partition animals into training and evaluation sets to assess the ability to generalize estimates of SNP effects and predictions of genetic merit on animals that were not used to estimate the SNP effects. For Red Angus, the first PC of the **A** matrix when considered as a data or covariance matrix explained 26.85% and 4.56% of the variation in the additive relationships, respectively, while the first PC of the **G** matrix when considered as a data or covariance matrix explained 19.97% and 1.86% of the

variation in the additive genetic relationships, respectively. The percentage of variation for the simulated data was averaged across 5 replicates. The first PC of the **A** matrix when considered as a data (covariance) matrix explained $12.60 \pm 1.89\%$ ($2.56 \pm 0.02\%$) and $9.33 \pm 1.56\%$ ($2.66 \pm 0.27\%$) of the variation in the additive relationships using random selection for genotyping and using animals with the top 25% of EBV, respectively. The first PC of the **G** matrix when considered as a data (covariance) matrix explained $5.80 \pm 0.80\%$ ($1.09 \pm 0.08\%$) and $6.20 \pm 0.73\%$ ($1.29 \pm 0.10\%$) of variation in the additive genetic relationships using random selection and selection of the top 25% animals for genotyping, respectively. It appears that a larger fraction of additive relationships was captured by the first PC using the **A** matrix compared to using the **G** matrix as generationally, more data are contained in the **A** matrix than the **G** matrix. Also, a data matrix explained a greater fraction of variation compared to a covariance matrix due to the covariance matrix as used herein being largely bounded by 0 and 1.

Average maximum relationships of animals within and between folds were calculated for each clustering method and shown in [Table 1](#) for Red Angus animals and [Table 2](#) for simulated animals. For Red Angus, the within cluster average maximum relationships were similar for the different clustering methods, ranging on average between 0.34 and 0.35 with the exception of random clustering which was lower (0.31) and k-means which was higher (0.37). The between cluster average maximum relationships were similar for the different clustering methods with averages ranging from 0.19 to 0.24 with the exception of random clustering (0.31). A similar pattern was observed when evaluating the simulated data. With the exception of random clustering, the within and between average maximum relationships were very similar across clustering methods with random clustering having a lower within cluster and higher between cluster average maximum relationship. The average maximum relationships overall were higher when the animals with the top EBV were chosen to be genotyped. This was expected given the trait was simulated to be moderately heritable and thus selective genotyping based on genetic merit is likely to choose more closely related individuals. The average number of progeny per sire within the animals that were genotyped increased from 10.87 to 11.71 between random genotyping and genotyping the top 25% of individuals. The maximum number of progeny included in the analysis for an individual

sire also doubled when genotyping the top 25% of animals compared to random genotyping.

Using registered Angus animals, [Boddhireddy et al. \(2014\)](#) compared k-means clustering, PC clustering based on an IBS **G** matrix, and random clustering for cross-validation. Their results showed that relationships were maximized within clusters and minimized across clusters with the exception of random clustering. Taken together, the results contained herein and previous work shows the ability of k-means, k-medoids, and PC analysis to partition animals with higher or lower degrees of relationship into different clusters.

[Tables 3](#) and [4](#) contain the adjusted Rand index values for the Red Angus and simulated data, respectively. For the Red Angus data, random clustering was clearly different as compared to any other clustering method, as expected. There was high agreement between PC using a data matrix or covariance matrix for **G** (0.67). Interestingly, high agreement was also found between k-means and PCN (Cov) clustering (0.45). Similarly, in the simulated data, random clustering compared to any other clustering method led to an index of approximately 0. Principal component methods across respective relationship matrices led to the highest indices. K-means also had high agreement with PC clustering on the **A** matrix, whether the data or covariance matrix was considered. These patterns were observed over both genotyping strategies. Overall, simulation tended to lead to slightly higher adjusted Rand indexes than Red Angus.

Estimated accuracies of GBV for each clustering method using the Red Angus animals and simulated data are shown in [Tables 5](#) and [6](#), respectively. In Red Angus, the average estimated accuracies across traits using DEBV were 0.58, 0.55, 0.61, 0.60, 0.60, 0.60, and 0.66 for the k-means, k-medoids, PCN (Data), PCN (Cov), PCG (Data), PCG (Cov), and random clustering methods, respectively. The average estimated accuracies across traits using adjusted phenotypes were 0.42, 0.45, 0.51, 0.50, 0.50, 0.52, and 0.59 for the k-means, k-medoids, PCN (Data), PCN (Cov), PCG (Data), PCG (Cov), and random clustering methods, respectively. Overall, random clustering led to the highest estimated accuracy while k-means and k-medoids consistently led to the lowest. Differences in estimated accuracies were negligible when comparing PC clustering on either the **A** or **G** matrix.

Using simulated data, random clustering led to the highest estimated accuracy. However, all other estimated accuracies were similar when comparing the other clustering methods. This was observed

Table 1. Red Angus average maximum relationships

Clustering method ¹	Fold	<i>N</i>	$a_{\max_within}^2$	$a_{\max_between}^3$
K-means	1	2,070	0.40	0.22
	2	615	0.41	0.21
	3	572	0.39	0.20
	4	3,592	0.31	0.14
	5	2,914	0.34	0.19
K-medoids	1	1,661	0.36	0.21
	2	1,783	0.32	0.17
	3	1,839	0.36	0.20
	4	3,803	0.35	0.19
	5	377	0.37	0.19
PCN (Data)	1	1,952	0.38	0.20
	2	1,952	0.36	0.23
	3	1,952	0.33	0.23
	4	1,952	0.34	0.21
	5	1,955	0.32	0.19
PCN (Cov)	1	1,952	0.42	0.21
	2	1,952	0.36	0.24
	3	1,952	0.32	0.24
	4	1,952	0.31	0.22
	5	1,955	0.29	0.17
PCG (Data)	1	1,952	0.40	0.27
	2	1,952	0.34	0.26
	3	1,952	0.33	0.25
	4	1,952	0.31	0.23
	5	1,955	0.31	0.18
PCG (Cov)	1	1,952	0.32	0.18
	2	1,952	0.31	0.22
	3	1,952	0.32	0.25
	4	1,952	0.35	0.26
	5	1,955	0.40	0.26
Random	1	1,994	0.31	0.31
	2	1,916	0.31	0.31
	3	1,893	0.31	0.31
	4	2,000	0.31	0.31
	5	1,960	0.31	0.31

¹K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix (**A** = Data); PCN (Cov) = principal component clustering using a numerator relationship matrix (**A** = Covariance matrix); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Data); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Covariance matrix); random = random clustering.

² $a_{\max_within}$ = average of the maximum relationship of an animal with other animals within its own fold.

³ $a_{\max_between}$ = average of the maximum relationship of an animal with other animals not within its own fold.

across both genotyping methods. However, no clustering method was consistently associated with more or less bias than the other clustering methods when comparing the difference between the estimated and true accuracy. Many studies have shown the relationships between the training and validation sets can impact the prediction accuracy. [Habier et al. \(2007\)](#) stated that the accuracies of genome-assisted breeding values (GEBV) are a result of the genetic relationships captured by markers. In a study of German Holstein cattle,

[Habier et al. \(2010\)](#) demonstrated the accuracy of GEBV decreased with decreasing additive genetic relationship values across training and validation sets with cross-validation. That is, the accuracies decreased as the training and validation sets became less related. Similar results were found by [Clark et al. \(2012\)](#) in both a simulated data set and data set containing Merino sheep. Moreover, similar results were reported by [Pszczola et al. \(2012\)](#) using simulated data as well as [Chen et al. \(2013\)](#) using purebred Angus and Charolais cattle. Interestingly,

Table 2. Simulated average maximum relationships and standard errors

Clustering method ³	Random selection ¹		Top EBV ²	
	a_{max_within} ⁴	$a_{max_between}$ ⁵	a_{max_within}	$a_{max_between}$
K-means	0.35 (0.002)	0.23 (0.004)	0.49 (0.002)	0.32 (0.003)
K-medoids	0.35 (0.002)	0.26 (0.001)	0.49 (0.001)	0.34 (0.003)
PCN (Data)	0.35 (0.003)	0.23 (0.003)	0.48 (0.003)	0.33 (0.003)
PCN (Cov)	0.34 (0.002)	0.24 (0.003)	0.47 (0.003)	0.33 (0.003)
PCG (Data)	0.35 (0.002)	0.25 (0.004)	0.47 (0.003)	0.35 (0.004)
PCG (Cov)	0.35 (0.002)	0.25 (0.004)	0.47 (0.002)	0.35 (0.003)
Random	0.32 (0.002)	0.31 (0.001)	0.41 (0.002)	0.41 (0.002)

¹Random selection = 5,000 animals randomly chosen across all 10 generations.

²Top EBV = the top 500 individuals from each of the 10 generations.

³K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix (**A** = Data); PCN (Cov) = principal component clustering using a numerator relationship matrix (**A** = Covariance matrix); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Data); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Covariance matrix); random = random clustering.

⁴ a_{max_within} = average of the maximum relationship of an animal with other animals within its own fold.

⁵ $a_{max_between}$ = average of the maximum relationship of an animal with other animals not within its own fold.

Table 3. Red Angus adjusted Rand index

Clustering method ¹	K-means	K-medoids	PCN (Data)	PCN (Cov)	PCG (Data)	PCG (Cov)	Random
K-means	1.00	0.14	0.26	0.45	0.22	0.21	0.00
K-medoids		1.00	0.10	0.07	0.08	0.08	0.00
PCN (Data)			1.00	0.23	0.15	0.14	0.00
PCN (Cov)				1.00	0.19	0.19	0.00
PCG (Data)					1.00	0.67	0.00
PCG (Cov)						1.00	0.00
Random							1.00

¹K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix (**A** = Data); PCN (Cov) = principal component clustering using a numerator relationship matrix (**A** = Covariance matrix); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Data); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Covariance matrix); random = random clustering.

maximum relationships within and between folds for random clustering in the simulation were more comparable to those obtained for other clustering methods while there was a larger difference between relationships within and between folds between random clustering and other clustering methods in the Red Angus data. Consequently, any estimate of bias is more likely related to the ability of clustering methods to minimize relationships between folds and maximize them within folds. Based on the comparison of maximum relationship values, random clustering was more comparable to other methods in simulation than it was in the Red Angus data at partitioning animals based on additive relationships.

The pattern of estimated accuracies using different clustering methods for cross-validation using Red Angus was also seen in previous studies. [Saatchi et al. \(2011\)](#) demonstrated the use of k-means clustering based on the additive genetic relationships

between animals as a means for clustering animals for cross-validation. They used registered Angus bulls and found that k-means clustering yielded lower estimated accuracies than random clustering for 16 traits. Similar results were seen using American Hereford animals ([Saatchi et al., 2013](#)). Additionally, [Boddhireddy et al. \(2014\)](#) compared random, k-means, and clustering on the first PC of the IBS genomic relationship matrix (data matrix) using registered Angus animals. Their results showed that PC clustering resulted in accuracy estimates that were intermediate to k-means and random clustering for BWT. The estimated accuracies across 15 additional traits showed that k-means clustering resulted in lower estimated accuracies compared to PC clustering.

In Red Angus, the average estimated regression coefficients of DEBV on GBV across traits were 0.83, 0.80, 0.89, 0.87, 0.89, 0.89, and 0.93 for the k-means, k-medoids, PCN (Data), PCN (Cov),

Table 4. Simulated adjusted Rand index and standard errors of randomly selected genotyped animals (above diagonal) and selection of top animals for genotyping (below diagonal)

Clustering method ¹	K-means	K-medoids	PCN (Data)	PCN (Cov)	PCG (Data)	PCG (Cov)	Random
K-means	1.00	0.09 (0.0259)	0.48 (0.0335)	0.45 (0.0180)	0.34 (0.0266)	0.32 (0.0278)	0.00 (0.0001)
K-medoids	0.15 (0.0194)	1.00	0.05 (0.0132)	0.02 (0.0054)	0.05 (0.0125)	0.06 (0.0133)	0.00 (0.0001)
PCN (Data)	0.39 (0.0052)	0.06 (0.0179)	1.00	0.56 (0.0298)	0.43 (0.0260)	0.40 (0.0323)	0.00 (0.0001)
PCN (Cov)	0.37 (0.0178)	0.02 (0.0075)	0.44 (0.0314)	1.00	0.35 (0.0283)	0.31 (0.0323)	0.00 (0.0001)
PCG (Data)	0.27 (0.0213)	0.09 (0.0172)	0.31 (0.0478)	0.19 (0.0393)	1.00	0.84 (0.0320)	0.00 (0.0001)
PCG (Cov)	0.26 (0.0183)	0.10 (0.0180)	0.28 (0.0377)	0.16 (0.0304)	0.85 (0.0210)	1.00	0.00 (0.0001)
Random	0.00 (0.0002)	0.00 (0.0001)	0.00 (0.0001)	0.00 (0.0001)	0.00 (0.0004)	0.00 (0.0004)	1.00

¹K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix ($\mathbf{A} = \text{Data}$); PCN (Cov) = principal component clustering using a numerator relationship matrix ($\mathbf{A} = \text{Covariance matrix}$); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix ($\mathbf{G} = \text{Data}$); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix ($\mathbf{G} = \text{Covariance matrix}$); random = random clustering.

PCG (Data), PCG (Cov), and random clustering methods, respectively. The average estimated regression coefficients of adjusted phenotypes on GBV across traits were 0.93, 0.91, 0.99, 0.97, 0.96, 0.97, and 1.04 for the k-means, k-medoids, PCN (Data), PCN (Cov), PCG (Data), PCG (Cov), and random clustering methods, respectively. K-means and k-medoids clustering led to the lowest regression coefficient estimates, whereas random clustering led to the largest regression coefficient estimates.

Table 7 contains the mean estimated regression coefficients of either phenotype or DEBV on GBV as well as the TBV on GBV using the simulated data. All estimated regression coefficients were similar across clustering methods and across genotyping methods. Additionally, all clustering methods underestimate performance as the estimated regression coefficients were below 1 across both genotyping methods.

Choice of Dependent Variable

With Red Angus, estimated accuracies were generally higher when DEBV were used compared to adjusted phenotypes and the associated standard errors were lower. Mean DEBV accuracies were significantly different ($P < 0.03$) from mean adjusted phenotype accuracies for all traits except YWT ($P = 0.464$). The differences in the standard errors between these 2 dependent variables demonstrate the additional information gained from the DEBV as compared to adjusted phenotypes. In contrast, phenotypes led to negligible numerical differences in mean estimated accuracies when compared to the DEBV in the simulated data for the selection of the top 25% of animals for genotyping ($P = 0.053$). However, there was a statistically significant difference between mean estimated accuracies of phenotypes compared to the DEBV for random genotyping ($P = 0.006$). The mean absolute differences, across replicates, between estimated and true accuracy were 0.05 and 0.06 for phenotypes and DEBV, respectively, within the random genotyping scenario. Additionally, the mean absolute differences were 0.12 and 0.20 for phenotypes and DEBV, respectively, within the selective genotyping scenario. This illustrated that the amount of bias, measured as the difference between estimated and true accuracy, was dependent upon the genotyping strategy. The discrepancy seen between the Red Angus and simulated data may be due to the population structures. The Red Angus had on average 7 progeny per sire. However, within the simulation, there were approximately 11 to 12 progeny per

Table 5. Average accuracy estimates and standard errors across all 5 folds for Red Angus

Trait ¹	Clustering method ²	Adjusted phenotypes ³			DEBV ⁴		
		<i>N</i>	$(r_{g,y})^5$	SE ⁶	<i>N</i>	$(r_{g,y})$	SE
BWT	K-means	9,282	0.49	0.06	7,214	0.69	0.05
	K-medoids		0.49	0.05		0.66	0.04
	PCN (Data)		0.56	0.06		0.68	0.04
	PCN (Cov)		0.60	0.06		0.68	0.04
	PCG (Data)		0.55	0.06		0.67	0.04
	PCG (Cov)		0.58	0.06		0.68	0.04
	Random		0.77	0.10		0.74	0.03
YWT	K-means	6,278	0.46	0.08	6,061	0.54	0.06
	K-medoids		0.54	0.09		0.48	0.05
	PCN (Data)		0.55	0.08		0.56	0.05
	PCN (Cov)		0.53	0.07		0.55	0.05
	PCG (Data)		0.57	0.08		0.58	0.04
	PCG (Cov)		0.57	0.08		0.57	0.04
	Random		0.57	0.04		0.63	0.04
MARB	K-means	5,582	0.40	0.09	5,275	0.52	0.08
	K-medoids		0.44	0.09		0.49	0.07
	PCN (Data)		0.54	0.12		0.59	0.06
	PCN (Cov)		0.49	0.10		0.54	0.07
	PCG (Data)		0.48	0.10		0.52	0.06
	PCG (Cov)		0.48	0.10		0.52	0.06
	Random		0.46	0.08		0.60	0.06
REA	K-means	5,582	0.32	0.07	5,115	0.55	0.08
	K-medoids		0.32	0.07		0.59	0.08
	PCN (Data)		0.38	0.06		0.62	0.07
	PCN (Cov)		0.38	0.06		0.64	0.07
	PCG (Data)		0.41	0.08		0.63	0.07
	PCG (Cov)		0.43	0.08		0.64	0.07
	Random		0.57	0.11		0.67	0.07

¹BWT = birth weight; YWT = yearling weight; MARB = marbling; REA = rib-eye area.

²K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix (**A** = Data); PCN (Cov) = principal component clustering using a numerator relationship matrix (**A** = Covariance matrix); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Data); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix (**G** = Covariance matrix); random = random clustering.

³Adjusted phenotypes for MARB and REA were the ultrasonically measured intramuscular fat percentage and rib-eye area, respectively.

⁴DEBV = deregressed estimated breeding value.

⁵ $(r_{g,y})$ = genetic correlation between GBV and either adjusted phenotype or DEBV.

⁶SE = average standard error across folds.

sire when averaged across replicates and genotyping scenarios. In the simulated data, the minimum number of progeny an animal could sire was 20 and it was assumed that all of them had a phenotype recorded. In contrast, in the Red Angus data, the sires included in the analysis ranged from having 1 to 822 progeny, leading to large differences in accuracy of EBV and necessitating deregression. Thus, the accuracy of EBV of the simulated animals was greater on average, and more homogeneous, than that of animals in the Red Angus data. Consequently, the deregression process did not aid in delineating information content in simulated data in the same fashion as in the real data.

For Red Angus, the average estimated regression coefficients of DEBV on GBV across clustering methods were 0.93, 0.90, 0.91, and 0.74 for BWT, MARB, REA, and YWT, respectively. The average estimated regression coefficients of adjusted phenotype on GBV across clustering method were 0.97, 0.98, 0.89, and 1.02 for BWT, MARB, REA, and YWT, respectively. Overall, the estimated regression coefficients of DEBV on GBV were lower than those of adjusted phenotypes. This pattern was also observed within the simulated data across both genotyping methods. However, estimated regression coefficients were more conservative when selective genotyping was used.

Table 6. Average estimated and true accuracy values and standard errors across all 5 simulations for cross validation.

Selection strategy ¹	Response variable ²	Clustering method ³	$(r_{g,y})^4$	SE ⁵	$r_{g,EBV}^6$	SE
Random selection	Phenotype	K-means	0.81	0.009	0.77	0.007
		K-medoids	0.81	0.016	0.78	0.010
		PCN (Data)	0.82	0.011	0.78	0.008
		PCN (Cov)	0.83	0.013	0.77	0.010
		PCG (Data)	0.84	0.016	0.78	0.009
		PCG (Cov)	0.84	0.015	0.78	0.008
		Random	0.85	0.015	0.80	0.008
		Random	0.90	0.010	0.82	0.007
	DEBV	K-means	0.84	0.009	0.78	0.006
		K-medoids	0.86	0.008	0.79	0.008
		PCN (Data)	0.86	0.009	0.80	0.006
		PCN (Cov)	0.85	0.011	0.79	0.008
		PCG (Data)	0.85	0.013	0.79	0.008
		PCG (Cov)	0.86	0.011	0.80	0.007
		Random	0.90	0.010	0.82	0.007
		Random	0.51	0.016	0.71	0.006
Top EBV	Phenotype	K-means	0.49	0.015	0.60	0.008
		K-medoids	0.47	0.012	0.61	0.005
		PCN (Data)	0.49	0.016	0.62	0.007
		PCN (Cov)	0.51	0.006	0.62	0.006
		PCG (Data)	0.49	0.017	0.62	0.007
		PCG (Cov)	0.49	0.011	0.62	0.006
		Random	0.52	0.020	0.64	0.006
		Random	0.51	0.016	0.71	0.006
	DEBV	K-means	0.45	0.024	0.66	0.009
		K-medoids	0.47	0.019	0.66	0.006
		PCN (Data)	0.48	0.019	0.67	0.006
		PCN (Cov)	0.44	0.018	0.68	0.004
		PCG (Data)	0.47	0.021	0.67	0.004
		PCG (Cov)	0.48	0.018	0.68	0.004
		Random	0.51	0.016	0.71	0.006
		Random	0.51	0.016	0.71	0.006

¹Random selection = 5,000 animals randomly chosen across all 10 generations; top EBV = 500 individuals from each of the 10 generations selected.

²Phenotype = raw phenotype; DEBV = deregressed estimated breeding value.

³K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix ($\mathbf{A}$ = Data); PCN (Cov) = principal component clustering using a numerator relationship matrix ($\mathbf{A}$ = Covariance matrix); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix ($\mathbf{G}$ = Data); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix ($\mathbf{G}$ = Covariance matrix); random = random clustering.

⁴ $(r_{g,y})$ = genetic correlation between GBV and either phenotype or DEBV.

⁵SE = standard deviation of correlations across replicates divided by the square root of the number of replicates.

⁶ $r_{g,EBV}$ = residual correlations between GBV and true breeding value including generation, fold, and generation * fold in the model.

Studies within other species have shown that choice of response variable can have an impact on prediction accuracy. Daetwyler et al. (2012) found phenotypes adjusted for fixed and breed effects led to higher estimated accuracies than nonadjusted phenotypes in sheep. van der Werf et al. (2010), in regards to a sheep information nucleus, stated that if an accurate EBV is used rather than a phenotype in training, it is like using a phenotype with higher heritability, in which the heritability of a trait also has an effect on the prediction accuracy (e.g., Goddard and Hayes, 2009). Guo et al. (2010) found that using EBV rather than daughter yield deviation (DYD) in simulation led to more reliable

predictions. Additionally, deregressed EBV have led to higher reliabilities of GBV than when EBV were used as response variables in pigs (Ostersen et al., 2011).

Forward in time validation was explored using simulation by using differing amounts of generational information to estimate prediction accuracy. Training sets included animals born in generations 6 to 11, 6 to 12, 6 to 13, and 6 to 14 to predict animals born in generation 15. The mean estimated and true accuracies of GBV for forward validation are presented in Table 8. As animals in generations closer to the selection candidates were included in the training set, the estimated accuracy increased.

Table 7. Average estimated and true regression coefficients and standard errors across all 5 simulations for cross validation

Selection strategy ¹	Response variable ²	Clustering method ³	$b_{Y,\hat{g}}$ ⁴	SE ⁵	$b_{TBV,\hat{g}}$ ⁶	SE
Random selection	Phenotype	K-means	0.91	0.005	0.85	0.008
		K-medoids	0.92	0.011	0.83	0.010
		PCN (Data)	0.92	0.004	0.86	0.005
		PCN (Cov)	0.90	0.008	0.86	0.006
		PCG (Data)	0.92	0.006	0.86	0.007
		PCG (Cov)	0.92	0.005	0.85	0.005
		Random	0.91	0.012	0.85	0.008
	DEBV	K-means	0.87	0.005	0.86	0.008
		K-medoids	0.86	0.008	0.84	0.009
		PCN (Data)	0.88	0.008	0.87	0.004
		PCN (Cov)	0.88	0.008	0.87	0.006
		PCG (Data)	0.89	0.006	0.87	0.006
		PCG (Cov)	0.89	0.007	0.87	0.005
		Random	0.90	0.010	0.86	0.007
Top EBV	Phenotype	K-means	0.55	0.018	0.73	0.013
		K-medoids	0.54	0.009	0.70	0.007
		PCN (Data)	0.54	0.014	0.74	0.009
		PCN (Cov)	0.53	0.005	0.75	0.014
		PCG (Data)	0.55	0.025	0.73	0.013
		PCG (Cov)	0.55	0.018	0.73	0.012
		Random	0.57	0.014	0.75	0.009
	DEBV	K-means	0.42	0.022	0.78	0.007
		K-medoids	0.44	0.018	0.75	0.008
		PCN (Data)	0.46	0.019	0.80	0.003
		PCN (Cov)	0.41	0.016	0.81	0.007
		PCG (Data)	0.44	0.024	0.78	0.005
		PCG (Cov)	0.44	0.019	0.78	0.004
		Random	0.46	0.016	0.83	0.006

¹Random selection = 5,000 animals randomly chosen across all 10 generations; top EBV = 500 individuals from each of the 10 generations selected.

²Phenotype = raw phenotype; DEBV = deregressed estimated breeding value.

³K-means = clustering based on k-means using the numerator relationship matrix; K-medoid = clustering based on k-medoids using the numerator relationship matrix; PCN (Data) = principal component clustering using a numerator relationship matrix ($A = \text{Data}$); PCN (Cov) = principal component clustering using a numerator relationship matrix ($A = \text{Covariance matrix}$); PCG (Data) = principal component clustering using an identical-by-state genomic relationship matrix ($G = \text{Data}$); PCG (Cov) = principal component clustering using an identical-by-state genomic relationship matrix ($G = \text{Covariance matrix}$); random = random clustering.

⁴ $b_{Y,\hat{g}}$ = regression coefficient of either phenotype or DEBV on GBV.

⁵SE = standard deviation of correlations across replicates divided by the square root of the number of replicates.

⁶ $b_{TBV,\hat{g}}$ = regression coefficient of true breeding value on GBV and including generation, fold, and generation * fold in model.

The accuracy of GBV is affected by the relationships between the training and evaluation sets. An increase in GBV accuracy as validation sets were generationally closer, thus more related, to testing sets was also found in other studies including Clark et al. (2011), and Pszczola et al. (2012) using simulated data, and Wolc et al. (2011) in a brown-egg layer line of chickens. As seen previously, the differences between estimated accuracy and true accuracy were negligible when using DEBV or phenotypes ($P = 0.318$ and $P = 0.178$ for random genotyping and selection of the top 25% of animals for genotyping, respectively). The slight

differences between estimated accuracy and true accuracy with DEBV as compared to phenotypes, although not statistically significant, suggest that the marker effects estimated from DEBV were more reliable for predicting the genetic merit of an animal.

Table 9 contains the regression coefficients of phenotype or DEBV on GBV as well as the regression coefficients of TBV on GBV. Smaller differences between the estimated and true regression coefficients were observed for DEBV than for phenotypes. These small differences between the estimated and true regression coefficients of DEBV

Table 8. Average estimated and true accuracy values and standard errors across all 5 simulations for forward validation

Selection strategy ¹	Response variable ²	Training population (generations) ³	$(r_{g,y})^4$	SE ⁵	$r_{g,TBV}^6$	SE
Random selection	Phenotype	6–11	0.77	0.024	0.77	0.008
		6–12	0.82	0.032	0.79	0.008
		6–13	0.82	0.019	0.80	0.007
		6–14	0.87	0.024	0.81	0.009
	DEBV	6–11	0.78	0.015	0.77	0.008
		6–12	0.80	0.014	0.79	0.007
		6–13	0.82	0.014	0.80	0.009
		6–14	0.83	0.018	0.82	0.006
Top EBV	Phenotype	6–11	0.76	0.033	0.73	0.009
		6–12	0.76	0.024	0.74	0.005
		6–13	0.76	0.017	0.75	0.011
		6–14	0.79	0.023	0.77	0.014
	DEBV	6–11	0.75	0.020	0.74	0.010
		6–12	0.77	0.015	0.75	0.007
		6–13	0.79	0.006	0.78	0.009
		6–14	0.82	0.011	0.81	0.004

¹Random selection = 5,000 animals randomly chosen across all 10 generations; top EBV = 500 individuals from each of the 10 generations selected.

²Phenotype = raw phenotype; DEBV = deregressed estimated breeding value.

³Training population (generations) = discrete generations used for the training data set. Evaluation set was always generation 15.

⁴ $(r_{g,y})$ = genetic correlation between GBV and either phenotype or DEBV.

⁵SE = standard error of estimated correlations across replicates divided by the square root of the number of replicates.

⁶ $r_{g,TBV}$ = residual correlations between GBV and true breeding value and including intercept in model.

and phenotypes further suggest that the use of DEBV as a dependent variable generates more reliable estimates of the cumulative SNP effects.

Genotyping Strategy

Randomly selecting individuals to genotype led to higher estimated and true accuracies than selection of the top 25% of individuals. The average estimated accuracies across clustering methods were 0.83 and 0.86 for phenotype and DEBV, respectively, when animals were randomly chosen for genotyping. When the top 25% of individuals were chosen for genotyping, the average estimated accuracies across clustering methods were 0.49 and 0.47 for phenotype and DEBV, respectively. The estimated accuracies underestimated the true accuracies when animals were chosen to be genotyped at random but overestimated the true accuracies when there was selective genotyping. The mean of the absolute differences between estimated accuracy and the true accuracy was 0.06 and 0.16 for random genotyping and selective genotyping, respectively, illustrating more bias associated with the estimated accuracy when animals were selectively genotyped as compared to being genotyped at random.

In a simulation study, [Ehsani et al. \(2010\)](#) demonstrated that random genotyping leads to higher reliability of estimated GBV as compared to only genotyping the top individuals. These conclusions were also found by [Boligon et al. \(2012\)](#) who compared 5 genotyping strategies in a simulation of a population undergoing selection. Members of the reference population were chosen to be genotyped at random, top individuals based on yield deviations, bottom individuals, top and bottom individuals, or least related individuals. The prediction accuracy was assessed in the selection candidates—progeny of the reference population—and it was found that selection of the top and bottom individuals based on yield deviation led to the highest accuracy and lowest predictive mean square error (PMSE). Random selection led to higher accuracy and lower PMSE than selecting just the top individuals. Within a simulated dairy system, [Jiménez-Montero et al. \(2012\)](#) implemented 5 genotyping strategies of dams in a forward in time validation. The 5 strategies included random, top and bottom individuals based on yield deviation values, top and bottom individuals based on EBV, highest yield deviation values, and highest EBV. The selection of top and bottom individuals led to the highest accuracies, followed by random, and the approaches

Table 9. Average estimated and true regression coefficients and standard errors across all 5 simulations for forward validation

Selection strategy ¹	Response variable ²	Training population (generations) ³	$b_{Y,g}$ ⁴	SE ⁵	$b_{TBV,g}$ ⁶	SE
Random selection	Phenotype	6–11	0.89	0.016	0.90	0.008
		6–12	0.91	0.011	0.92	0.013
		6–13	0.89	0.014	0.91	0.010
		6–14	0.89	0.008	0.91	0.002
	DEBV	6–11	0.86	0.011	0.90	0.007
		6–12	0.88	0.007	0.91	0.012
		6–13	0.88	0.007	0.92	0.005
		6–14	0.89	0.011	0.92	0.004
Top EBV	Phenotype	6–11	1.14	0.037	1.18	0.027
		6–12	1.10	0.026	1.12	0.022
		6–13	1.10	0.022	1.11	0.017
		6–14	1.06	0.011	1.09	0.010
	DEBV	6–11	1.10	0.020	1.15	0.014
		6–12	1.07	0.013	1.10	0.011
		6–13	1.07	0.019	1.11	0.015
		6–14	1.06	0.014	1.10	0.010

¹Random selection = 5,000 animals randomly chosen across all 10 generations; top EBV = 500 individuals from each of the 10 generations selected.

²Phenotype = raw phenotype; DEBV = deregressed estimated breeding value.

³Training population (generations) = discrete generations used for the training data set. Evaluation set was always generation 15.

⁴ $b_{Y,g}$ = regression coefficient of either phenotype or DEBV on GBV.

⁵SE = standard deviation of correlations across replicates divided by the square root of the number of replicates.

⁶ $b_{TBV,g}$ = residual regression coefficient of true breeding value on GBV and including intercept in model.

where only the highest individuals (based on yield deviations or EBV) were used produced the lowest accuracies. Additionally, the random approach and selection of top and bottom individuals led to the least amount of bias. The pattern of decreased accuracy when going from genotyping the extreme phenotypes (both top and bottom), to random genotyping, to genotyping top individuals was demonstrated within a Guernsey cattle herd when selective genotyping methods of females were compared to genotyping all females (Jenko et al., 2017). Pszczola et al. (2012) concluded animals that were randomly selected for the reference (i.e., training) population led to higher average reliabilities, measured as the squared correlation between the true and estimated BV, than when the reference population consisted of highly, moderately, or lowly related animals in simulation. Calus (2010) suggested a reference population with a wide range of genotypes and phenotypes would be optimal for reliable predictions.

The structure of the population simulated differed from that of the Red Angus data. With the simulation used in this study, the assumptions were that of a purebred cattle population, all relationships were known, there were no systematic effects, and phenotypes were measured without error. Full pedigrees of the Red Angus animals were not

available, which could have led to some of the discrepancies seen for the clustering methods between the real and simulated data because these clustering methods are dependent on the relatedness of animals to other individuals. Attempts to adjust for the systematic effects within the Red Angus phenotypes were made through preadjustments of sex, age, and breed composition as well as contemporary group deviations. Even with these adjustments, there is likely additional “noise” associated with the phenotypes because of other systematic effects that are hard to account for and the fact that phenotypes are often measured with some degree of error. Also, a systematic genotyping strategy is not currently employed in the cattle industry. Consequently, the collection of genotypes for Red Angus is likely somewhere between random selection and genotyping only the top individuals.

Overall, random clustering led to the highest estimated accuracy and k-means and k-medoids led to the lowest estimated accuracy within the Red Angus population. The estimated accuracies when DEBV were used to estimate SNP effects were generally higher, and associated with smaller standard errors, than when adjusted phenotypes were used. When simulation was used, random clustering led to the highest estimated accuracy while there was

no difference in the estimated accuracy between the other clustering methods. Based on the forward validation, use of DEBV to estimate marker effects was associated with less bias than phenotypes. Randomly genotyping animals to ensure representation of animals across the spectrum of EBV, not just choosing the animals with the top EBV, appeared to also be associated with the least amount of bias in the GBV and also the highest estimated accuracies.

ACKNOWLEDGMENTS

This project is based on research that was partially supported by the Nebraska Agricultural Experiment Station with funding from the Hatch Act (accession number 1011203) through the USDA National Institute of Food and Agriculture. The authors would like to thank the Red Angus Association of America for providing data used for this research effort.

LITERATURE CITED

- Beef Improvement Federation. 2010. Guidelines for uniform beef improvement programs. 9th ed. Beef Improv. Fed., Raleigh, NC. <http://www.beefimprovementorg> (Accessed 13 July 2017).
- Bodhireddy, 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
- Boligon, A. A., N. Long, L. G. Albuquerque, K. A. Weigel, D. Gianola, and G. J. Rosa. 2012. Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. *J. Anim. Sci.* 90:4716–4722. doi:10.2527/jas.2012-4857
- Bos_taurus_UMD_3.1 Genome Assembly NCBI. 2009. https://www.ncbi.nlm.nih.gov/assembly/GCF_000003055.4/#/st (Accessed 14 May 2018).
- Calus, M. P. L. 2010. Genomic breeding value prediction: methods and procedures. *Animal* 4:157–164. doi:10.1017/S1751731109991352
- Coster, A. 2012. pedigree: pedigree functions. R package version 1.4. <https://CRAN.R-project.org/package=pedigree>. (Accessed 1 September 2018)
- Chen, G. K., P. Marjoram, and J. D. Wall. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res.* 19:136–142. doi:10.1101/gr.083634.108
- Chen, L., F. Schenkel, M. Vinsky, D. H. Crews Jr, and C. Li. 2013. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *J. Anim. Sci.* 91:4669–4678. doi:10.2527/jas2013-5715
- Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* 44:4. doi:10.1186/1297-9686-44-4
- Clark, S. A., J. M. Hickey, and J. H. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. *Genet. Sel. Evol.* 43:18. doi:10.1186/1297-9686-43-18
- Daetwyler, H. D., A. A. Swan, J. H. van der Werf, and B. J. Hayes. 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genet. Sel. Evol.* 44:33. doi:10.1186/1297-9686-44-33
- Ehsani A., Janss L., and O. F. Christensen. 2010. Effects of selective genotyping on genomic prediction. In: 9th World Congress on Genetics Applied to Livestock Production. p. 145.
- Fraley, C., A. E. Rafter, T. B. Murphy, L. Scrucca. 2012. mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report No. 597. Department of Statistics, University of Washington, Seattle, WA, USA.
- Garrick, D. J., and R. L. Fernando. 2013. Implementing a QTL detection study (GWAS) using genomic prediction methodology. In: C. Gondro, J. H. van der Werf, and B. Hayes, editors, Genome-wide association studies and genomic prediction. Springer Series, Berlin. p. 275–298.
- 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.* 42:55. doi:10.1186/1297-9686-41-55
- Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2008. ASReml User Guide Release 3.0. VSN Int. Ltd., Hemel Hempstead, UK.
- Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Publ. Gr.* 10:381–391. doi:10.1038/nrg2575
- 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. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397. doi:10.1534/genetics.107.081190
- 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. doi:10.1186/1297-9686-42-5
- Howard, J. T., F. Tiezzi, J. E. Pryce, and C. Maltecca. 2017. GenoDiver: a combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits. *J. Anim. Breed. Genet.* 134:553–563. doi:10.1111/jbg.12277
- Hubert, L., and P. Arabie. 1985. Comparing partitions. *J. Classif.* 2:193–218. doi:10.1007/BF01908075
- Jenks, J., G. R. Wiggins, T. A. Cooper, S. A. E. Eaglen, W. G. L. Luff, M. Bichard, R. Pong-Wong, and J. A. Woolliams. 2017. Cow genotyping strategies for genomic selection in a small dairy cattle population. *J. Dairy Sci.* 100:439–452. doi:10.3168/jds.2016-11479
- Jiménez-Montero, J. A., O. González-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. *Anim. Anim. Consort.* 6:1216–1224. doi:10.1017/S1751731112000341
- 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, et al. 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 genotypes. *J. Anim. Sci.* 88:544–551. doi:10.2527/jas.2009-2064
- Lee, J., S. D. Kachman, and M. L. Spangler. 2017. The impact of training strategies on the accuracy of genomic predictors in United States Red Angus cattle. *J. Anim. Sci.* 95:3406–3414. doi:10.2527/jas.2017.1604
- Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, et al. 2014. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. *J. Dairy Sci.* 97:3200–3212. doi:10.3168/jds.2013-7745
- Liu, T., H. Qu, C. Luo, D. Shu, J. Wang, M. S. Lund, and G. Su. 2014. Accuracy of genomic prediction for growth and carcass traits in Chinese triple-yellow chickens. *BMC Genet.* 15:110. doi:10.1186/s12863-014-0110-y
- Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. Meuwissen. 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183:1119–1126. doi:10.1534/genetics.109.107391
- Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2018. cluster: cluster analysis basics and extensions. R package version 2.0.7-1. <https://cran.r-project.org/web/packages/cluster/index.html> (Accessed 1 September 2018).
- Ostensen, 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
- Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95:389–400. doi:10.3168/jds.2011-4338
- R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. (Accessed 1 September 2018)
- de Roos, A. P., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503–1512. doi:10.1534/genetics.107.084301
- 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, et al. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-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
- Saatchi, M., J. Ward, and D. J. Garrick. 2013. Accuracies of direct genomic breeding values in Hereford beef cattle using national or international training populations. *J. Anim. Sci.* 91:1538–1551. doi:10.2527/jas.2012-5593
- 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
- Thallman, R. M., K. J. Hanford, R. L. Quass, S. D. Kachman, R. J. Templeman, R. L. Fernando, L. A. Kuehn, E. J. Pollak. 2009. Estimation of the proportion of genetic variation accounted for by DNA tests. *Proceedings of the Beef Improvement Federation 41st Annual Research Symposium and Annual Meeting: April 30–May 3, 2009: Sacramento, CA.* p. 184–209.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. doi:10.3168/jds.2007-0980
- Weber, K. L., R. M. Thallman, J. W. Keele, W. M. Snelling, G. L. Bennett, T. P. Smith, T. G. McDanel, M. F. Allan, A. L. Van Eenennaam, and L. A. Kuehn. 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. *J. Anim. Sci.* 90:4177–4190. doi:10.2527/jas.2011-4586
- van der Werf, J. H. J., B. P. Kinghorn, and R. G. Banks. 2010. Design and role of an information nucleus in sheep breeding programs. *Anim. Prod. Sci.* 50:998–1003. doi:10.1071/an10151
- Wolc, A., J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, D. Habier, R. Fernando, D. J. Garrick, and J. C. Dekkers. 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet. Sel. Evol.* 43:23. doi:10.1186/1297-9686-43-23