12-2020

Improving the Accuracy of Genomic Predictions: Investigation of Training Methods and Data Pooling

Johnna Baller
University of Nebraska-Lincoln, jballer2@unl.edu

IMPROVING THE ACCURACY OF GENOMIC PREDICTIONS: INVESTIGATION OF TRAINING METHODS AND DATA POOLING

by

Johnna Lynn Baller

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Animal Science

(Animal Breeding and Genetics)

Under the Supervision of Professor Matthew L. Spangler

Lincoln, Nebraska

December, 2020

IMPROVING THE ACCURACY OF GENOMIC PREDICTIONS: INVESTIGATION OF TRAINING METHODS AND DATA POOLING

Johnna Lynn Baller, Ph.D.

University of Nebraska, 2020

Advisor: Matthew L. Spangler

One of the primary factors in the response to selection is the accuracy of selection. This study focused on methodologies to predict breeding values (BV) accurately within multi- and single-step genomic evaluations. Factors including cross-validation methods, dependent variables, and genotyping strategies were assessed on the accuracy of genomic BV while using multi-step prediction in real and simulated data. In both cases, random clustering led to largest estimated accuracies compared to clusters based on k-means, k-medoids, and principle component analysis, but differences in bias were not detected. Using deregressed estimated BV (EBV) to estimate SNP effects led to larger accuracies and smaller standard errors than adjusted phenotypes. Randomly genotyping animals instead of selectively genotyping the top 25% was associated with highest accuracies and least amount of bias.

Genetic improvement of economically relevant traits (ERT) should be the goal of breeding programs. Although generally absent in seedstock herds, ERT are routinely collected within commercial sectors; therefore, pooling data was proposed to include commercial information in a cost-effective manner. Pooling involves collecting tissue

samples from a group of animals and then combining the DNA to be genotyped as one. The accuracy of EBV when pooled data were used within single-step analysis was investigated through simulation. For a single trait, pool sizes of 2, 10, 20 or 50 did not generally lead to differences in EBV accuracy compared to using individual data when pools were constructed to minimize phenotypic variation. Low accuracy sires benefited the most from pooling, while EBV for the pools could be used for management purposes. For a bivariate analysis, pool sizes of at least 20 were recommended in combination with minimizing phenotypic variation. Additionally, if pools were constructed to minimize phenotypic variation, pooling could be used across a range of genetic correlations (0.1, 0.4, and 0.7) and ways in which missing values arise (randomly missing records or sequential culling). Collectively, these results suggest pooling can be used to include commercial data within genetic evaluations.

Acknowledgements

I would first and foremost like to thank my parents, Jimmer and Jerri, and sister, Codi, for the support they have provided me. They have always encouraged me to pursue my dreams, no matter how big or small, and they have been there for me every step of the way, helping in any way they could. Thank you, Kyle, for your motivation and for always putting up with me.

I would also like to thank many people within the University of Nebraska system. It has been a privilege to be a part of the UNL atmosphere for over 5 years and across two different departments. I would like to thank Dr. Matt Spangler for his guidance through my doctoral program. All of our discussions regarding research projects and the beef industry extended my education beyond the classroom. I would also like to thank my committee members Dr. Steve Kachman, Dr. Larry Kuehn, and Dr. Ron Lewis for their input on my dissertation. All three have been instrumental in bettering my education, my projects, and understanding of breeding and genetics. I am very grateful for Sherri Pitchie and all of her help with travel and conference arrangements, scheduling, and paperwork.

I would not have been able to make it through this Ph.D. journey without the help and support of the other students, visiting researchers, and post-docs within the Animal Breeding and Genetic group. Lastly, I would like to thank Jessica Hauschild, Kelsey Karnik, and Tiffany Sunderland for encouragement, advice, statistics help, and most of all friendship.

Preface

Chapter 2 has been published in Journal of Animal Science (Baller J. L., J. T. Howard, S. D. Kachman, and M. L. Spangler. 2019. The impact of clustering methods for cross-validation, choice of phenotypes, and genotyping strategies on the accuracy of genomic predictions. J. Anim. Sci. 97:1534–1549. doi:10.1093/ jas/skz055.)

Chapter 3 has been published in Journal of Animal Science (Baller, J. L., S. D. Kachman, L. A. Kuehn, and M. L. Spangler. 2020. Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J. Anim. Sci. 98. doi:10.1093/jas/skaa184.)

The content discussed in Chapter 4 is expected to be included in a manuscript that is currently in preparation for publication under the lead of Johnna L. Baller. (Baller, J. L., S. D. Kachman, L. A. Kuehn, and M. L. Spangler. “Using pooled data for genomic prediction in a bivariate framework with missing data” (in preparation for submission to Journal of Animal Breeding and Genetics).)

Table of Contents

List of Tables

List of Figures

Chapter 1

LITERATURE REVIEW

1.1 Introduction

Animal populations have changed over time due to artificial selection, and the tools used to help aid in the selection of animals have continued to evolve. Animals were first appraised by phenotypic selection, in which animals were judged based on their own performance for traits of interest. Later, using Henderson's mixed model equations (Henderson, 1975), an animal's own performance records were combined with pedigree information and the performance of the animal's relatives using best linear unbiased prediction (BLUP). With the inclusion of DNA information in the prediction of estimated breeding values (EBV), more accurate selection decisions could be taken before an individual even produced progeny, therefore, increasing overall accuracy of selection and shortening the generation interval, both of which in turn increase the rate of genetic change per year.

Traits for which EBV are calculated can include economically relevant traits (ERT) that directly affect the profitability of a commercial system because they relate to either a cost or source of income (Golden et al., 2000). It is important to note that ERT are only measured within the commercial sectors of livestock industries. Thus, it is important to include data from the commercial segments in genetic evaluations. The lack of integration in the beef industry makes including commercial animal data into genetic evaluations challenging.

1.2 Brief history of genetic evaluations

1.2.1. Estimation of Breeding Values

A breeding value (BV) is a measure of an animals' additive genetic merit deviated from a population mean for a given trait. A true BV is never known but can be approximated using EBV. It is expected that half of an animal's genetic merit will be passed to its progeny, which in the US beef industry is known as an expected progeny difference (EPD). An EPD is one-half of an EBV. In a simplistic case, an EPD of an individual can be thought of as the average of the genetic merit of its parents: $EPD_{Individual} = \frac{1}{2}EPD_{Sire} + \frac{1}{2}EPD_{Dam}$. However, because genes are randomly sampled during the formation of gametes, the offspring do not inherit exactly one-half of the cumulative genetic merit of each of the parents – it could be more, or it could be less. This random sampling of genes is known as Mendelian sampling and is used to describe the deviation from the parental average. The genetic merit can now be described as: $EPD_{Individual} = \frac{1}{2}EPD_{Sire} + \frac{1}{2}EPD_{Dam} + \varphi$ where φ is the Mendelian sampling term. Genomic data can be used to help capture and quantify the Mendelian sampling term, leading to more accurate EPD of individuals.

1.2.2 Introduction of genomic data

A quantitative trait locus (QTL) is a region within a DNA sequence that influences the phenotype of a particular trait of interest. However, the exact causative mutation within the QTL is often unknown. Spread across the genome are additional genetic markers that can be identified and genotyped. Sometimes these genetic markers and a QTL allele are inherited together more often than expected, leading to linkage disequilibrium (LD). The incorporation of a few direct (known QTL) and indirect

markers (those in LD with QTL but not necessarily a causative marker) into the traditional selection decisions is known as marker assisted selection (MAS).

One of the first instances of identifying the genotype of animals was documented by Bouw et al. (1974) in which the blood groups were used as markers. This early work in blood serums and other protein work seemed impractical because of the lack of polymorphisms and genome coverage associated with these structures (Drinkwater and Hetzel, 1991). Microsatellites, repetitive sequences of DNA in which the unit consists of one to six base pairs, were recognized as a useful tool, especially compared to protein markers, as a means for studying genetic relationships in cattle (Arranz et al., 1996). Microsatellites were also used for association studies for quantitative traits (Georges et al., 1993; Napolitan et al., 1996) to be used for MAS. Another goal of microsatellite usage was within family linkage tracking (e.g. Bowling et al., 1997; Glowatzki-Mullis et al., 1995; Heyen et al., 1997)

The downfall of using microsatellites in the 1990s was that very few markers were initially identified and used with MAS. These markers generally explained only a small proportion of the additive variation of traits of interest as only a few markers were statistically significant during testing. With qualitative traits, in which one or very few genes determine the outcome of the phenotype, this was not a problem (e.g., double muscling in cattle (Grobet et. al., 1997; McPherron and Lee, 1997) and DGAT1 that affects milk-fat content (Grisart et al., 2001)). However, most traits of interest are complex in nature, meaning that the few QTL and markers used in MAS were inadequate to explain a large proportion of variation in the traits. To combat this problem, a method called fine mapping was employed in order to increase marker density around the

previously documented QTL to better enhance QTL mapping and provide for better effect estimates to be used in MAS (Pollak et al., 2015).

As technology improved, the genotyping of single nucleotide polymorphisms (SNP), single base changes in the DNA sequence, became available. Single nucleotide polymorphisms are easy to evaluate because they are bi-allelic in nature and are spread across the genome (Fan et al., 2010). Meuwissen et al. (2001), proposed genomic selection (GS), an extension of MAS, where the effects of thousands of markers spread across the genome are estimated and then summed up to predict an animal's genetic merit. Because the true QTL are not likely to be genotyped, the premise behind GS is the LD between SNP and QTL. As the density of SNP genotyped increases, the probability that the marker and QTL will be in close proximity to each other and even be in LD rises. With high density SNP panels, it is expected that at least one SNP is in LD with a QTL (Hayes and Goddard, 2010). Therefore, if the genotype of an animal at the SNP marker is known and is in LD with a QTL, it may be possible to predict the breeding value at that locus, and cumulatively across all loci.

1.3 Methods for genomic prediction

Since the discovery of so many SNP, the relative cost effectiveness of genotyping animals, and the introduction of GS, genomic information has seen widespread use in livestock evaluations (Meuwissen et al., 2013).

1.3.1 Best Linear Unbiased Predictions

Relationships between individuals can be quantified using pedigrees, which are then summarized by a numerator relationship matrix (**A**). These are the expected relationships between two individuals. For example, a pair of full-sibs are expected to share one-half of their genome while the relationship between an individual and their grandparent is expected to be one-quarter. This relationship matrix would then be used in BLUP evaluations, leading to estimates deemed as "traditional EBV", commonly called PBLUP. Assume observations are modeled by $y = \mathbf{X}b + \mathbf{Z}u + e$ where y is a vector of observations, b is a vector of fixed effects, u is a vector of random genetic effects, **X** and **Z** are incidence matrices, and e is a vector of random residuals. The solutions for the fixed and random effects can be obtained by solving

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z + G^{-1}} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}}y \\ \mathbf{Z'R^{-1}}y \end{bmatrix}.$$

It is also assumed that $V(u) = \mathbf{G} = \mathbf{A}\sigma_a^2$ and $V(e) = \mathbf{R} = \mathbf{I}\sigma_e^2$. Substituting in these variances and multiplying by σ_e^2 throughout leads to

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A^{-1}} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} \mathbf{X'}y \\ \mathbf{Z'}y \end{bmatrix}$$

where λ is equal to $\frac{\sigma_e^2}{\sigma_a^2}$. It is important to note that this lambda simplification only applies to single trait case while the original mixed model equations above are generalizable.

However, it is known that realized relationships can deviate from these expectations. Genomic relationships can be calculated as the covariance of the genetic effects of two individuals, where the genetic effects are measured as the genotypes of the individuals. The resulting genomic relationship matrix (**G**) can be easily substituted into BLUP evaluations, resulting in genomic best linear unbiased prediction (**GBLUP**) in which the random genetic effects are now genomic EBV (**GEBV**). With the inclusion of

$\mathbf{G}$ instead of $\mathbf{A}$, $V(u) = \mathbf{G} = \mathbf{G}\sigma_g^2$, and λ is equal to $\frac{\sigma_e^2}{\sigma_g^2}$. The assumptions of GBLUP are an infinitesimal model, meaning that there a very large number of loci each with small effects that influence a quantitative trait. Because the $\mathbf{G}$ matrix can partially account for Mendelian sampling and pedigrees are oftentimes missing or incorrect, genomic relationships provide more accurate estimates of relationship and thus increased accuracy of EBV (Hayes et al., 2009).

1.3.2 Regression

In some cases, it is appropriate to assume the not all SNP have the same effect, thus it is useful to estimate the effect of each locus. Ordinary least squares (OLS), or the regression of phenotype on genotype, have been used in order to estimate the effects of each SNP. The model can be described as $y = \mathbf{X}\beta + e$ where y is the vector of observations, β is the vector of effect sizes of the SNP, $\mathbf{X}$ is a *nxk* matrix denoting the genotype of n^{th} individual at the k^{th} SNP, and e is the random residual. The SNP effects are estimated by $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$.

More than two million SNP have been found in the beef genome (Zimin et al., 2009) through the discovery methods such as whole genome sequencing, HapMap projects, and completing reduced representation library (RRL) sequencing (Fan et al., 2010). With the identification of millions of SNP, it has been possible to develop a variety of commercially available high-density panels that range from 3k to 777k.
As the number of available SNP become much larger than the number of individuals available to train the model (estimate SNP effects), problems arise with unique solutions and poor prediction (Zhang & Smith, 1992). To overcome this, subsets of SNP data could

be used in which some of the SNP are excluded from the analysis. There have been many ways of choosing the subsets of SNP including best subset method or a forward/backward stepwise method (Breiman, 1995), or more recently machine learning methods (Li et al., 2018).

Another alternative to using a subset of SNP is to use a technique called ridge regression (Whittaker et al., 2000). All SNP are included in the model, but the effects are estimated by $\hat{\beta} = (\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}'y$ where the λ parameter shrinks the estimates towards zero. Effects attained by ridge regression were associated with smaller standard errors than those from a subset of markers, thus leading to more reliable response to selection (Whittaker et al., 2000). Ridge regression is equivalent to GBLUP when the markers are assumed to be independent draws from a normal distribution. Then, λ is equal to $\frac{\sigma_e^2}{\sigma_g^2}$ (Piepho, 2009).

1.3.3 Bayesian

Another technique commonly used to obtain estimates of breeding values and parameter estimates is Bayesian analysis. Bayesian methods are attractive to animal breeders in many ways. First, these types of models have the ability to overcome the "small n, large p" problem where the number of markers can exceed the number of observations available (Gianola et al., 2009). It may be unreasonable to assume that all markers have an effect. Mixture models are easily implemented within a Bayesian framework in which one can assume some distribution is actually a mixture of two or more distributions. This may be warranted when it is reasonable to assume some markers have a null effect while others are non-zero. These types of models are also called

variable selection models because the markers with a non-zero effect are "selected" to be in the model. Additionally, those effects don't have to be distributed Normally.

The Bayes Theorem states that the probability of two events, θ and X, occurring together is: $P(\theta, X) = P(\theta)P(X|\theta) = P(X)P(\theta|X)$, where $P(X|\theta)$ and $P(\theta|X)$ are conditional probabilities. From this relationship, it can be stated that $P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)}$ which is proportional to $P(\theta)P(X|\theta)$ since $P(X)$ does not depend on θ. Commonly, $P(\theta)$ is known as the prior distribution of θ, $P(X|\theta)$ is the likelihood of θ, and $P(\theta|X)$ is the posterior distribution of θ given X. For example, let X represent a set of phenotypic and genotypic observations and θ represent the set of parameters to estimate. In animal breeding, these parameters include all variance components, "fixed effects" including regression coefficients, and marker effects. Thus, all of the unknown parameters, θ, are treated as random and have their own distribution phenotypic and genotypic observations are a function of the prior. The posterior distribution is the conditional distribution of the parameters given the phenotypes and genotypes. Unfortunately, to make inferences of the parameters from this posterior distribution usually requires high-dimensional integrals that often-times are not expressed in a closed form (Garrick et al., 2014). Instead, the Markov Chain Monte Carlo (MCMC) sampling technique can be used to draw samples from the posterior distribution to estimate posterior means and variances. The Gibbs sampler is one of the most widely used MCMC techniques in animal breeding (Garrick et al., 2014).

Many variations of the Bayesian methods are available, often referred to as the "Bayesian alphabet". Meuwissen et al. (2001) proposed a model, called BayesA, that assumes all of the markers have an effect. It also assumes that the prior distribution of

these effects is Normal, with each marker having its own variance modeled with a scaled inverse chi-square distribution. BayesB, also proposed by Meuwissen et al. (2001), is a mixture model. This model allows for a proportion of the markers (π) to have a null-effect while another proportion of the markers ($1 - \pi$) to have a non-zero effect. The proportion of markers that do have an effect is assumed to have a distribution equal to those of BayesA. The proportion of makers that have a null-effect (π) is chosen *a priori*. Habier et al. (2011) developed a model, BayesC, which is very similar to BayesB except that there is a common marker variance instead of marker-specific variances. BayesC can be extended to a model known as BayesCπ. This model assumes π is also unknown and is also estimated from the data (Habier et al., 2011). Another extension to BayesC is BayesC0 in which π is set to zero, and thus all markers have a non-zero effect. Bayesian ridge regression assumes the same genetic variance for all markers, which is the same model as BayesC0, and equivalent to GBLUP. While each of these methods are similar, the most important aspect about the differences is how they assume the variances of the marker affects are distributed, the presence or absence of the mixture model, and the degree of the mixture.

1.4 Methods for combining pedigree and genomic data

Two methods, multi-step and single-step, have been used to combine all information into one estimate.

1.4.1 Multi-step methods

In multi-step methods, the SNP effects are estimated using all available animals with genotypes and a phenotype. The resulting marker effects are multiplied by the SNP

genotypes of the animals and summed across all markers, resulting in the estimate of molecular breeding values (MBV). The estimated marker effects can also be applied to animals with genotypes but no phenotypes - usually young selection candidates. The MBV are then combined with traditional EBV to attain a final genome enhanced EBV (GE-EBV).

The variance/covariance estimates and the genetic correlation between the MBV and phenotypes are very important for the combination of the MBV with EBV. To estimate this correlation, a method called cross-validation is used. This technique is often used to assess the predictive ability of a model on data that was not used to estimate the model. In other words, it tests how well a model works in practice. A population of animals that have known genotypes and phenotypes are divided into *k* independent groups, where *k* is specified *a priori*. Then, one of the *k* groups is deemed as a validation set and the remaining *k-1* groups are specified as a training set. The training set is used to estimate SNP marker effects. The resulting marker effects are applied to the validation set to estimate the MBV of the animals. Using a bivariate animal model, genetic variance and covariances are estimated to calculate the genetic correlation between the predicted MBV and the phenotype of the animals within the validation set (Kachman et al., 2013). This whole sequence of specifying training and validation sets to estimating the genetic correlation is repeated *k* times, such that each group is used once and only once for validation. The genetic correlation is then averaged over the *k* cross-validations for the final genetic correlation. This genetic correlation is also known as accuracy. An alternative to estimating the genetic correlation for each fold separately and then averaging over the folds would be to estimate the genetic correlation across all folds

using a bivariate model and fit a random effect of fold. The square of the genetic correlations (r_g^2) estimate the proportion of genetic variance explained by the MBV (Thallman et al., 2009).

In order to combine the MBV and EBV, two approaches have typically been used. The first method is called "blending", in which MBV and EBV are weighted proportional to their reliabilities. These weightings will be different for each trait depending on r_g^2, and for each animal depending on its EBV reliability (Garrick and Saatchi, 2013). Traditionally, blending was done post-evaluation, and so only genotyped animals were affected (Spangler, 2013). The American Hereford Association released its first GE-EPDs, using the blending approach, in the fall of 2012 (Ward, 2013). The "correlated trait approach", introduced by Kachman (2008), incorporated MBV as a correlated indicator trait within a multi-trait model. This approach is particularly attractive because the predictions of animals that were in the pedigree but not genotyped were still influenced by the genomic information (Spangler, 2013). MacNeil et al. (2010) utilized the correlated trait approach to incorporate ultrasound intramuscular fat (IMF) and MBV indicator traits for the prediction of a marbling EBV for the American Angus Association.

1.4.2 Single-step methods

Two methods qualify as single-step methods for combing genomic and pedigree data. The first is single-step genomic best linear unbiased prediction (ss-GBLUP) method (Legarra et al., 2009). As described previously, relationships between animals can be derived in two ways. Pedigree-based relationships are included in the numerator

relationship matrix (**A**), which estimates the expected relationship between two individuals. Genomic relationships are included in the genomic relationship matrix (**G**), which are derived using the SNP markers individuals share in common. Previously, genotyped and non-genotyped animals were not included in the same model because methods did not exist to combine all the information into one relationship matrix for use in BLUP. Use of the ss-GBLUP combines phenotypic information as well as genotypic and pedigree-based relationships into one step in order to estimate GEBV. As with other BLUP methods, it is assumed the marker effects have a Normal distribution. During this process, the **A** and **G** matrices are combined in order to estimate the matrix **H** such that

$$\mathbf{H} = \begin{bmatrix} \mathbf{G}_{\mathrm{w}} & \mathbf{G}_{\mathrm{w}}\mathbf{A}_{22}^{-1}\mathbf{A}_{12} \\ \mathbf{A}^{\mathrm{T}}\mathbf{A}_{22}^{-1}\mathbf{G}_{\mathrm{w}} & \mathbf{A}_{11} + \mathbf{A}_{12}^{\mathrm{T}}\mathbf{A}_{22}^{-1}(\mathbf{G}_{\mathrm{w}} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{12} \end{bmatrix}$$

where $\mathbf{A}_{11}, \mathbf{A}_{12}$, and $\mathbf{A}_{22}$ are submatrices of **A** containing the relationships of among the non-genotyped animals, genotyped and non-genotyped animals, and genotyped animals, respectively, and $\mathbf{G}_{\mathrm{w}}$ is a weighted genomic relationship matrix such that

$$\mathbf{G}_{\mathrm{w}} = (1 - \mathrm{w})\mathbf{G} + \mathrm{w}\mathbf{A}_{22}$$

where **G** is computed as $\frac{\mathbf{MM}'}{2\Sigma p_i(1-p_i)}$, **M** is the centered genotype incidence matrix for individuals, p_i is the allelic frequency of the second allele of the *i*th SNP is the genomic relationship matrix (VanRaden, 2008), and w is the relative weight on the polygenic effect, or some small value (Christensen and Lund, 2010). The purpose of weighting **G** is to obtain a non-singular matrix and the weight has been suggested to be equal to 0.05 (VanRaden, 2008). Christensen et al. (2012) suggested **G** and $\mathbf{A}_{22}$ be compatible prior to weighting such that

$$\mathbf{G}_{\mathrm{a}} = \beta\mathbf{G} + \alpha$$

Where β and α can be found by solving the system of linear equations:

$$\overline{\text{diag}(\mathbf{G})}\beta + \alpha = \overline{\text{diag}(\mathbf{A}_{22})}$$

$$\bar{\mathbf{G}}\beta + \alpha = \overline{\mathbf{A}_{22}}$$

where the bars denote average values. The matrix $\mathbf{G}_a$ would then be weighted with $\mathbf{A}_{22}$to form $\mathbf{G}_w$. The final matrix $\mathbf{H}$ can be easily substituted into the BLUP evaluations, and the random genetic effects are again GEBV. However, solving the MME require the inverse of the relationship matrix. Aguilar et al. (2010) reduced the complexity of computing $\mathbf{H}^{-1}$ by giving

$$\mathbf{H^{-1}} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Matrix $\mathbf{G}$ is usually singular, and so the genomic matrix should again be compatible with $\mathbf{A}_{22}$ and weighted. The inverse of $\mathbf{G}$ is needed for the computation of $\mathbf{H}^{-1}$, but as the number of genotyped animals increases, the ability to compute $\mathbf{G^{-1}}$ becomes more expensive computationally.

To combat this problem, a few solutions have been proposed including an approximating $\mathbf{G}$ so that its inverse could be found efficiently (Faux et al., 2012) or obtaining the solutions of the MME without inverting $\mathbf{G}$ explicitly (Legarra and Ducrocq, 2012). Misztal et al. (2014) proposed the algorithm for proven and young (APY) that splits genotyped animals into core (proven) and non-core (young) groups and uses recursion to approximate $\mathbf{G^{-1}}$. For this method, the only direct inversion needed is for core animals, whereas all other coefficients are estimated by recursion (Lourenco et al., 2015). The resulting $\mathbf{G^{-1}}$ is a sparce matrix with non-zero coefficients in an "L" shape and diagonal elements. The core animals and core size can be determined in a variety of ways. Largest accuracies of GEBV, assessed as the correlation of GEBV and true

breeding value, in a simulation were attained by a core size determined by the largest eigenvalues explaining 98% of the variation in $\mathbf{G^{-1}}$ and when animals were randomly selected from all genotyped animals (Bradford et al., 2017). Other scenarios explored by Bradford et al. (2017) were core sizes equal to the largest number of eigenvalues explaining 90% and 95% of the variation in $\mathbf{G^{-1}}$ and core animals being sampled from distinct generations of parents and from the youngest animals without progeny. Overall, it was found that the core size and definition was robust, but became more important as pedigrees became more incomplete (Bradford et al., 2017). In an analysis using Holstein cows, Fragomeni et al. (2015) found the definition of animals as core or non-core does not necessarily matter (definitions of core animals were only sires, sire and cows, only cows, and only sires with 5 or more progeny), but the core size does matter when comparing the correlation of GEBV from ss-GBLUP direct inversion to GEBV from ss-GBLUP using the APY algorithm on almost 49,611 young animals (neither a bull or a cow that had records). Optimal core size was between 10,000 and 20,000 animals while the number of genotyped animals was 100,000 for this analysis.

Similarly, single-step Bayesian Regression (SSBR) also combines phenotype, genotype and pedigree information in one step (Fernando et al., 2014). One of the major disadvantages of ss-GBLUP is the inversion of relationship matrices, particularly **G,** even though work has been done to reduce this computational burden. However, SSBR does not need to invert the dense genomic relationship matrix. The model for SSBR is:

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \begin{bmatrix} \mathbf{X}_n \\ \mathbf{X}_g \end{bmatrix} \beta + \begin{bmatrix} \mathbf{Z}_n & 0 \\ 0 & \mathbf{Z}_g \end{bmatrix} \begin{bmatrix} \mathbf{M}_n \alpha + \epsilon \\ \mathbf{M}_g \alpha \end{bmatrix} + e$$

where y_n and y_g are vectors of phenotypes for non-genotyped and genotyped animals, $\mathbf{X}_n$ and $\mathbf{X}_g$ are incidence matrices for fixed effects for non-genotyped and genotyped animals,

β is a vector of fixed effects, $\mathbf{Z}_n$ and $\mathbf{Z}_g$ are incidence matrices that related the breeding values of non-genotyped and genotyped animals to the phenotypic values, $\mathbf{M}_n\alpha + \epsilon$ and $\mathbf{M}_g\alpha$ are the breeding values of non-genotyped and genotyped animals, $\mathbf{M}_g$ is the centered marker matrix for genotyped animals, $\mathbf{M}_n$ is the imputed marker matrix for non-genotyped animals, α is the vector of random marker effects, ϵ is a vector of imputation residuals, and e is a vector of residuals. In summary, the model is written in terms of two random factors – one for marker effects and one for the residual breeding values of non-genotyped animals. This algorithm involves predicting genotypes for the non-genotyped animals from their genotyped relatives using regression and then the residual breeding value accounts for the EBV information that is not explained by these predicted genotypes (Garrick et al., 2014). One of the advantages of this model is that only the $\mathbf{A}^{-1}$ is needed, which is computationally easy. Fernando et al. (2014) further explains centering the genotype matrix is not needed as long as an additional covariate is added to model the mean of the breeding values. During the computation of **G**, the allelic frequencies are needed. If selection has taken place, these frequencies should be estimated using founder animals, however, founder animal genotypes are usually not available. If the frequencies are estimated using the genotyped animals, the evaluation could be biased, especially in multi-breed evaluations (Fernando et al., 2014).

Using the assumptions of BayesC0 and assuming no APY, ssGBLUP and SSBR are equivalent models. Unlike ss-GBLUP, SSBR is not limited to normally distributed marker effects, it can be extended to other models such as t-distributed marker effects and with mixture models, depending on the prior used for the random marker effects. In beef cattle evaluations, ss-GBLUP has been utilized by the Angus Genetics Inc. while SSBR

has been utilized by International Genetic Solutions and the American Hereford Association (e.g. Misztal and Lourenco, 2018; Golden et al., 2018).

1.5 Commercial data in genetic evaluations

Genetic evaluations produce EBV for traits using data largely generated by the nucleus/seedstock sector of livestock industries. Some of these traits target economically relevant traits (ERT). By definition, true ERT are measured within the commercial sectors. Thus, the EBV produced using nucleus data are either for "presumed" ERT or indicator traits. Millions of records that represent the true ERT are recorded within the commercial industry every year. However, these records rarely make it into genetic evaluations because relationships that tie the commercial animals to the nucleus selection candidates are missing. Relationships between these groups exist, but pedigree information is often missing or incomplete. Nonetheless, inclusion of commercial data has enormous potential to increase the response to selection for traits that are economically important to the livestock industries. An optimal solution would be to collect the true ERT from commercial herds and estimate relationships between commercial animals and seedstock animals in an economical manner for use in genetic evaluations.

1.5.1. Economically relevant traits

Economically relevant traits are traits that directly affect the profitability of a commercial system because they relate to either a cost or source of income (Golden et al., 2000). Examples of ERT include, but are not limited to, weight at time of sale (e.g.

weaning weight direct, weaning weight maternal, carcass weight, salvage cow weight), calving ease, maintenance feed requirement, stayability, heifer pregnancy rate, tenderness, and days to finish (e.g. Golden et al., 2000). Enterprises may only identify a subset of these traits as ERT, which is specific to the production system. Take for example a producer who sells calves at weaning, and the price is determined by weight. An obvious ERT in this system would be weaning weight. However, if another producer determines profit based on carcass weight, weaning weight is no longer an ERT, but could be indicative of carcass weight. Thus, not all traits that are recorded directly affect profitability, but are instead considered indicator traits of the ERT. These indicator traits are genetically correlated with the ERT. In the latter example, the ERT would be carcass weight whereas weaning weight would be considered an indicator trait.

Even though indicator traits do not directly affect the overall profitability of an enterprise, they are measured because the associated ERT are hard to measure or are expressed later in life. Furthermore, most data collection and selection decisions usually take place in the seedstock sector of the beef industry (Garrick, 2018). This has resulted in the collection of phenotypes that are convenient and easy to validate in resulting progeny (Garrick, 2018). Because true ERT are only expressed in commercial animals, the data collected from seedstock animals represent presumed ERT. Additionally, many ERT such as disease susceptibility and survival cannot be collected within seedstock herds, due to increased health conditions and more rapid replacements rates, or there is a genetic by environmental interaction between these traits within the commercial and seedstock herds.

When breeding objectives are defined and selection decisions are taken based on those objectives, only ERT should be included in the decision-making process. In fact, when ERT and indicator traits are used in combination to attain the same selection decision for one trait, the accuracy of that decision is decreased (Golden et al., 2009; Enns, 2013). Oftentimes, merit of an animal is not defined by just one trait, rather a combination of multiple traits. To combine multiple traits into one succinct value to inform the overall genetic merit of an animal, selection indices can be used in order to correctly weight the information (Hazel, 1943). When creating a selection index, typically two sets of traits are needed: objective traits – the ERT defined in the breeding objective, and selection criteria – the traits that are actually measured. Ideally, selection criteria would consist entirely of ERT. Sometimes these ERT are not measured or readily available, and so indicator traits are used as selection criteria (Ochsner et al., 2017).

1.6 Examples from other species

Crossbred animals play an important role in the commercial sectors of some livestock industries (e.g. poultry and swine), but it is known that the same traits in commercial and purebred animals can be genetically different. Therefore, it is important to include commercial data into genomic evaluations, especially when the goal of purebred selection is to increase crossbred performance. Many methods have been evaluated; however, most of this research and implementation has been conducted within the swine and poultry industries where crossbreeding schemes are much more structured than with beef.

1.6.1 Crossbreeding

Crossbreeding has been used in commercial livestock production systems to exploit heterosis and breed complementarity. The goal of many selection programs is to maximize crossbred performance through purebred selection; however, traits that are recorded in purebred animals can be genetically different than those recorded in crossbred animals. Núñez-Dominguez et al. (1993) found the correlation of genetic expression between crossbred and purebred performance (r_{PC}) for growth traits averaged across progeny sired by three breeds of cattle (Angus, Hereford, and Polled Hereford) to be 0.93, 0.77, and 0.76 for weights at birth, 200 days, and 365 days, respectively. Newman et al. (2002) also found r_{PC} less than 1 for post-weaning growth and carcass traits using progeny from five sire breeds (Angus, Hereford, Shorthorn, Belmont Red, and Santa Gertrudis) mated to Brahman dams. These deviations of r_{PC} from 1 are likely to be caused by non-additive effects and genotype by environment interactions (Wei and van der Steen, 1991).

Historically most selection decisions have been taken in the purebred lines/breeds using primarily purebred data. Additionally, many economically relevant traits, such as disease susceptibility and survival, cannot be collected in purebred herds. This is especially true in the swine and poultry industries where nucleus herds are under strict bio-security measures (Ibañez-Escriche and Gonzalez-Recio, 2011). Therefore, methods are needed in order to reflect crossbred performance within purebred selection.

1.6.2 Combining crossbred and purebred selection

One such method involves combining crossbred and purebred selection (CCPS) proposed by Wei and van der Steen (1991) and Lo et al. (1993). This entails collection of phenotypic data on both crossbred and purebred performance and combining the information with a selection index (Wei and van der Werf, 1994). Animals evaluated with this methodology receive breeding values for both purebred and crossbred performance, where they are considered different but correlated traits. Crossbred performance has been shown to increase with CCPS in contrast with purebred line selection in pigs (Bijma and van Arendonk, 1998). However, CCPS leads to an increase in inbreeding because of the increased probability of selecting family members (Bijma et al., 2001; Dekkers, 2007). This strategy requires not only collection of phenotypic data at the commercial level, but also the pedigree information to connect the crossbred animals to their purebred ancestors which has hindered the adoption of CCPS in the industry (Dekkers, 2007). Genomic selection, proposed by Meuwissen et al. (2001), helps to alleviate these downfalls of CCPS.

1.6.3 Genomic selection for crossbred performance

Purebred selection has seen promising results from genomic selection and has been well documented (Meuwissen et al., 2001; Muir, 2007; Hayes et al., 2009). In simulation of a swine production system, Dekkers (2007), demonstrated the use of genomic selection to increase crossbred performance in purebred selection. Crossbred animals were used in the training set in order to estimate marker effects. This led to a higher response in crossbred performance than purebred selection or CCPS. Dekkers (2007) also demonstrated this concept led to lower rates of inbreeding compared to

CCPS. The collection of pedigree data connecting crossbred animals to their purebred ancestors is also no longer needed. Once marker effects have been estimated, they do not need to be re-estimated for several generations (Meuwissen et al., 2001). Taken together, genomic selection was superior to CCPS.

In a simulation by Toosi et al. (2010), training sets included admixed and crossbred populations while the validation set was made entirely of purebred animals of one breed. Results suggested the accuracy of prediction using admixed or crossbred animals for training was similar or slightly less compared to the accuracy when purebred animals of the same breed were used for training and validation. When this methodology was applied to a beef population, average accuracy of MBV from progeny phenotypes of Angus bulls mated to commercial cows were 0.26 and 0.24 when prediction equations were trained on the 2,000 Bull Project (approximately 2,000 influential bulls representing 16 different breeds) and a subset of the 2,000 Bull Project including only Angus individuals, respectively (Weber et al., 2012a).

As the number of breeds included in the crossbred or admixed populations increased, the accuracy decreased, and if the breed used for validation was not included in the admixed or crossbred population, accuracy declined drastically (Toosi et al., 2010). Even if the breed was in the training set, breed composition of the crossbred training set can also have an impact of the accuracy of prediction. Weber et al. (2012b) used the U.S. Meat Animal Research Center Germplasm Evaluation Program (USMARC_GPE) for training while individual breeds from the 2,000 Bull Project were used for validation, both of which are multi-breed populations. Accuracies were higher for breeds that were more represented in USMARC_GPE, namely Angus and Hereford which contributed

almost 28 and 23 percent breed composition respectively, while breeds that were in lower proportions in USMARC_GPE had lower accuracies, particularly Charolais which contributed 6.6 percent.

1.6.4 Modeling breed specific effects

Given that crossbred animals result from parents of different breeds, a model fitting a common additive effect may not be the most suitable approach. Because persistence of LD across breeds may be small, especially for greater divergence between breeds (de Roos et al., 2008), SNP effects may be breed specific. Therefore, alternative models have been proposed to fit breed specific SNP effects (BSAM). Dekkers (2007) proposed a method for a cross of two breeds, but could be extended to more breed crosses, increasing in complexity with the addition of every breed. High-density genotypes were collected on a sample of crossbred animals in commercial herds and their ancestors in the nucleus herds. Marker haplotypes in the crossbred animals were traced back to their purebred parental populations. Haplotype effects were estimated with the high-density genotypes in combination with phenotypes of the crossbred animals, and finally, the estimated haplotype effects were used to help with the selection of purebred animals. With the addition of more breeds, the breed specific haplotypes may be harder to identify and effects to be estimated, therefore accuracy may suffer (Dekkers, 2007).

Ibanez-Escriche et al. (2009) and explored the use of BSAM in simulation in which breed-specific marker effects were explicitly fit in the model and compared to an across-breed SNP genotype model (ASGM) in which a common allele substitution effect was fit. It was hypothesized BSAM would outperform ASGM, however the accuracy of

ASGM was equal to or greater than the accuracy of BSAM for a variety of scenarios including 2-, 3-, and 4-way crosses when only additive gene action was considered (Ibanez-Escriche et al., 2009). As the marker density increased, the need for BSAM decreased because the probability of SNP markers in the model being closer to the QTL increased (Ibanez-Escriche et al., 2009). Additionally, BSAM models included many more parameters in the model that needed to be estimated, and this grew as the number of breeds increased. However, BSAM was advantageous over ASGM when the number of animals in the training set increased; therefore, with more animals, small differences in effect sizes could be estimated especially if parental breeds were distantly related (Ibanez-Escriche et al., 2009). Within simulation, Kinghorn et al. (2010) also modeled the genotypes of the gametes contributing to the crossbred animal, a method they called reciprocal recurrent genomic selection (RRGS), and showed that RRGS led to higher responses in crossbred populations over when common additive allelic substitutions were modeled. Kinghorn et al. (2010) warned the application of this methodology was more suited for swine or poultry rather than sheep or beef industries because the swine and poultry industries have lower generation intervals and crossbreeding systems are already well defined. Kinghorn et al. (2010) also warned the use of RRGS would ultimately push the beef and sheep industries into having more specialized maternal and paternal lines and more structured crossbreeding systems like those found in the swine and poultry industries.

1.6.5 Modeling Dominance

Heterosis is likely due to dominant gene action (Falconer and Mackay, 1996), thus including dominance in a model has the potential to increase the accuracy of crossbred performance in purebred selection. Despite this fact, most published work has only considered models with additive effects (Calus, 2010). This lack of dominance modeling could be due to three reasons. First, many animals with both genotypes and phenotypes are needed in order to estimate dominance effects. This problem has been overcome recently with genotyping becoming more affordable. Secondly, deregressed EBV (DEBV) have been widely used as phenotypes for many traits within genomic evaluations (Garrick et al., 2009). With DEBV, the difference between additive and dominance effects are indistinguishable (Sun et al., 2014). However, with the collection of raw phenotypes, this limitation may also be alleviated. In addition, complex calculations are needed in order to estimate dominance effects (e.g. Misztal et al., 1998). This problem has been combatted by the increased computational power.

Within purebred populations, dominance variance has been shown to account for a small proportion of total variation of yield traits in dairy (Sun et al., 2014), and within phenotypes in swine (Su et al., 2012; Nishio and Satoh, 2014). Prediction accuracy was shown to increase with models that explicitly modeled dominance. However, this was not seen across all traits, especially those that have small dominance variation (Nishio and Satoh, 2014). Because crossbred animals show considerable heterosis, the product of dominance, Nishio and Satoh (2014) predict the inclusion of dominance effects in models would benefit crossbred performance, but this was not proven in their research as they used populations of purebred pigs.

Accuracy in crossbred performance in pigs was shown to increase when dominance was included in the model; however, only genomic information from the purebred parental lines was included (Esfandyari et al., 2016). The collection of information only in parental lines was performed in order to display the usefulness of including dominance in a model when collection of genotypes and phenotypes in the crossbred animals was difficult or infeasible. In contrast, Hidalgo (2015) included genotypes from the crossbred information for training, and was still able to show an increase in accuracy when dominance was included as compared to models with additive effects only.

1.6.6 ssGBLUP for crossbred performance

In multi-step evaluations and its extensions which have been described previously, it was imperative that crossbred animals be genotyped. This was especially true when models depended on crossbred animals for training in order to estimate SNP marker effects. This may not be a realistic approach since economically it may not make sense to genotype a large number of commercial animals.

Based on the work of Wei and van der Werf (1994) with CCPS, Christensen et al. (2014) extended ssGBLUP for the specific use of improving crossbred performance for a two-breed crossbreeding system. This methodology takes advantage of partial relationship matrices that describe relationships based on genetic origin since crossbred animals are derived from more than one breed. This assumes that the SNP markers can be phased - alleles inherited from the sire and dam can be identified. Three traits were modeled, one for the phenotype of the first purebred, one for the phenotype of the second

purebred, and one for the crossbred. This differs from the original ssGBLUP in which a single model is used in which these three traits are assumed to be the same, thus one relationship matrix is used instead of partial relationship matrices. This new ssGBLUP model was implemented in a real two breed swine crossbreeding system for total number of piglet born (Xiang et al., 2016). It was concluded that the new model improved reliabilities of crossbred performance in purebred animals in comparison to a single trait ssGBLUP. The model was later updated by Christensen et al. (2015) to include a three-way crossbreeding system that is commonly used for terminal crossbreeding in swine.

1.7 Current U.S. beef cattle genetic evaluations

Even though the difference between seedstock and commercial herds does not necessarily reduce to purebred and crossbred animals within the beef industry as it does with swine and poultry, it does begin to demonstrate the need for the utilization of commercial phenotypes within genetic evaluations. By the year 2000, more than fifteen different EPD were produced within the national cattle evaluations. At that time, many of those EPD were for traits that addressed the same breeding goal, such as separate EPD for ultrasonically measured carcass traits and actual carcass traits, but often could have led to selection decisions that were in contradiction of each other (Golden et al., 2009). Golden et al. (2000) realized the need to incorporate indicator traits into the analysis of EPD for ERT during genetic evaluations and that the EPD for indicator traits should not be published. This strategy would have eliminated the problem of which EPD to use for selection decisions. Unfortunately, today not all traits published are ERT (e.g. birth weight). Also, the number of published traits has increased, not decreased.

During the estimation of EPD, multivariate models are used to combine information from both the ERT and indicator traits. Because most phenotypes collected are from the seedstock industry, some indicator traits are more convenient, cheaper, or simply more practical to collect than the ERT. For example, ultrasound measurements from seedstock are collected more often than carcass data from progeny tests. The ultrasound measurements generally include intramuscular fat percentage, back fat thickness, and ribeye area which are indicator traits of carcass marbling, back fat, and ribeye area, respectively. The industry has taken a general consensus that ultrasound measurements of carcass traits are reliable indicators of actual carcass data. Literature generally reports moderate to relatively high genetic correlations between the ultrasound and carcass traits (e.g. Moser et al., 1998; Reverter et al., 2000; Devitt and Wilton, 2001; Kemp et al., 2002). This literature justifies the use of ultrasound measurements in seedstock animals to inform selection criteria instead of collecting only actual carcass measurements from progeny test individuals, in which progeny tests are expensive and time consuming to develop. However, Reverter et al. (2000) cautions that genetic correlations are not always consistent across breeds or even between sexes within breeds. The genetic correlation between ultrasound and carcass rib fat thickness was estimated as 0.79, 0.99, 0.87, and 0.02 for Angus bulls, Angus heifers, Hereford bulls, and Hereford heifers (Reverter et al., 2000). Although generally high, genetic correlations between ultrasound and carcass data can differ, thus varying in the validity as adequate indicators.

Additional indicator traits include scrotal circumference as an indicator for age at puberty of a sire's daughter, which is an indicator trait for heifer pregnancy (Golden et al., 2009). Vargas et al. (1998) estimated the genetic correlations between scrotal

circumference and age at puberty to be -0.31, which in this case is favorable; bulls with a larger scrotal circumference tend to have daughters that reach puberty earlier. However, Evans et al. (1999) and McAllister et al. (2011) both found the genetic correlation between scrotal circumference and heifer pregnancy to be near zero. This suggests scrotal circumference is not a reliable indicator of the ERT heifer pregnancy. Therefore, heifer pregnancy phenotypes should be reported for genetic evaluations.

Many traits have a large economic impact within the cattle industry but do not have a breed-wide EPD associated with them. One of these traits is bovine respiratory disease (BRD), which has a large economic impact in the feedlot sector (Snowder et al., 2006). Griffin (1997) estimated BRD accounts for approximately 7% of the total production cost from weaning until the animal is received at the packer. When included in a terminal index, BRD morbidity had an economic value 10.65 times greater than days to finish (Buchanan et al., 2016). Hot carcass weight was the only other trait in the index to have a greater economic value than BRD morbidity; hot carcass weight was 11.47 times more important than days to finish (Buchanan et al., 2016). Other traits included in the index were yield grade, marbling, dry matter intake, and weaning weight. In regards to the lack of collection of disease susceptibility in seedstock or nucleus herds, this is especially true in the swine and poultry industries where nucleus herds are under strict bio-security measures (Ibañez-Escriche and Gonzalez-Recio, 2011). Although beef seedstock herds are not under such strict bio-security measures, true collection of disease phenotypes would mean introducing the pathogen of interest into seedstock herds, which is undesirable for breeding stock (Garrick, 2018).

Given the genetic correlations between indicator traits and the associated ERT are not one, data from the indicator traits do not explain all variation of the ERT. Thus, collection and utilization of ERT phenotypes in genetic evaluations would aid in faster genetic response. Millions of true ERT records (disease incidence, female fertility, growth traits, and carcass traits) are collected within the commercial sectors - cow/calf herds, feedlots, and packing plants - every year. However, most of this data does not make it into the genetic evaluations. This is simply because relationships are needed in order to connect information from family members' performance. There are pedigree ties between seedstock and commercial individuals, but they are often not known for a variety of reasons. Sometimes pedigrees are not recorded, group mating leads to unknown parentage, or pedigree information does not follow the animals as they move along into different segments of the industry. Relationships could be estimated using genomics, but that would require every animal with a record to be genotyped. This is not an economical option even as genotyping costs have decreased. Therefore, most of the phenotypes we are truly interested in are not included in the genetic evaluations.

1.8 Pooling genotypes and phenotypes

1.8.1 Use in GWAS

Genome wide association studies (GWAS) are used in order to discover genetic variations that are associated with traits. These studies typically require a large number of individuals to be genotyped, which can often be in the hundreds or thousands (Huang et al., 2010). Genotyping these large sample sizes can be one of the major limitations of this

research even as the cost of genotyping has decreased over the years. However, pooling DNA for GWAS has been shown to reduce the cost associated with genotyping (Sham et al., 2002). This is done by selectively grouping animals based on a phenotype and then genotyping a combined pool of DNA (Darvasi and Soller, 1994).

Many studies have identified candidate quantitative trait loci through pooling DNA in humans and livestock alike. Huang et al. (2010) used pools of Holstein cattle that exhibited high and low blastocyst rate or fertilization rate. A total of 589 and 571 samples were available for fertilization and blastocyst rate, respectively. Two pools each of high and low rate were constructed for each phenotype, where pool sizes ranged in size from 42 to 49 animals. When testing the association between allelic frequencies and blastocyst rate or fertilization rate, 22 and 5 SNP were found significant, respectively. Results were validated with individual genotypes and found only six of the previously significant SNP were insignificant (P-value > 0.10). Importantly, the signs of the allelic effects were the same between the pooled and individual samples. Many other studies have also shown the use of pooled DNA for GWAS including low reproductive cattle with the presence of SNP mapped to the Y chromosome (McDaneld et al., 2012) and somatic cell score in Valdostana Red Pied cattle (Strillacci et al., 2014). These studies clearly demonstrate the power of pooled DNA testing and their ability to genotype a fraction of samples that would otherwise be needed for individual testing.

1.8.2 Use in prediction

Pooled data for prediction has also been used in a variety of ways. Olson et al. (2006) investigated the use of pooled phenotypes and their effects on prediction accuracy

using simulated data. Work such as this is practical when the phenotype of interest is inherently measured on a group or pen level or when group phenotypes are more cost effective than individual phenotypes. Several other studies have also investigated the use of pooled phenotypes for prediction in simulation and with real data sets. For example, Biscarini et al. (2008) used total body weight and total egg production in laying hens in cages of four, Biscarini et al. (2010) looked at total early egg production in laying hens in cages of four, Cooper et al. (2010) explored total pen intake with steers in pens of six to nine, and Su et al. (2018) used simulation with varying group sizes from three to thirty. One of the major drawbacks of these studies was that all animals within the group or pen must be identified and connected to other animals with a pedigree. Additionally, results showed that pooled observations led to lower accuracies than when individual data were available and utilized (Biscarini et al., 2008; Cooper et al., 2010; Olson et al., 2006; Su et al., 2018). Nonetheless, pooled phenotypes could be effectively utilized in evaluations.

Moving from pooled phenotypes and known pedigrees, research has also been conducted in the use of pooled DNA and a categorical or mean phenotype to predict EBV. Within a simulation mimicking an aquaculture scheme, Sonesson et al. (2010) pooled test individuals two groups based on phenotype, high or low, and marker effects were estimated based on the pooled genotypes of the groups. The marker effects were then used to estimate EBV of individual selection candidates. The accuracy of selection, estimated by the correlation of TBV and EBV, was high when large numbers of test individuals were used in order to estimate the marker effects. Henshall et al. (2012) used the logistic regression of estimated pooled allelic frequencies on mean phenotype to inform marker effects, which were then compared to the effects estimated from

individual genotypes and phenotypes within a simulated data set in beef cattle where the trait was hip height.

As seen previously, when pedigree information is not known, relationships can be derived through the use of genomics. Just as with GWAS, even as genotyping has become cheaper over the years, it still not economical to genotype every commercial animal we would like to include into the genetic evaluations. Recently, the innovative approach of using pooled phenotypic and genotypic data has been used for genetic prediction. Reverter et al. (2016) performed DNA testing on a group of animals based on results of a pregnancy test, and created a "hybrid" genomic relationship matrix (h-GRM) consisting of pooled and non-pooled animals. Genotypes of the pooled animals were given as the B-allele frequencies rather than traditional 0, 1, or 2 for AA, AB, or BB genotypes, respectively. It was concluded that the pooled genomic data can provide estimates of relationships with individual bulls currently in the herd or previously used, and the resulting h-GRM can be used to calculate GEBVs incorporating data from pooled, commercial level herds. Sheep were pooled based on dag scores and sex, and pooled DNA was used in order to estimate an h-GRM (Bell et al., 2017). Contributions of sires to each pool were estimated using simple linear regression and were shown to be equivalent to the GEBV that were estimated using GBLUP (Bell et al., 2017). Alexandre et al. (2019) simulated two traits and pooled animals based on trait one, trait two, a combination of both traits, or randomly and estimated the prediction accuracies of both traits. The highest prediction accuracy of a trait resulted from pooling based on the trait itself and lowest when the pools were constructed randomly. Using Angus data (phenotypes and genotypes), Alexandre et al. (2020) constructed pools in silico with

yearling weight, coat score, and marbling score. Again, prediction was highest when pools were based on the traits themselves, and lowest when the pools were constructed randomly.

A concern with pooled DNA is the addition of pool construction and genotyping errors. Kuehn et al. (2018) investigated the efficiency of estimated genomic relationship of pools to the animals contained in the pools and other potentially related individuals. It was found that the technical error, the error associated with genotyping the intensity of the florescent dye used to estimate the B-allele frequencies, provided a minimal contribution to the total pooled error. Additionally, it was suggested that large pools be utilized because they are less prone to pool construction error – the planned representation of individual DNA to the pool. Thus, if large pools are used, minor errors in pooling allelic frequency can be assumed small. Kuehn et al. (2018) suggested pool sizes of at least 20. On the other hand, Alexandre et al. (2019) suggested pool sizes of 10 in order to retain prediction accuracy and save on the cost of genotyping. Alexandre et al. (2020) confirmed their recommendations of pool sizes of at least 10, and observed consistent accuracies when pool sizes of 15, 20, and 25 were used with in silico beef data.

1.9 Literature Cited

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769. doi:10.1093/jas/skz344.

Alexandre, P. A., A. Reverter, S. A. Lehnert, L. R. Porto-Neto, and S. Dominik. 2020. In silico validation of pooled genotyping strategies for genomic evaluation in Angus cattle. J. Anim. Sci. 98. doi:10.1093/jas/skaa170.

Arranz, J. J., Y. Bayon, and F. San Primitivo. 1996. Comparison of protein markers and microsatellites in differentiation of cattle populations. Anim. Genet. 27: 415-419. doi:j.1365-2052.1996.tb00508.x.

Bell, A. M., J. M. Henshall, L. R. P. Neto, S. Dominik, R. Mcculloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7. doi:10.1186/s12711-017-0303-8.

Bijma, P., and J. A. M. Van Arendonk. 1998. Maximizing genetic gain for the sire line of a crossbreeding scheme utilizing both purebred and crossbred information. Anim. Sci. 66:529–542. doi:10.1017/S135772980000970X.

Bijma, P., J. A. Woolliams, and J. A. M. Van Arendonk. 2001. Genetic gain of pure line selection and combined crossbred purebred. Anim. Sci. 72:225–232. doi:10.1017/S1357729800055715.

Biscarini, F., H. Bovenhuis, and J. A. M. Van Arendonk. 2008. Estimation of variance components and prediction of breeding values using pooled data. J. Anim. Sci. 86:2845–2852. doi:10.2527/jas.2007-0757.

Biscarini, F., H. Bovenhuis, E. D. Ellen, S. Addo, and J. A. M. Van Arendonk. 2010. Estimation of heritability and breeding values for early egg production in laying hens from pooled data. Poult. Sci. 89:1842–1849. doi:10.3382/ps.2010-00730.

Bouw, J., C. Buys, and I. Schreuder. 1974. Further studies on the genetic control of the blood group system C of cattle. Anim. Blood Grps. and Biochem. Genet. 5: 105-114. doi:10.1111/j.1365-2052.1974.tb01318.x.

Bowling, A.T., M. L. Eggleston-Stott, G. Byrns, R. S. Clark, S. Dileanis, E. Wictum. 1997. Validation of microsatellite markers for routine horse parentage testing. Anim. Genet. 28:247-252. doi:10.1111/j.1365-2052.1997.00123.x

Bradford, H. L., I. Pornic, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. J. Anim. Breed. and Genet. 134:545-552. doi:10.1111/jbg.12276.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. Technometrics, 37:373-384. doi: 10.1080/00401706.1995.10484371.

Buchanan, J. W., M. D. Macneil, R. C. Raymond, A. R. McClain, and A. L. Van Eenennaam. 2016. Rapid Communication: Variance component estimates for Charolais-sired fed cattle and relative economic impact of bovine respiratory disease. J. Anim. Sci. 94:5456–5460. doi:10.2527/jas2016-1001.

Calus, M. P. L. 2010. Genomic breeding value prediction: methods and procedures. Animal. 4:157–164. doi:10.1017/S1751731109991352.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su. 2015. Genetic evaluation for three - way crossbreeding. Genet. Sel. Evol. 47:1–13. doi:10.1186/s12711-015-0177-6.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2. doi:10.1186/1297-9686-42-2.

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genet. Sel. Evol. 46:1–9. doi:10.1186/1297-9686-46-23.

Christensen O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal. 6:1565–1571. doi:10.1017/S1751731112000742.

Cooper, A. J., C. L. Ferrell, L. V Cundiff, and L. D. Van Vleck. 2010. Prediction of genetic values for feed intake from individual body weight gain and total feed intake of the pen. J. Anim. Sci. 88:1967–1972. doi:10.2527/jas.2009-2391.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics. 138:1365–1373.

Dekkers, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85:2104–2114. doi:10.2527/jas.2006-683.

Devitt, C. J. B., and J. W. Wilton. 2001. Genetic correlation estimates between ultrasound measurements on yearling bulls and carcass measurements on finished steers. J. Anim. Sci. 79:2790–2797. doi:10.2527/2001.79112790x.

Drinkwater, R. D. and D. J. S. Hetzel. 1991. Applications of molecular biology to understanding genotype-environment interactions in livestock production Proc. of an International Symposium on Nuclear Techniques in Animal Production and Health. April 15–19. Vienna, p. 437-452.

Enns, R. M. 2013. Understanding and applying economically relevant traits (ERT) and indices for the commercial cattle rancher. In: Proc. Range Beef Cow Symposium XXIII, Rapid City, SD. p. 103-107.

Esfandyari, H., P. Bijma, M. Henryon, O. F. Christensen, and A. C. Sørensen. 2016. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet. Sel. Evol. 1–9. doi:10.1186/s12711-016-0220-2.

Evans, J. L., B. L. Golden, R. M. Bourdon, and K. L. Long. 1999. Additive genetic relationships between heifer pregnancy and scrotal circumference in Hereford cattle. J. Anim. Sci. 77:2621–2628. doi:10.2527/1999.77102621x.

Falconer D. S., T. F. C. Mackay TFC. 1996. Introduction to quantitative genetics. Ed. 4. Longman, New York.

Fan, B., Z. Du, D. M. Gorbach, and M. F. Rothschild. 2010. Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australas J. Anim. Sci. 23:833-847. doi:10.5713/ajas.2010.r.03.

Faux, P., N. Gengler, and I. Misztal. 2012. A recursive algorithm for decomposition and creation of the inverse of the genomic relationship matrix. J. Dairy. Sci. 95: 6093-6102. doi:10.3168/jds.2011-5249.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analysis. Genet. Sel. Evol. 46:50-62. doi:10.1186/1297-9686-46-50.

Fragomeni, B. O., D. A. L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J. Dairy Sci. 98:4090-4094. doi:10.3168/jds.2014-9125.

Garrick, D. 2018. Focus on traits not considered. In: Proc. Beef Improv. Federation Ann. Meet. & Symp., Loveland, CO. p. 36-39.

Garrick D., J. Dekkers, and R. Fernando. 2014. The evolution of methodologies for genomic prediction. Livestock Sci. 166:10–18. doi:10.1016/j.livsci.2014.05.031.

Garrick, D. J. and M. Saatchi. 2013. Practical experiences in developing breed-specific predictions for genome-enhanced EPDs. Proc. 10th Genetic Prediction Workshop. Kansas City, MO. p. 24-34.

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.

Georges. M., A. B. Dietz, A. Mishra, D. Nielsen, L. S. Sargeant, A. Sorensen, M. R. Steele, X. Zhao, H. Leipold, J. E. Womack, and M. Lathrop. 1993. Microsatellite mapping of the gene causing weaver disease in cattle will allow the study of an associated quantitative trait locus. Proc. Natl. Acad. Sci. 90:1058-1062. doi:10.1073/pnas.90.3.1058.

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics. 183:347-363. doi.org.10.1534/genetics.109.103952.

Glowatzki-Mullis M.L., C. Gaillard, G. Wigger, R. Fries. 1995. Microsatellite-based parentage control in cattle. Anim. Genet. 26:7–12. doi:10.1111/j.1365-2052.1995.tb02612.x

Golden, B. L., D. J. Garrick, and L. L. Benyshek. 2009. Milestones in beef cattle genetic evaluation. J. Anim. Sci. 87:E3–E10. doi:10.2527/jas.2008-1430.

Golden, B. L., D. J. Garrick, S. Newman, and R. M. Enns. 2000. Economically relevant traits: A framework for the next generation of EPDs. In: Proc. of the Beef Improv. Federation Ann. Meet. & Symp., Witchita, KS. p. 2-13.

Golden, B. L., M. L. Spangler, W. M. Snelling, and D. J. Garrick. 2018. Current single-step national beef cattle evaluation models used by the American Hereford Association and International Genetic Solutions, computational aspects, and implications of marker selection. Proc. 11th Genetic Prediction Workshop. Kansas City, MO. p. 14-22.

Griffin, D. 1997. Economic impact associated with respiratory disease in beef cattle. Vet. Clin. Food Anim. Pract. 13:367–377. doi:10.1016/S0749-0720(15)30302-9.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2001. Partial candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 12:222-231. doi:10.1101/gr.224202.

Grobet, L., L. J. R. Martin, D. Poncelet, D. Pirottin, B. Brouwers, J. Riquet, A. Schoeberlein, S. Dunner, F. Menissier, J. Massabanda, R. Fries, R. Hanset, and M.

Georges. 1997. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. Nat. Genet. 17:71–74. doi:10.1038/ng0997-71.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443. doi:10.3168/jds.2008-1646.

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome. 53:876–883. doi:10.1139/G10-076.

Hayes, B., P. Visscher, and M. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res., 91:47-60. doi:10.1017/S0016672308009981.

Hazel, L. N. 1943. The genetic basis for constructing selection indexes. Genetics. 28:476–490.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31, 423–447. doi:10.2307/2529430.

Henshall, J. M., R. J. Hawken, S. Dominik, and W. Barendse. 2012. Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. Genet. Sel. Evol. 44:1–13. doi:10.1186/1297-9686-44-12.

Heyen, D. W., J. E. Beever, Y. Da, R. E. Evert, C. Green, S. R. E. Bates, J. S. Ziegle, and H. A. Lewin, 1997. Exclusion probabilities of 22 bovine microsatellite markers in fluorescent multiplexes for semi-automated parentage testing. Anim. Genet. 28:21-27. doi:j.1365-2052.1997.t01-1-00057.x.

Hidalgo, A.M. 2015. Exploiting genomic information on purebred and crossbred pigs. PhD thesis. Swedish University of Agricultural Sciences, Uppsala, Sweden and Wageningen University, Wageningen, the Netherlands.

Huang, W., B. W. Kirkpatrick, G. J. M. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Anim. Genet. 41:570–578. doi:10.1111/j.1365-2052.2010.02046.x.

Ibánẽz-Escriche, N., R. L. Fernando, A. Toosi, and J. C. M. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genet. Sel. Evol. 41:1–10. doi:10.1186/1297-9686-41-12.

Ibánẽz-Escriche, N. and O. Gonzalez-Recio. 2011. Review. Promises, pitfalls and challenges of genomic selection in breeding programs. Spanish J. of Ag. Res. 9:404-413. doi:10.5424/sjar/20110902-447-10.

Kachman, S. D. 2008. Incorporation of marker scores into national genetic evaluation. Proc. 9th Genetic Prediction Workshop. Kansas City, MO. pp. 92-98.

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, E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet. Sel. Evol. 45. doi:10.1186/1297-9686-45-30.

Kemp, D. J., W. O. Herring, and C. J. Kaiser. 2002. Genetic and environmental parameters for steer ultrasound and carcass traits. J. Anim. Sci. 80:1489–1496. doi:10.2527/2002.8061489x.

Kinghorn B.P., J. M. Hickey, J. H. J. Van Der Werf. 2010. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proc. 9th WCGALP p. 36.

Kuehn, L. A., McDaneld, T. G., Keele, J. W. 2018 Quantification of genomic relationship from DNA pooled samples. In: Proceedings of the World Congress on Genetics Applied to Livestock Production. February 12-16. Auckland, New Zealand.

Legarra, A., I. Aguilar, I. Misztal. 2009. A relationshiop matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663. doi:10.3168/jds.2009-2061.

Legarra, A. and V Ducrocq. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy. Sci. 95: 4629-4645. doi:10.3168/jds.2011-4982.

Li, B., N. Zhang, Y. Wang, A. W. George, A. Reverter, and Y. Li. 2018. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. Front. Genet. 9. doi:10.3389/fgene.2018.00237.

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations: additive model. Theoret. Appl. Genet. 87, 423–430. doi.org:10.1007/BF00215087.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Lagarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653-2662. doi:10.2527/jas2014-8836.

MacNeil, M. D., J. D. Nkrumah, B. W. Woodward, and S. L. Northcutt. 2010. Genetic evaluation of Angus cattle for carcass marbling using ultrasound and genomic indicators. J. Anim. Sci. 88:517-522. doi:10.2527/jas.2009-2022.

McAllister, C. M., S. E. Speidel, D. H. J. Crews, and R. M. Enns. 2011. Genetic parameters for intramuscular fat percentage , marbling score , scrotal circumference , and heifer pregnancy in Red Angus cattle. J. Anim. Sci. 89:2068–2072. doi:10.2527/jas.2010-3538.

McDaneld, T. G., L . A. Kuehn, M. G. Thomas, W. M. Snelling, T. S. Sonstegard, L. K. Matukumalli, T. P. L. Smith, E. J. Pollak, and J. W. Keele. 2012.Y are you not pregnant: Identification of Y chromosome segments in female cattle with decreased reproductive efficiency. J. Anim. Sci. 90:2142-2151. doi:10.2527/jas.2011-4536.

McPherron, A. C. and S. Le. 1997. Double muscling in cattle due to mutations in the myostatin gene. Proc. Natl. Acad. Sci. 94:12457-12461. doi:10.1073/pnas.94.23.12457.

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.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating improvement of livestock with genomic selection. Annu. Rev. Anim. Biosci 1:221-237. doi: 10.1146/annurev-animal-031412-103705

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. Dairy. Sci. 97: 3943-3952. doi:10.3168/jds.2013-7752.

Misztal, I. and D. Lourenco. 2018. Current research in unweighted and weighted ssGBLUP. Proc. 11th Genetic Prediction Workshop. Kansas City, MO. p. 1-13.

Misztal, I., L. Varona, M. Culbertson, J. K. Bertrand, J. Mabry, T. J. Lawlor, C. P. Van Tassel, and N. Gengler. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol. Agron. Soc. Env. 2:227–233.

Moser, D. W., J. K. Bertrand, I. Misztal, L. A. Kriese, and L. L. Benyshek. 1998. Genetic parameter estimates for carcass and yearling ultrasound measurements in Brangus cattle. J. Anim. Sci. 76:2542–2548. doi:10.2527/1998.76102542x.

Muir, W. M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J. Anim. Breed. Genet. 124:342–355. doi:10.1111/j.1439-0388.2007.00700.x.

Napolitano, F., P. Leone, S. Puppo, B. M. Moioli, F. Pilla, S. Comincini, L. Ferretti, and A. Carretta. 1996. Exploitation of microsatellites as genetic markers of beef-performance traits in Piemontese x Chianina crossbred cattle. J. Anim. Breed. Genet. 113:157-162. doi:10.1111/j.1439-0388.1996.tb00601.x.

Nishio, M., and M. Satoh. 2014. Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS One. 9:1–6. doi:10.1371/journal.pone.0085792.

Nunez-Dominguez, R., L. D. Van Vleck, K. G. Boldman, and L. V Cundiff. 1993. Correlations for genetic expression for growth of calves of Hereford and Angus dams using a multivariate animal model. J. Anim. Sci. 71:2330–2340. doi:10.2527/1993.7192330x.

Newman, S., A. Reverter, and D. J. Johnston. 2002. Purebred-crossbred performance and genetic evaluation of postweaning growth and carcass traits in Bos indicus × Bos taurus crosses in Australia. J. Anim. Sci. 80:1801–1808. doi:10.2527/2002.8071801x.

Ochsner, K. P., M. D. Macneil, R. M. Lewis, and M. L. Spangler. 2017. Economic selection index development for Beefmaster cattle I: Terminal breeding objective. J. Anim. Sci. 95:1063–1070. doi:10.2527/jas2016.1231.

Olson, K. M., D. J. Garrick, and R. M. Enns. 2006. Predicting breeding values and accuracies from group in comparison to individual observations. J. Anim. Sci. 84:88–92. doi:10.2527/2006.84188x.

Peipho, H. P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49:1165–1176. doi:10.2135/cropsci2008.10.0595.

Pollak, E. J., T. P. L. Smith, and W. M. Snelling. 2015. Historical overview and current status of genomic technology and marker assisted selection in beef cattle. In Proceedings of the American Meat Science Association. 68th Annual Reciprocal Meat Conference. June 14 –17. Lincoln, NE. pg. 32-36.

Reverter, A., D. J. Johnston, H. Graser, M. L. Wolcott, and W. H. Upton. 2000. Genetic analyses of live-animal ultrasound and abattoir carcass traits in Australian Angus and Hereford cattle. J. Anim. Sci. 78:1786–1795. doi:10.2527/2000.7871786x.

Reverter, A., L. R. Porto-Neto, M. R. S. Fortes, R. Mcculloch, R. E. Lyons, S. Moore, D. Nicol, J. Henshall, and S. A. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree 1. J. Anim. Sci. 94:4096–4108. doi:10.2527/jas2016-0675.

de Roos, A. P. W., 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. 179:1503–12. doi:10.1534/genetics.107.084301.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: A tool for large-scale association studies. Nat. Rev. Genet. 3:862–871. doi:10.1038/nrg930.

Snowder, G. D., L. D. Van Vleck, L. V Cundiff, and G. L. Bennett. 2006. Bovine respiratory disease in feedlot cattle : Environmental , genetic , and economic factors. J. Anim. Sci. 84:1999–2008. doi:10.2527/jas.2006-046.

Sonesson, A. K., T. H. E. Meuwissen, and M. E. Goddard. 2010. The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. Genet. Sel. Evol. 42:1–9. doi:10.1186/1297-9686-42-41.

Spangler, M. 2013. Strengths and weaknesses of methods of incorporating genomics into genetic evaluations. Proc. 10th Genetic Prediction Workshop. Kansas City, MO. p. 1-4.

Strillacci, M. G., E. Frigo, F. Schiavini, A. B. Samoré, F. Canavesi, M. Vevey, M. C. Cozzi, M. Soller, E. Lipkin, and A. Bagnato. 2014. Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. BMC Genet. 15. doi:10.1186/s12863-014-0106-7.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One. 7:1–7. doi:10.1371/journal.pone.0045293.

Su, G., P. Madsen, B. Nielsen, T. Ostersen, M. Shirali, J. Jensen, and O. F. Christensen. 2018. Estimation of variance components and prediction of breeding values based on group records from varying group sizes. Genet. Sel. Evol. 50:1–12. doi:10.1186/s12711-018-0413-y.

Sun, C., P. M. Vanraden, J. B. Cole, and J. R. O. Connell. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS One. 9. doi:10.1371/journal.pone.0103934.

Thallman, R. M., K. J. Hanford, R. L. Quass, S. D. Kachman, R. J. Templeman, R. L. Fernando, L. A. Kuehn, and 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, California, USA. p. 184-209.

Toosi, A., R. L. Fernando, and J. C. M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88:32–46. doi:10.2527/jas.2009-1975.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Vargas, C. A., M. A. Elzo, C. C. Chase, P. J. Chenoweth, and T. A. Olson. 1998. Estimation of genetic parameters for scrotal circumference, age at puberty in heifers, and hip height in Brahman cattle. J. Anim. Sci. 76:2536–2541. doi:10.2527/1998.76102536x.

Ward, J. 2013. Incorporating genomics into genetic evaluation, Hereford. Proc. 10th Genetic Prediction Workshop. Kansas City, MO. pp. 8-9.

Weber, K. L., D. J. Drake, J. F. Taylor, G. D. J, L. A. Kuehn, R. M. Thallman, R. D. Schnabel, W. M. Snelling, E. J. Pollak, and A. L. Van Eenennaam. 2012a. The accuracies of DNA-based estimates of genetic merit derived from Angus or multibreed beef cattle training populations. J. Anim. Sci. 90:4191–4202. doi:10.2527/jas2011-5020.

Weber, K. L., R. M. Thallman, J. W. Keele, W. M. Snelling, G. L. Bennett, T. P. L. Smith, T. G. McDaneld, M. F. Allan, A. L. Van Eenennaam, and L. A. Kuehn. 2012b. 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/jas2011-4586.

Wei, M. and H. A. M. van der Steen. 1991. Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (a review). Anim. Breed. Abstr. 59:281-298.

Wei, M., and J. H. J. van der Werf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim. Genet. 59:401–413. doi:10.1017/S0003356100007923.

Whittaker, J. C., R. Thompson. and M. C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249-252. doi:10.1017/S0016672399004462.

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. J. Anim. Sci. 94:936–948. doi:10.2527/jas2015-9930.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. Salzeberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurs*. Genome Biol. 10. doi:10.1186/gb-2009-10-4-r42.

Zhang, W. and C. Smith. 1992. Computer simulation of marker-assisted selection utilizing linkage disequilibrium. Theor. And Appl. Genet. 83:813-820. doi:10.1007/BF00226702.

Chapter 2

THE IMPACT OF CLUSTERING METHODS FOR CROSS-VALIDATION, CHOICE OF PHENOTYPES, AND GENOTYPING STRATEGIES ON THE ACCURACY OF GENOMIC PREDICTIONS

2.1 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, principle 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 BayesC 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 rib eye area. 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 five 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 two 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 over-predict true accuracy while selective genotyping tended to under-predict 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.

2.2 Introduction

Many clustering methods have been proposed for cross-validation to assess the accuracy of genomic breeding values (GBV). Legarra et al. (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; Saatchi et al., 2012; Saatchi et al., 2013; Boddhireddy et al., 2014). However, Boddhireddy et al. (2014) showed that principle 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 markers effects have also differed. Adjusted phenotypes led to higher estimated accuracies than non-adjusted phenotypes in sheep (Daetweler et al., 2012). Deregressed Expected Progeny Differences (DEPD) have been used in the past to develop genomic predictors in US beef breeds given genotyped animals were limited and DEPD 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.

2.3 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.

2.3.1 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. BIF 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 rib eye area, and YWT were also obtained from RAAA. These phenotypes were pre-adjusted for sex, age, and breed composition. The final response variable used for analysis was the contemporary group deviation from

these pre-adjusted 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 five were excluded from further analysis. The number 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.

2.3.2 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 three 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 one 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 Geno-Diver, 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 2.1 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 generation 6 to 15 were chosen as animals to have genotypes available. The genotyped animals were chosen using two scenarios: 25% of the animals born in generation 6-15 were chosen at random, or the top 25% of animals born in generations 6-15 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 6-15 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 was 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 one 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.

2.3.3 Cross-validation methods

Seven different clustering methods were employed for cross-validation: k-means, k-medoids, principal component (PC) analysis of the numerator relationship matrix (**A**) and the identical-by-state (IBS) genomic matrix (**G**) assuming the matrices were either a data matrix or a covariance matrix, and random clustering. Each method used five folds/clusters 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, three folds were used for the simulated data given the reduced number of animals in the simulated data compared to the real cattle dataset. For both data sets, training and evaluation sets were arranged using the evaluation set as one 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 $\frac{\mathbf{MM'}}{2\Sigma p_i(1-p_i)}$, 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 one fold and animals with the lowest coefficients being in another.

Random clusters were determined by randomly assigning animals to one of five clusters for the Red Angus cattle or to one of three 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 seven 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 6-11, 6-12, 6-13, or 6-14. The evaluation set consisted of all 2,000 selection candidates in generation 15.

2.3.4 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:

$$y_i = \mu + \Sigma_{j=1}^{k} \boldsymbol{Z}_{ij} u_j \delta_j + e_i$$

where y_i is the DEBV or the adjusted phenotype for animal *i* for each of the four traits, μ is the overall mean, $\boldsymbol{Z}_{ij}$ 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.

2.3.5 Genetic correlation and regression coefficients

Estimates of the genetic correlations ($r_{\hat{g},Y}$) 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 under-predict 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 five 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.

2.4 Results and Discussion

2.4.1 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 five 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 two SNP across a range of distances at generation 15 were consistent with having generated a large amount of short-range LD (results not shown).

2.4.2 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±1.89% (2.56±0.02%) and 9.33±1.56% (2.66±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±0.80% (1.09±0.08%) and 6.20±0.73% (1.29±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 is 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 2.1 for Red Angus animals and Table 2.2 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 2.3 and 2.4 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 zero. Principal component methods across respective relationship matrices led to the highest indices. K-means also had high agreeance 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 2.5 and 2.6, 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 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, 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 birth weight. 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), 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 2.7 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.

2.4.3 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 two 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 11-12 progeny per 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.

Studies within other species have shown that choice of response variable can have an impact on prediction accuracy. Daetweler et al. (2012) found phenotypes adjusted for fixed and breed effects led to higher estimated accuracies than non-adjusted 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 6-11, 6-12, 6-13, and 6-14 to predict animals born in generation 15. The mean estimated and true accuracies of GBV for forward validation are presented in Table 2.8. As animals in generations closer to the selection candidates were included in the training set, the estimated accuracy increased. 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 2.9 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 and phenotypes further suggests that the use of DEBV as a dependent variable generates more reliable estimates of the cumulative SNP effects.

2.4.4 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 genomic breeding values as compared to only genotyping the top individuals. These conclusions were also found by Boligon et al. (2012) who compared five 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, Jimenez-Montero et al. (2012) implemented five genotyping strategies of dams in a forward in time validation. The five 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 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 pre-adjustments 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.

2.5 Literature Cited

Beef Improvement Federation. 2010. Guidelines for Uniform Beef Improvement Programs, 9th ed. Beef Improv. Fed., Raleigh, NC. Available: http://www.beefimprovementorg. Accessed 13 July 2017.

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 1. J. Anim. Sci. 92:485–497. doi:10.2527/jas2013-6757.

Boligon, A. A., N. Long, L. G. Albuquerque, K. A. Weigel, D. Gianola, and G. J. M. 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. Available: 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. 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. 19:136–42. doi:10.1101/gr.083634.108.

Chen, L., F. Schenkel, M. Vinsky, D. H. C. Jr, and C. Li. 2013. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. 4669–4678. doi:10.2527/jas2013-5715.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. 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. 44:1–9. doi:10.1186/1297-9686-44-4.

Clark, S. A., J. M. Hickey, and J. H. J. 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. HJ 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., 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.

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. Available from: http://dx.doi.org/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. 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. doi:10.1534/genetics.107.081190.

Habier, D., J. Tetens, F. 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. Geno-Diver: 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.

Jenko, J., G. R. Wiggans, T. A. Cooper, S. A. E. Eaglen, W. G. de 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, 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 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, and L. Mintegi. 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, and T. H. E. 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)

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.

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. Available from: http://dx.doi.org/10.3168/jds.2011-4338

R Core Team. 2017. R: a language and environment for statis- tical 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, S. Bauck, B. Woodward, J. C. M. 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 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 populations1. 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. L. Smith, T. G. McDaneld, 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/jas2011-4586.

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. M. 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.

Table 2.1. Red Angus average maximum relationships

Clustering Method[1]	Fold	N	$a_{\max\_within}$[2]	$a_{\max\_between}$[3]
K-means	1	2070	0.40	0.22
	2	615	0.41	0.21
	3	572	0.39	0.20
	4	3592	0.31	0.14
	5	2914	0.34	0.19
K-medoids	1	1661	0.36	0.21
	2	1783	0.32	0.17
	3	1839	0.36	0.20
	4	3803	0.35	0.19
	5	377	0.37	0.19
PCN (Data)	1	1952	0.38	0.20
	2	1952	0.36	0.23
	3	1952	0.33	0.23
	4	1952	0.34	0.21
	5	1955	0.32	0.19
PCN (Cov)	1	1952	0.42	0.21
	2	1952	0.36	0.24
	3	1952	0.32	0.24
	4	1952	0.31	0.22
	5	1955	0.29	0.17
PCG (Data)	1	1952	0.40	0.27
	2	1952	0.34	0.26
	3	1952	0.33	0.25
	4	1952	0.31	0.23
	5	1955	0.31	0.18
PCG (Cov)	1	1952	0.32	0.18
	2	1952	0.31	0.22
	3	1952	0.32	0.25
	4	1952	0.35	0.26
	5	1955	0.40	0.26
Random	1	1994	0.31	0.31
	2	1916	0.31	0.31
	3	1893	0.31	0.31
	4	2000	0.31	0.31
	5	1960	0.31	0.31

[1]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle 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

Table 2.2. 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)

[1]Random Selection= 5,000 animals randomly chosen across all 10 generations

[2]Top EBV = The top 500 individuals from each of the 10 generations

[3]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle 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 2.3. 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

[1]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Covariance matrix); Random = random clustering

Table 2.4. 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

[1] K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Covariance matrix); Random = random clustering

Table 2.5. Average accuracy estimates and standard errors across all 5 folds for Red Angus.

		Adjusted Phenotypes[3]			DEBV[4]		
Trait[1]	Clustering Method[2]	N	$r_{\hat{g},Y}$[5]	SE[6]	N	$r_{\hat{g},Y}$	SE
BWT	Kmeans	9,282	0.49	0.06	7,214	0.69	0.05
	Kmedoids		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	Kmeans	6,278	0.46	0.08	6,061	0.54	0.06
	Kmedoids		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	Kmeans	5,582	0.40	0.09	5,275	0.52	0.08
	Kmedoids		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	Kmeans	5,582	0.32	0.07	5,115	0.55	0.08
	Kmedoids		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

[1]BWT = birth weight; YWT = yearling weight; MARB = marbling; REA = ribeye area
[2]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Covariance matrix); Random = random clustering
[3]Adjusted Phenotypes for MARB and REA were the ultrasonically measured intramuscular fat percentage and rib eye area, respectively

[4]DEBV = Deregressed Estimated Breeding Value
[5]$r_{\hat{g},Y}$ = genetic correlation between GBV and either adjusted phenotype or DEBV
[6]S.E. = average standard error across folds

Table 2.6. 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]	$r_{\hat{g},Y}$[4]	S.E.[5]	$r_{\hat{g},TBV}$[6]	S.E.
Random Selection	Phenotype	Kmeans	0.81	0.009	0.77	0.007
		Kmedoids	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
	DEBV	Kmeans	0.84	0.009	0.78	0.006
		Kmedoids	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
Top EBV	Phenotype	Kmeans	0.49	0.015	0.60	0.008
		Kmedoids	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
	DEBV	Kmeans	0.45	0.024	0.66	0.009
		Kmedoids	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

[1]Random Selection = 5,000 animals randomly chosen across all 10 generations; Top EBV = 500 individuals from each of the 10 generations selected

[2]Phenotype = raw phenotype; DEBV = Deregressed Estimated Breeding Value

[3]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Covariance matrix); Random = random clustering

[4]$r_{\hat{g},Y}$ = genetic correlation between GBV and either phenotype or DEBV

[5]S.E. = Standard deviation of correlations across replicates divided by the square root of the number of replicates
[6]$r_{\hat{g},TBV}$ = Residual correlations between GBV and true breeding value including generation, fold, and generation*fold in the model

Table 2.7. 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]	S.E.[5]	$b_{TBV,\hat{g}}$[6]	S.E.
Random Selection	Phenotype	Kmeans	0.91	0.005	0.85	0.008
		Kmedoids	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	Kmeans	0.87	0.005	0.86	0.008
		Kmedoids	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	Kmeans	0.55	0.018	0.73	0.013
		Kmedoids	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	Kmeans	0.42	0.022	0.78	0.007
		Kmedoids	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

[1]Random Selection = 5,000 animals randomly chosen across all 10 generations; Top EBV = 500 individuals from each of the 10 generations selected

[2]Phenotype = raw phenotype; DEBV = Deregressed Estimated Breeding Value

[3]K-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) = Principle component clustering using a numerator relationship matrix(**A** = Data); PCN (Cov) = Principle component clustering using a numerator relationship matrix(**A** = Covariance matrix); PCG (Data) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Data); PCG (Cov) = Principle component clustering using an identical by state genomic relationship matrix (**G** = Covariance matrix); Random = random clustering

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

[5]S.E. = 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

Table 2.8. 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]	$r_{\hat{g},Y}$[4]	S.E.[5]	$r_{\hat{g},TBV}$[6]	S.E.
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

[1]Random Selection = 5,000 animals randomly chosen across all 10 generations; Top EBV = 500 individuals from each of the 10 generations selected

[2]Phenotype = raw phenotype; DEBV = Deregressed Estimated Breeding Value

[3]Training Population (Generations) = Discrete generations used for the training data set. Evaluation set was always generation 15

[4]$r_{\hat{g},Y}$ = genetic correlation between GBV and either phenotype or DEBV

[5]S.E. = Standard error of estimated correlations across replicates divided by the square root of the number of replicates

[6]$r_{\hat{g},TBV}$ = Residual correlations between GBV and true breeding value and including intercept in model

Table 2.9. Average estimated and true regression coefficients and standard errors across all 5 simulations for forward validation.

Selection Strategy[1]	Response Variable[2]	Training populations (Generations)[3]	$b_{Y,\hat{g}}$[4]	S.E.[5]	$b_{TBV,\hat{g}}$[6]	S.E.
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

[1]Random Selection = 5,000 animals randomly chosen across all 10 generations; Top EBV = 500 individuals from each of the 10 generations selected

[2]Phenotype = raw phenotype; DEBV = Deregressed Estimated Breeding Value

[3]Training Population (Generations) = Discrete generations used for the training data set. Evaluation set was always generation 15

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

[5]S.E. = Standard deviation of correlations across replicates divided by the square root of the number of replicates

[6]$b_{TBV,\hat{g}}$ = Residual regression coefficient of true breeding value on GBV and including intercept in model

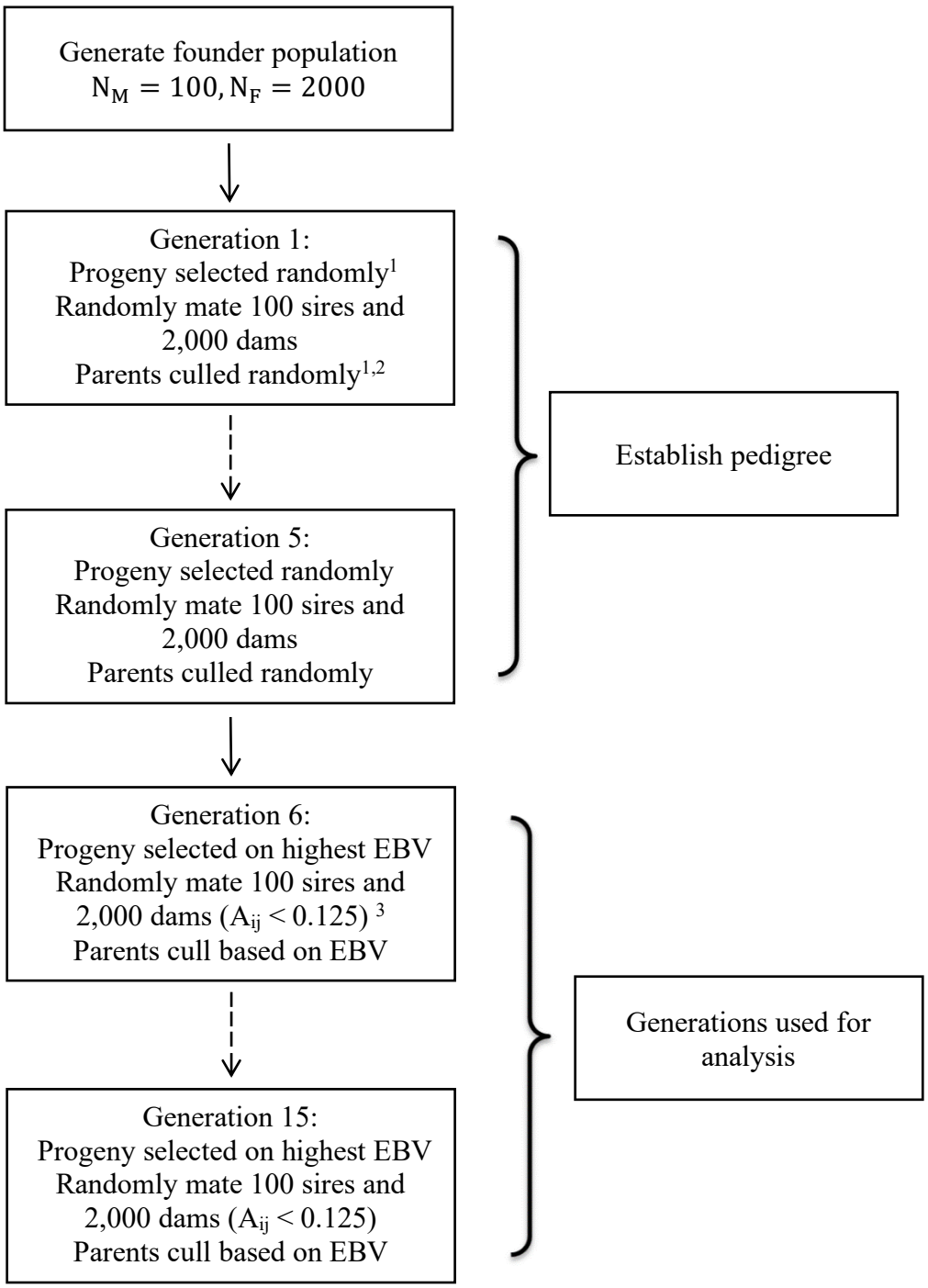


Figure 2.1: Schematic of simulation process. [1]Replacement rates: 0.4 for sires; 0.2 for dams. [2]Animals culled randomly or based on EBV or when they were in the population as a parent for 12 generations. [3]Sires and dams mated randomly with the caveat A_{ij} (additive relationship between animals *i* and *j*) was less than 0.125

Chapter 3

GENOMIC PREDICTION USING POOLED DATA IN A SINGLE-STEP GENOMIC BEST LINEAR UNBIASED PREDICTION FRAMEWORK

3.1 Abstract

Economically relevant traits are routinely collected within the commercial segments of the beef industry but are rarely included in genetic evaluations because of unknown pedigrees. Individual relationships could be resurrected with genomics, but this would be costly; therefore, pooling DNA and phenotypic data provides a cost-effective solution. Pedigree, phenotypic, and genomic data were simulated for a beef cattle population consisting of 15 generations. Genotypes mimicked a 50k marker panel (841 quantitative trait loci were located across the genome, approximately once per 3 Mb) and the phenotype was moderately heritable. Individuals from generation 15 were included in pools (observed genotype and phenotype were mean values of a group). Estimated breeding values (EBV) were generated from a single-step GBLUP model. The effects of pooling strategy (random and minimizing or uniformly maximizing phenotypic variation within pools), pool size (1, 2, 10, 20, 50, 100, or no data from generation 15), and generational gaps of genotyping on EBV accuracy (correlation of EBV with true breeding values) were quantified. Greatest EBV accuracies of sires and dams were observed when there was no gap between genotyped parents and pooled offspring. The EBV accuracies resulting from pools were usually greater than no data from generation 15 regardless of sire or dam genotyping. Minimizing phenotypic variation increased EBV accuracy by 8% and 9% over random pooling and uniformly maximizing phenotypic variation,

respectively. A pool size of 2 was the only scenario that did not significantly decrease EBV accuracy compared to individual data when pools were formed randomly or by uniformly maximizing phenotypic variation ($P>0.05$). Pool sizes of 2, 10, 20, or 50 did not generally lead to statistical differences in EBV accuracy than individual data when pools were constructed to minimize phenotypic variation ($P>0.05$). Largest numerical increases in EBV accuracy resulting from pooling compared to no data from generation 15 were seen with sires with prior low EBV accuracy (those born in generation 14). Pooling of any size led to larger EBV accuracies of the pools than individual data when minimizing phenotypic variation. Resulting EBV for the pools could be used to inform management decisions of those pools. Pooled genotyping to garner commercial-level phenotypes for genetic evaluations seems plausible although differences exist depending on pool size and pool formation strategy.

3.2 Introduction

Millions of phenotypic records are collected annually within commercial sectors of livestock industries including commercial herds, feedlots, and abattoirs. However, most of these records are not included in genetic evaluations because of the lack of available pedigree ties between the commercial and nucleus (seedstock) animals. Examples of traits routinely recorded in commercial settings include carcass merit, disease incidence, female fertility, and growth traits. Many of these trait complexes represent economically relevant traits, those that have a direct source of revenue or cost at the commercial level. Pedigree ties inherently exist between commercial and seedstock animals, but they are often unknown due to lack of recording, group mating, or

the pedigree simply does not follow an animal through the entire production system (Bell et al., 2017). Kinship ties can be resurrected through genomic relationships; but even with the decreasing cost of genotyping, it is still not economically feasible to genotype all commercial animals.

Pooling DNA for genome-wide association studies (GWAS) has been shown to reduce the cost of genotyping (Sham et al., 2002) by selectively grouping animals based on phenotype and then genotyping a combined pool of DNA (Darvasi and Soller, 1994). Many studies have identified candidate quantitative trait loci (QTL) for traits using this approach – e.g., general cognitive ability in children (Fisher et al., 1999), fertility in Holstein cattle (Huang et al, 2010), low reproductive cattle with the presence of single nucleotide polymorphism (SNP) mapped to the Y chromosome (McDaneld et al., 2012), colorectal and prostate cancer in a Polish population (Gaj et al, 2012), and somatic cell score in Valdostana Red Pied cattle (Strillacci et al., 2014). Recently, pooled data has also been used for genetic prediction within a simulated aquaculture population (Sonesson et al., 2010), Brahman and Tropical composite cattle (Henshall et al., 2012; Reverter et al., 2016), Merino sheep (Bell et al., 2017), and a simulated cattle data set (Alexandre et al., 2019). Pooling data, genotypes and thus phenotypes, not only reduces the cost of genotyping, but also allows the inclusion of phenotypes that are typically only observed at the commercial level in genetic evaluations.

The aims of large-scale genetic evaluations should be to improve commercial-level phenotypes that directly impact the profitability of commercial enterprises. However, the majority of, if not all, phenotypes recorded in nucleus (seedstock) settings are indicator traits. A comprehensive genetic evaluation would combine indicator traits

from nucleus animals with the target phenotypes from commercial animals, and to do so would require the use of individual and pooled data simultaneously. However, in some species (e.g., beef cattle) not all parent animals are genotyped thus necessitating the use of both pedigree and genomic kinship as in single-step genomic best linear unbiased prediction (ssGBLUP). Moreover, the estimated breeding values (EBV) of pools could themselves be used to inform management-level decisions. To our knowledge, previous literature has not investigated the accuracy of EBV of the pools themselves. Consequently, the objectives of this paper were to quantify the impact of pool size, method of assigning animals to pools, and generational gaps between the genotyped nucleus (seedstock) and commercial animals on the resulting accuracy of EBV of parents and grand-parents and of the pools in a ssGBLUP framework utilizing simulation.

3.3 Materials and Methods

Animal care and use committee approval was not obtained for this study as all data were simulated.

3.3.1 Simulation

The simulated data used for the analysis was previously described by Baller et al., 2019. Briefly, a purebred beef cattle population was simulated using Geno-Diver (Howard et al., 2017). Five replicates were simulated, each with a different founder genome. Individuals contained 29 chromosomes, with 29 QTL per chromosome. Markers mimicked those from a 50k SNP panel and were randomly distributed across the genome. Locations of the markers and QTL were randomly drawn from separate uniform

distributions. A phenotype with a heritability of 0.4 was simulated. The Markovian Coalescence Simulator (MaCS) program (Chen et al., 2009) generated a founder genome in which a large amount of short-range linkage disequilibrium was created. The founder population was assumed to have an effective population size of 70. Founder animals were randomly selected and mated for five generations in order to establish a pedigree. For an additional ten generations, individuals were randomly mated with the caveat that individuals with an additive relationship of 0.125 or greater were not mated together. These last 10 generations were selectively replaced based on the highest EBV determined by pedigree-based BLUP with replacement rates of 0.2 and 0.4 for dams and sires, respectively. Animals remained within the breeding population until they were culled for low EBV or until they had been a parent for 12 generations.

3.3.3 Pooling

Individuals born in generation 15 (n = 2,000) were assigned to pools, where each individual was included in only one pool per scenario. Pool sizes included 2, 10, 20, 50, or 100 individuals, resulting in 1,000, 200, 100, 40, or 20 pools, respectively. The pool size was consistent within each scenario. Pool assignments were determined in three ways: randomly, minimize phenotypic variation within a pool, and uniformly maximize phenotypic variation within a pool. In order to construct random pools, individuals were randomly assigned a pool number where the only caveat was consistent pool sizes. To minimize phenotypic variation within pools, individuals were ranked based on phenotype and then grouped together dependent on pool size such that for pool size 2, for example, the first two ranked animals were grouped together. This resulted in individuals with the

smallest phenotype in one pool while individuals with the largest phenotype were in another pool. To uniformly maximize phenotypic variation within pools, individuals were again ranked based on phenotype. Individuals with ranks i, i+r, …, i+r(q-1) were assigned to pool i, where r was the number of pools and q was the pool size. For example, when pool size was 2 and thus 1,000 pools were constructed, individuals with ranks one and 1,001 were assigned to pool one and individuals with ranks 1,000 and 2,000 were assigned to pool 1,000. Minimizing and uniformly maximizing phenotypic variation within pools were chosen to demonstrate extreme cases of pooling strategies. Minimizing variation within pools increases variation between pools. Alternatively, maximizing variation within pools decreases variation between pools.

Once the individuals were assigned to pools, the phenotypic record for a given pool was determined as the average of the individuals contributing to the pool. Pooling allele frequency (PAF) for each SNP is based on the normalized intensity of red and green signals from the genotyping assay and is an estimate of the proportion of alternate alleles at every SNP locus (McDaneld et al., 2012). These PAF can be used instead of traditional genotype calls of “0”, “1”, or “2” of individual animals. Pooled genotypes were constructed by averaging the genotype calls across the SNP for all individuals in a pool, resulting in quantitative PAF ranging from 0 to 2. In the current study, all genotypes were assumed to be known without error. Additionally, error associated with the formation of pooled genotypes was also ignored, for example, no over- or under-representation of one individual’s DNA in a pool. Thus, it was assumed that no additional residual variation was introduced through the process of generating pooled genotypes or genotyping. In real populations, PAF can only range from 0 to 1. Within real data, a

minor adjustment can be made to genotype calls and PAF so that they are on the same scale (Bell et al., 2017).

To mimic a commercial setting where pedigree ties are known to exist between the commercial and seedstock individuals but are not often recorded, the animals in generation 15 were not included in the pedigree. Therefore, the only ties between the pools and individuals in the rest of the population were quantified through genomic relationships.

As a means of comparison, pool sizes of 1 in generation 15 were also considered, which is equivalent to individuals having their own phenotypes and genotypes included in the analysis. In this case, PAF was not needed; genotypes entered the evaluation as the typical calls of "0", "1", or "2". Scenarios in which no information was included from generation 15 was also considered to serve as the alternate extreme comparison. This set of scenarios enables the illustration of the EBV accuracy gained with individual or pooled data compared to no data being utilized, which represents the current situation for many livestock industries, particularly those that are non-integrated.

3.3.4 Missing generations of genotypes

All individuals (n=32,000) from the 15 generations had a genotype retained. However, in real livestock populations, genotypes of founder individuals are usually missing and there can be a generational gap between genotyped seedstock animals (e.g., natural service sires in beef cattle, an initial reference population) and commercial animals due to the cost of genotyping. Additionally, genotyped ancestors might be sparse because animals selected for genotyping may be superior or may have an associated

phenotype of particular interest or importance (Boligon et al., 2012). Thus, generational gaps of genotyping were induced. For all scenarios, genotypes were retained once selection began (individuals born in generation 6 or after). Four scenarios were considered: individuals up to and including those born in generation 11 were genotyped (Gen11); up to and including those born in generation 12 were genotyped (Gen12); up to and including those born in generation 13 were genotyped (Gen13); and up to and including those born in generation 14 were genotyped (Gen14). All individuals in generations 6 thru 14, no matter what scenario was considered, had a recorded phenotype as well as known pedigree relationships. Individuals born in generations 12, 13, or 14 that were not genotyped were included in the pedigree and were phenotyped. Individuals born in generations 0 thru 5 that appeared in a three-generation pedigree of the individuals born in generation 15 were included in the pedigree and phenotyped, whereas all others were excluded from the analysis.

3.3.5 Analysis

Single-step GBLUP, which combines genomic and pedigree information in a kinship matrix typically known as **H** (Aguilar et al., 2010; Christensen and Lund, 2010) was used in order to calculate EBV. The model used when only individual data were included in the analysis was $\mathrm{y} = \mathbf{X}\mathrm{b} + \mathbf{Z}\mathrm{u} + \mathrm{e}$ where y was a vector of individual phenotypic observations, **X** was a known incidence matrix relating observations to fixed effects, b was a vector of fixed effects, **Z** was a known incidence matrix relating observations to random additive genetic effects, u was a vector of random additive genetic effects, and e was a vector of random residuals. It was assumed $\mathbf{var}[\mathrm{u}] = \mathbf{G} =$

$\mathbf{H}\sigma_u^2$ and $\mathbf{var}[e] = \mathbf{R} = \mathbf{I}\sigma_e^2$. The only fixed effect considered was the intercept because no other systematic effects were simulated. The inverse of $\mathbf{H}$ ($\mathbf{H^{-1}}$) was constructed as:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

where $\mathbf{A^{-1}}$ was the inverse of the numerator relationship matrix constructed using all animals in the pedigree using the principles derived by Hendersen (1976). Matrix $\mathbf{A_{22}}$ was the pedigree-based relationship matrix of only the genotyped animals and was constructed according to Colleau (2002). The genomic relationship matrix, **G**, was calculated in the following way. First, a genomic relationship matrix ($\mathbf{G_{raw}}$) was computed as $\frac{\mathbf{MM'}}{2\Sigma p_i(1-p_i)}$, where **M** is the centered genotype incidence matrix for individuals and p_i is the allelic frequency of the second allele of the *i*th SNP (VanRaden, 2008). Christensen et al. (2012) formulated a matrix ($\mathbf{G_{scale}}$) in order to make $\mathbf{G_{raw}}$ and $\mathbf{A_{22}}$ compatible by forcing the mean off-diagonal and diagonal elements of $\mathbf{G_{raw}}$ to equal the mean off-diagonal and diagonal elements of $\mathbf{A_{22}}$. This was done by setting $\mathbf{G_{scale}} = \beta\mathbf{G_{raw}} + \alpha$, where β and α are found by solving the following system of linear equations:

$$\overline{\text{diag}(\mathbf{G_{raw}})}\beta + \alpha = \overline{\text{diag}(\mathbf{A_{22}})}$$

$$\overline{\mathbf{G_{raw}}}\beta + \alpha = \overline{\mathbf{A_{22}}}$$

Lastly, matrix $\mathbf{G_{scale}}$ was blended with $\mathbf{A_{22}}$ with coefficients of 0.95 and 0.05, respectively, as suggested by VanRaden (2008) to produce the final genomic relationship matrix (**G**).

When pooled data were added to the analysis and following the notation established by Su et al. (2018), the underlying model was $\mathbf{T}[y = \mathbf{X}b + (\mathbf{ZS})(\mathbf{W}u) + e]$ where vectors y, u, and e and matrices **X** and **Z** were defined the same as above. Let *m*

equal the number of individuals that were not pooled, and again q equal the number of individuals in a pool and r equal the number of pools. Matrix **T** had dimensions (m+r)x(m+rq) and was a design matrix that linked individual observations to the individuals in the pools they were contained in. Matrix **S** had dimensions (m+rq)x(m+r) and was an indicator matrix that linked individual genotypes to pooled genotypes. Matrix **W** had dimensions (m+r)x(m+rq) and was also a design matrix that linked individual breeding values to the breeding values of individuals in the pools they were contained in. Let j denote an animal and k denote a pool. Elements $\mathbf{T}_{kj}$, $\mathbf{W}_{kj}$, and S_{jk} were 1 when j = k for individuals in generations 0 through 14, $\frac{1}{q}$ if the j^{th} animal in generation 15 belonged to the k^{th} pool, and 0 otherwise. The matrices T and W average phenotypes and breeding values within pools. Elements $\mathbf{S}_{jk}$ were 1 if the j^{th} animal in generation 15 belonged to the k^{th} pool and 0 otherwise.

Given the assumptions that individual data (genotypes and phenotypes) were unknown for individuals contained in pools, as could be the case in practice, the final prediction model was $y^* = \mathbf{X}^*b + \mathbf{Z}^*u^* + e^*$ where y^* was a vector of individual observations of animals in generations 0 through 14 and pooled phenotypic observations of animals in generation 15, $\mathbf{X}^*$ was a known incidence matrix relating individual and pooled observations to fixed effects, b was a vector of fixed effects, $\mathbf{Z}^*$ was a known incidence matrix relating individual or pooled observations to random additive genetic effects, u^* was a vector of random additive genetic effects of the individual animals in generations 0 through 14 and pooled animals in generation 15, and e^* was a vector of random residuals. It was assumed $\mathbf{var}[u^*] = \mathbf{G}^* = \mathbf{H}^*\sigma_u^2$, and $\mathbf{var}[e^*] = \mathbf{R}^* = \text{diag}(\frac{1}{q})\sigma_e^2$ because the observations in y^* are heterogeneous in information content given

some phenotypes are individuals and others are means of groups of individuals. The inverse of $\mathbf{H}^*$ was constructed in the same fashion above except that the allelic frequencies, p_i, were estimated from individuals in generations 0 through 14 as well as the pools. The inverse of $\mathbf{H}$ and $\mathbf{H}^*$ was constructed within R (R Core Team, 2017) and then used within ASReml v4.1 software (Gilmour et al., 2009) for the estimation of breeding values.

Accuracy of EBV for sires and dams was estimated as the correlation between true breeding value (TBV) and predicted EBV. The EBV accuracies were estimated for each sex and the generation in which they were born. Accuracy of EBV for pools were estimated as the correlation between the average TBV of the individuals within the pool and the predicted EBV of the pool. To determine the significance of effects on the EBV accuracy, Analysis of Variance tests were performed with the following model:

$$y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + b_l + e_{ijklm}$$

where y_{ijklm} was the EBV accuracy of sires/dams born in generations 11, 12, 13 or 14 or pools; μ was the overall mean; α was the effect of generational gap; β was the effect of pooling strategy; γ was the effect of pool size; b was the random effect of replicate; and e was the random residual. It was assumed b and e were distributed normally with a mean of zero and variance of σ_b^2 and σ_e^2, respectively. Significance was determined at the 0.05 level.

3.3.6 Expectations of pooled genomic relationships

Let $\mathbf{G}^0$ represent a genomic relationship matrix with no pooling. Let $\mathbf{G}^{\mathrm{p}}$ represent the expectation of the genomic relationship matrix when considering pooled and non-

pooled individuals. The expected genomic relationship matrix is a function of $\mathbf{G}^0$ and can be partitioned into four distinct submatrices such that $\mathbf{G}^{\mathrm{p}} = \begin{bmatrix} \mathbf{G}^{\mathrm{p}}_{11} & \mathbf{G}^{\mathrm{p}}_{12} \\ \mathbf{G}^{\mathrm{p}}_{21} & \mathbf{G}^{\mathrm{p}}_{22} \end{bmatrix}$ where $\mathbf{G}^{\mathrm{p}}_{11}$ is the submatrix of relationships between individuals in generations 1 through 14, $\mathbf{G}^{\mathrm{p}}_{12}$ and $\mathbf{G}^{\mathrm{p}}_{21}$ are the submatrices of relationships between individuals in generations 1 through 14 and the pools, and $\mathbf{G}^{\mathrm{p}}_{22}$ is the submatrix of relationships between the pools. Similarly, the genomic relationship matrix can be partitioned into four distinct submatrices such that $\mathbf{G}^0 = \begin{bmatrix} \mathbf{G}^0_{11} & \mathbf{G}^0_{12} \\ \mathbf{G}^0_{21} & \mathbf{G}^0_{22} \end{bmatrix}$. Again, let *q* equal the pool size. The expectations of $\mathbf{G}^{\mathrm{p}}$ are as follows:

1. $\mathbf{G}^{\mathrm{p}}_{11} = \mathbf{G}^0_{11}$.
2. $\{\mathbf{G}^{\mathrm{p}}_{22}\}_{kk'} = \left(\frac{1}{\mathrm{q}}1'\right)\{\mathbf{G}^0_{22}\}_{kk'}\left(\frac{1}{\mathrm{q}}1\right)$ where $\{\mathbf{G}^{\mathrm{p}}_{22}\}_{kk'}$ is the *kk'* element of $\mathbf{G}^{\mathrm{p}}_{22}$ corresponding to pools *k* and *k'* and $\{\mathbf{G}^0_{22}\}_{kk'}$ is the *kk'* submatrix of $\mathbf{G}^0_{22}$ corresponding to individuals in pools *k* and *k'*.
3. $\{\mathbf{G}^{\mathrm{p}}_{12}\}_{jk} = \{\mathbf{G}^0_{12}\}_{jk}\left(\frac{1}{q}1\right)$ where $\{\mathbf{G}^{\mathrm{p}}_{12}\}_{jk}$ is the *jk* element of $\mathbf{G}^{\mathrm{p}}_{12}$ corresponding to individual *j* and pool *k* and $\{\mathbf{G}^0_{12}\}_{jk}$ is the *jk* submatrix of $\mathbf{G}^0_{12}$ corresponding individual *j* and to individuals in pool *k.*

From the expectations above it can be seen that for a pool of unrelated individuals, the diagonal elements of $\mathbf{G}_{22}$ are equal to $\frac{1}{\mathrm{q}}$, the off-diagonals of $\mathbf{G}_{22}$ are proportional to $\frac{1}{\mathrm{q}^2}$, and the elements of $\mathbf{G}_{12}$ and $\mathbf{G}_{21}$ are proportional to $\frac{1}{\mathrm{q}}$. However, as individuals in pools become more related, the diagonal of $\mathbf{G}^{\mathrm{p}}_{22}$ is expected to be greater than $\frac{1}{\mathrm{q}}$.

3.4 Results and Discussion

3.4.1 Pooling

The number of dams contributing to a pool was equal to the pool size because dams have one progeny per generation. However, sires have 20 progeny per generation and so the number of contributing sires to a pool depended on pool size. The average number of contributing sires to a pool across pooling scenarios were 1, 1.99, 9.57, 18.22, 39.76, and 63.96 for pools of 1, 2, 10, 20, 50, and 100, respectively. On average, random assignment led to the most sires contributing to a pool whereas minimizing phenotypic variation led to the smallest. However, these differences were small. The largest discrepancy was seen with a pool size of 100; random assignment led to an average of 0.96 more contributing sires than when minimizing phenotypic variation within pools.

The correlations of the average phenotype and the average TBV within pools are depicted in Figure 3.1. Three distinct patterns emerge when considering pool formation. Randomly assigning individuals to pools led to approximately the same correlation between the average phenotype and average TBV regardless of pool size. When minimizing phenotypic variation within pools, the smallest correlation between average phenotype and average TBV was observed with a pool size of 1 and increased as pool size increased. A large increase was observed between pool sizes of 1 and 2, and again between pool sizes of 2 and 10. After pools of size 10, the gain in the correlation between average phenotype and average TBV plateaued with increasing pool size and approached 1. When considering uniformly maximizing phenotypic variation within pools, the largest correlation was observed with a pool size of 1 and the smallest with a pool size of 100.

Figures 3.2 and 3.3 represent the average relationships of individuals across pools and within pools, respectively. Regardless of the pooling strategy or pool size, the

average relationship of individuals across different pools was approximately equal. Relative to relationships of individuals within pools, random assignment led to approximately equal relationships regardless of pool size with the exception of pool sizes of 2, due to random chance. When minimizing phenotypic variation within pools, relationships were the lowest for pool sizes of 2, the highest for pool sizes of 10, and intermediate for pools of 20, 50, and 100. Grouping individuals together with the same sire based on similar phenotypes was unlikely, especially with groups of two. Grouping some half-sibs together was more likely with pool sizes of 10, which led to the increase in average relationship within pools. The average relationships declined again with pools sizes of 20, 50, and 100 because of the large number of individuals in the pools. When uniformly maximizing phenotypic variation within pools, average relationships within pools were approximately equal with the exception with pools of 2, which led to the lowest relationships. This was because individuals with differing phenotypic values were grouped together and given a moderate heritability it was expected that they would not be highly related.

If individuals in pools were unrelated, expected values of the diagonal of $\mathbf{G}_{22}^{P}$ were $\frac{1}{q_k}$, where q_k was the size of the pool. The average realized values of the diagonal elements of $\mathbf{G}_{22}^{P}$ were 0.99, 0.50, 0.12, 0.07, 0.04, and 0.03 for pool sizes of 1, 2, 10, 20, 50, and 100, respectively. Slight deviations of realized values are due to the fact that some related individuals were pooled together.

3.4.2 EBV accuracies of sires and dams

Figures 3.4 and 3.5 depict the EBV accuracies of sires and dams, respectively, by generation of birth that resulted from different generational gaps in genotyping, pooling strategies, and pool sizes. Results of grand sires/dams are not shown as they follow the same patterns as sires/dams except delayed by one generation. Across all scenarios, the only significant effect was the generational gap in genotyping with the exception of sires and dams born in generation 11. The EBV accuracies of sires born in generation 11 were not significantly impacted by any effects while EBV accuracies of dams born in generation 11 were significantly impacted by both genotyping gaps and pool sizes.

3.4.3 Generational gaps of genotyping

Across all scenarios, the lowest EBV accuracies were observed when genotyping occurred only through generation 11 and the largest were observed when genotyping occurred through generation 14. Increases in EBV accuracy due to larger reference populations have been well documented in literature (e.g. Hayes et al. 2009; Daetwyler et al., 2010). Additionally, in a simulated data set, Lourenco et al. (2017) found that the accuracy of GEBV when using single-step GBLUP increased as more genotyped individuals were used. Note that when genotyping occurred through generation 14, this represented a situation where all information was used. Accuracies of EBV by year of birth for sires and dams were impacted by the generation in which genotyping stopped and EBV accuracies were highest when the genotyping occurred through or past the generation considered. Table 3.1 provides the least-squares means of EBV accuracies when different generational gaps in genotyping were considered. All differences of least-squares means were significant.

The increase in EBV accuracy from when the sires and dams in a generation were genotyped versus when they were not was dependent on sex and the total number of progeny they had contributing to the evaluation. The largest increase in EBV accuracy resulting from additional genotypes was observed with sires and dams born in generation 14. Accuracy of EBV increased by 70% and 54% for sires and dams, respectively, from when genotyping stopped at generation 13 to 14. Accuracy of EBV increased by 9% and 47% for sires and dams born in generation 13, respectively, from when genotyping stopped at generation 12 to 13. Sires born in generation 14 only had progeny that were born in generation 15, which were those that were pooled. Sires born in generation 13 had 20 individually genotyped/pedigreed progeny in addition to the progeny that were pooled in generation 15. The increase in EBV accuracy from when sires were and were not genotyped was not as large for sires born in generation 13 as those born in generation 14 because EBV accuracy of the sires were already relatively high due to the 20 individual progeny born in generation 14 that were at least in the pedigree. The same concept applied to sires born in generations 11 and 12. Dams, on the other hand, had large increases in EBV accuracy from when they were and were not genotyped compared to sires born in the same generation because they had only one progeny per generation. Predictive ability of young animals for growth traits, measured as the correlation between corrected phenotypes and genomic EBV (GEBV), increased from when reference populations included only top bulls with accuracy for birth weight greater than 0.85 (n=1,628) to when all genotyped animals were included (n=33,162) for an Angus population (Lourenco et al., 2015). The gaps in genotyping in the current research could reflect a similar situation in which the top accuracy animals (accuracy accumulated

because of more progeny) were included in the evaluation. From this result, it can be concluded that the quantity and quality of the information used for evaluation matters.

Connectedness between individuals – deduced from pedigrees or genotypes – impacted EBV accuracies, with the latter giving rise to higher EBV accuracies. Additionally, the number of pedigreed progeny also impacted the EBV accuracies. With more pedigreed progeny already in the evaluation, EBV accuracies of sires did not increase as substantially from when individuals themselves were genotyped and when they were not genotyped. The EBV accuracies of sires and dams as a result of pooling were generally higher than if no data from generation 15 entered the evaluation. This was consistent whether the sires or dams in question were genotyped or were not.

3.4.4 Pooling strategy

Although not significant overall, significant differences were found when looking at pairwise differences in least-squares means of different pooling strategies. Differences were not significant between random assignment and uniformly maximizing phenotypic variation but were significant for the other pairwise comparisons. Minimizing phenotypic variation within pools led to larger EBV accuracies than the other two scenarios. The largest differences in least-squares means were found in sires born in generation 14 where minimizing phenotypic variation resulted in an increase of EBV accuracy of 8% and 9% compared to random assignment and uniformly maximizing variation, respectively. Although other comparisons between these pooling scenarios were statistically significant when sires/dams were born in other generations, the difference

may not be practically different. The average increase across generations born and sires/dams was approximately 1% (results not shown).

Henshall et al. (2012) concluded that pooling by the rank of phenotype within contemporary groups led to results more correlated with individual genotyping than pooling based on ranked, pre-adjusted phenotypes across contemporary groups. The current study did not include designed systematic effects, therefore contemporary groups were not considered when constructing pools. Within simulation, Alexandre et al. (2019) pooled individuals based on two traits, one with a heritability of 0.1 (trait 1) and the other of 0.4 (trait 2). The pools were constructed based on trait 1, trait 2, a combination of both, or randomly. Relationships between pools and 200 sires were estimated by genomic relationships alone. Construction of pools based on a single trait was similar to minimizing phenotypic variation within pools in the current study. Accuracies of GEBV, estimated as the correlation of GEBV and TBV, for a single trait were greatest when pools were constructed based on the trait itself and lowest when pools were constructed randomly. Therefore, the ways in which pools are constructed does impact the EBV accuracies of prediction.

3.4.5 Pooling size

Again, while the effect of pool size was not significant overall, some pairwise comparisons of least-squares means did show significant differences. Least-squares means of sire EBV accuracies are presented in Table 3.2. The EBV accuracies of sires resulting from pool sizes of 10, 20, 50, or 100 were not significantly different from those when no information from generation 15 was included in the evaluation when pools were

constructed randomly or by maximizing phenotypic variation. Exceptions to this were for pool sizes of 10 and 20 using either pooling strategy (sires born in generation 14 had significantly increased accuracy) and when pools of size 20 uniformly maximized variation (sires born in generation 13 had significantly increased accuracy). Estimated BV accuracies resulting from pool sizes of 2 were intermediate to situations in which progeny in generation 15 were individually genotyped and when no information from generation 15 was used. Additionally, the only differences in EBV accuracies resulting from pooling and individual data that were not significantly different were with pool sizes of 2. The gain in additional information when pooling randomly or by uniformly maximizing phenotypic variation within pools was not significant when progeny were grouped in pool sizes greater than 10 compared to when data from generation 15 was not used at all, often a numerical gain in accuracy was not even observed. A pooling size of 2 was the only scenario that did not decrease the EBV accuracy significantly when pools were formed randomly or by uniformly maximizing phenotypic variation within pools.

When minimizing phenotypic variation within pools, EBV accuracies of sires resulting from pool sizes of 50 or 100 were not significantly different than those when no information from generation 15 was included. Additionally, EBV accuracies from all pool sizes were not significantly different than individual information from generation 15 with the exception of pool sizes of 10, 20, 50, and 100 when sires were born in generation 12 and genotyping stopped at generation 12 or with pool sizes of 100 when genotyping stopped at either generations 13 or 14. These results also show that EBV accuracies from large pool sizes (50 or 100) show no improvement compared to when data from generation 15 was excluded completely. It also shows that overall, even though there is a

reduction in EBV accuracy resulting from pooling compared to individual data, the reduction is not statistically significant. These results are consistent with Alexandre et al. (2019) who suggested pool sizes of 10 in order to retain EBV accuracy but also save on genotyping costs. However, Kuehn et al. (2018) suggested pool sizes of at least 20. In a study investigating the efficiency of estimated genomic relationships of pools to the animals that make up the pools and to other potentially related individuals, Kuehn et al. (2008) found that technical error (error due to the genotyping of the intensity of the florescent dye) was a minimal contribution to the total pooled error. It was also suggested the use of large pools because they are less prone to pool construction error – the planned representation of individual DNA to the pool. Thus, the impact of errors associated with PAF and pool construction decrease with large pool sizes.

Although some statistically significant differences were found for pairwise comparisons of least-squares means of EBV accuracy of dams, differences in EBV accuracy did not exceed 0.02, and thus results are not presented.

When comparing the decrease in EBV accuracy due to pooling compared to individual data, Alexandre et al. (2019) reported larger decreases compared to those presented herein and were dependent on the heritability of the trait. Alexandre et al. (2019) reported large drops in GEBV accuracy from individual data to pool sizes of 2 and 10, but began to plateau with pool sizes of 20, 25, 50, and 100 for the trait with a heritability of 0.4, when pools were constructed based on the trait itself. The same authors reported that when pools were constructed randomly, GEBV accuracy of the trait with a heritability of 0.4 resulting from pool sizes of 10 was comparable to the GEBV accuracy of the lowly heritable trait. The more dramatic decreases in GEBV accuracy

observed by Alexandre et al. (2019) may be caused by the fact that only a sire's own phenotype and the pools' phenotypes were entered into the evaluation. In the current study other relatives' information also entered into the evaluation, so the decrease in information as pool sizes became larger were not as detrimental, justifying the use of single-step evaluation.

Presumably, results from when no information from generation 15 was included in the evaluation would serve as a lower boundary for EBV accuracy and the upper boundary would be defined by the case when progeny born in generation 15 were genotyped individually. However, when sires/dams were not genotyped and pools were constructed to minimize phenotypic variation within pools, EBV accuracies resulting from pooling were actually higher than if generation 15 had individual data. The EBV accuracies were maximized at pool sizes of 10. This phenomenon was likely a result of both the increased relationship within pools and the confidence in the average phenotype representing the pooled phenotype, determined by the correlation of average phenotype and average TBV in pools. These differences in EBV accuracy from individual data from generation 15 to any pool size were not significant except pools of 10, 20, 50 and 100 for sires born in generation 12 and genotyping stopped at generation 12, as already noted previously.

3.4.6 EBV accuracy of pools

The EBV accuracy of pools are given in Figure 3.6. An Analysis of Variance showed the effects of pool size and the interaction between pool size and pooling strategy to be significant. Pools sizes of 100 had the lowest EBV accuracy and pool size of 1 had

the largest EBV accuracy using random assignment and uniformly maximizing phenotypic variation within pools. However, the effect of pooling when uniformly maximizing variation had larger effects on the EBV accuracy compared to random assignment to pools, seen by larger decreases in EBV accuracy as pool sizes increased. When pools were formed by minimizing phenotypic variation, pool sizes of 100 led to the largest EBV accuracies for the pools while individual data led to the lowest EBV accuracy. Accuracies of EBV resulting from pool sizes of 10 were significantly different compared to pool sizes of 2 and 1. However, EBV accuracies resulting from pools of 10 compared to pool sizes of 20, 50, or 100 were not significantly different.

Practical applications of pooling phenotypes and genotypes have been used before. Bell et al. (2017) used dag scores in Merino sheep to pool individuals in commercial flocks, resulting in categorical phenotypes and PAF for each of the pools. These PAF were combined with individual sire genotypes into a hybrid genomic relationship matrix **(**h-GRM**)** for the use in GBLUP estimations of GEBV of the sires. Pregnancy and lactation status, a categorical phenotype, in Brahman cows were used to pool cattle (Reverter et al., 2016). The resulting PAF from the pools were combined with individual genotypes of herd and stud bulls into an h-GRM for use in GBLUP estimations of GEBV for the fertility of bulls. The bulls were not the sires of the cows in the pools. Within both studies, pedigrees were unknown for the animals used for pooling. These studies showed the potential use of pooling to estimate GEBV of direct parents (Bell et al., 2017) or of seedstock individuals (Reverter et al. 2016). The work of Bell et al. (2017) and Reverter et al. (2016) represent the practical applications of the current study. However, because individual genotypes were not available, the loss of GEBV accuracy

was unknown, warranting further research in this area. Additionally, both Bell et al. (2017) and Reverter et al. (2016) pooled individuals based on similar categorical phenotypes, which would be similar to minimizing phenotypic variation within pools using a quantitative phenotype. The current research demonstrates the validity of work such as Bell et al. (2017) and Reverter et al. (2016), especially when pools are constructed in order to minimize phenotypic variation within the pools and pool size is less than 50. Results from such studies should lead to EBV accuracy that is not significantly different than when individual data is included. Further research with single-step GBLUP and pooling DNA and phenotypic data are needed within real populations.

3.5 Conclusions

Accuracies of EBV from this simulation represent theoretical maximum EBV accuracies; realized EBV accuracies resulting from pooling could be less due to lab and genotyping errors. However, the results presented in this paper show the potential use of pooling data in order to economically make use of commercial data in genetic evaluations. The use of pooled phenotypes and genotypes in combination with a single-step GBLUP evaluation can be a potential way to economically leverage the plethora of phenotypes from commercial sectors in combination with the individual level data (genotypes and phenotypes) from nucleus (seedstock) animals. When pools were constructed in such a way that minimized the phenotypic variation within pools, pool sizes of 2, 10, 20, or 50 did not generally lead to differences in EBV accuracy that are statistically different than when individual progeny data were used. Sires with prior low EBV accuracy benefited the most from pooled observations. Additionally, the resulting

EBV for the pools could be used to inform management decisions. Such examples would be using the EBV for marketing purposes or specialized feeding programs.

3.6 Literature Cited

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769. doi:10.1093/jas/skz344.

Baller, J. L., J. T. Howard, S. D. Kachman, and M. L. Spangler. 2019. The impact of clustering methods for cross-validation , choice of phenotypes , and genotyping strategies on the accuracy of genomic predictions. J. Anim. Sci. 97:1534–1549. doi:10.1093/jas/skz055.

Bell, A. M., J. M. Henshall, L. R. P. Neto, S. Dominik, R. Mcculloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7. doi:10.1186/s12711-017-0303-8.

Boligon, A. A., N. Long, L. G. Albuquerque, K. A. Weigel, D. Gianola, and G. J. M. 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.

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.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2. doi:10.1186/1297-9686-42-2.

Christensen O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal. 6:1565–1571. doi:10.1017/S1751731112000742.

Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genet. Sel. Evol. 34:409–421. doi:10.1051/gse:2002015.

Daetwyler, H. D., R. Pong-wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 185:1021–1031. doi:10.1534/genetics.110.116855.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics. 138:1365–1373. doi:10.1007/bf00222881

Fisher, P. J., D. Turic, N. M. Williams, P. Mcguffin, P. Asherson, D. Ball, I. Craig, T. Eley, L. Hill, K. Chorney, M. J. Chorney, C. P. Benbow, D. Lubinski, R. Plomin, and M. J. Owen. 1999. DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. Hum. Mol. Genet. 8:915–922. doi:10.1093/hmg/8.5.915.

Gaj, P., N. Maryan, E. E. Hennig, J. K. Ledwon, A. Paziewska, A. Majewska, J. Karczmarski, M. Nesteruk, J. Wolski, A. A. Antoniewicz, K. Przytulski, A. Rutkowski, A. Teumer, G. Homuth, T. Starzynska, J. Regula, and J. Ostrowski. 2012. Pooled sample-based GWAS : A cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population. PLoS One. 7. doi:10.1371/journal.pone.0035307.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2015. ASReml User Guide Release 4.1 Functional Specification. VSN International. Hemel Hempstead, United Kingdom. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-4.1-Functional-Specification.pdf

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review : Genomic selection in dairy cattle : Progress and challenges. J. Dairy Sci. 92:433–443. doi:10.3168/jds.2008-1646.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 32:69–83. doi:10.2307/2529339.

Henshall, J. M., R. J. Hawken, S. Dominik, and W. Barendse. 2012. Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. Genet. Sel. Evol. 44:1–13. doi:10.1186/1297-9686-44-12.

Howard, J. T., F. Tiezzi, J. E. Pryce, and C. Maltecca. 2017. Geno-Diver: 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.

Huang, W., B. W. Kirkpatrick, G. J. M. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Anim. Genet. 41:570–578. doi:10.1111/j.1365-2052.2010.02046.x.

Kuehn, L. A., T. G. McDaneld, J. W. Keele. 2018 Quantification of genomic relationship from DNA pooled samples. In: Proceedings of the World Congress on Genetics Applied to Livestock Production; February 12 to 16; Auckland, New Zealand. http://www.wcgalp.org/proceedings/2018/quantification-genomic-relationship-dna-pooled-samples. Accessed 11 June 2020.

Lourenco, D. A. L., B. O. Fragomeni, H. L. Bradford, I. R. Menezes, J. B. S. Ferraz, I. Aguilar, S. Tsuruta, and I. Misztal. 2017. Implications of SNP weighting on single-

step genomic predictions for different reference population sizes. J. Anim. Breed. Genet. 134:463–471. doi:10.1111/jbg.12288.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653–2662. doi:10.2527/jas2014-8836.

McDaneld, T. G., L . A. Kuehn, M. G. Thomas, W. M. Snelling, T. S. Sonstegard, L. K. Matukumalli, T. P. L. Smith, E. J. Pollak, and J. W. Keele. 2012.Y are you not pregnant: Identification of Y chromosome segments in female cattle with decreased reproductive efficiency. J. Anim. Sci. 90:2142-2151. doi:10.2527/jas.2011-4536.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available https://www.R-project.org/.

Reverter, A., L. R. Porto-Neto, M. R. S. Fortes, R. Mcculloch, R. E. Lyons, S. Moore, D. Nicol, J. Henshall, and S. A. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree 1. J. Anim. Sci. 94:4096–4108. doi:10.2527/jas2016-0675.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: A tool for large-scale association studies. Nat. Rev. Genet. 3:862–871. doi:10.1038/nrg930.

Sonesson, A. K., T. H. E. Meuwissen, and M. E. Goddard. 2010. The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. Genet. Sel. Evol. 42:1–9. doi:10.1186/1297-9686-42-41.

Strillacci, M. G., E. Frigo, F. Schiavini, A. B. Samoré, F. Canavesi, M. Vevey, M. C. Cozzi, M. Soller, E. Lipkin, and A. Bagnato. 2014. Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. BMC Genet. 15. doi:10.1186/s12863-014-0106-7.

Su, G., P. Madsen, B. Nielsen, T. Ostersen, M. Shirali, J. Jensen, and O. F. Christensen. 2018. Estimation of variance components and prediction of breeding values based on group records from varying group sizes. Genet. Sel. Evol. 50:1–12. doi:10.1186/s12711-018-0413-y.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

Table 3.1. Least-squares means estimates of EBV accuracies due to generational gaps of genotyping

	Sires[2]				Dams[3]			
Generation Genotyping Stops[1]	14	13	12	11	14	13	12	11
Gen11	0.38	0.82	0.83	0.76	0.48	0.53	0.60	0.83
Gen12	0.41	0.83	0.87	0.72	0.50	0.54	0.83	0.85
Gen13	0.46	0.90	0.90	0.79	0.53	0.82	0.84	0.85
Gen14	0.79	0.91	0.90	0.83	0.82	0.83	0.84	0.86
Std. Error	0.064	0.013	0.022	0.090	0.020	0.016	0.005	0.008

[1]Gen11 = individuals up to and including those born in generation 11 were genotyped; Gen12 = individuals up to and including those born in generation 12 were genotyped; Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped

[2] Sires born in generations 14, 13, 12, or 11

[3] Dams born in generations 14, 13, 12, or 11

Table 3.2. Least-squares means estimates of EBV accuracies of sires due to pooling strategy, pool size, and generational gaps in genotyping

		Born in Generation[3]															
		14				13				12				11			
Pooling Strategy[1]	Pool Size[2]	Gen 11[4]	Gen 12[5]	Gen 13[6]	Gen 14[7]	Gen 11	Gen 12	Gen 13	Gen 14	Gen 11	Gen 12	Gen 13	Gen 14	Gen 11	Gen 12	Gen 13	Gen 14
Random	1	0.40	0.45	0.52[b]	0.87[b]	0.83	0.83	0.92[b]	0.93[b]	0.84	0.91[b]	0.92[b]	0.92[b]	0.82	0.80	0.85	0.88
	2	0.38	0.42	0.48	0.82	0.82	0.83	0.91	0.91	0.83	0.89[b]	0.91	0.90	0.78	0.72	0.80	0.84
	10	0.37	0.39	0.44	0.77[a]	0.82	0.83	0.90[a]	0.90[a]	0.83	0.86[a]	0.89[a]	0.89	0.72	0.69	0.77	0.80
	20	0.37	0.39	0.43	0.75[a]	0.82	0.82	0.89[a]	0.90[a]	0.83	0.86[a]	0.89[a]	0.89[a]	0.73	0.68	0.76	0.79
	50	0.36	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.90[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.74	0.69	0.76	0.80
	100	0.37	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.74	0.69	0.76	0.80
	0	0.37	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.77	0.70	0.77	0.80
Minimize	1	0.40	0.45	0.52[b]	0.87[b]	0.83	0.83	0.92[b]	0.93[b]	0.84	0.91[b]	0.92[b]	0.92[b]	0.82	0.80	0.85	0.88
	2	0.40	0.46	0.54[b]	0.86[b]	0.83	0.83	0.92[b]	0.93[b]	0.84	0.90[b]	0.92[b]	0.91[b]	0.80	0.78	0.84	0.88
	10	0.41	0.47	0.54[b]	0.85[b]	0.83	0.84	0.92[b]	0.92[b]	0.84	0.88[a]	0.91[b]	0.90	0.77	0.76	0.82	0.87
	20	0.40	0.46	0.53[b]	0.84[b]	0.83	0.84	0.92[b]	0.92[b]	0.84	0.87[a]	0.90	0.90	0.77	0.76	0.82	0.87
	50	0.39	0.44	0.50	0.82	0.83	0.83	0.91	0.91	0.83	0.87[a]	0.90	0.90	0.77	0.74	0.81	0.85
	100	0.38	0.42	0.47	0.80	0.83	0.83	0.91	0.91	0.83	0.87[a]	0.90[a]	0.90[a]	0.76	0.72	0.80	0.84
	0	0.37	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.77	0.70	0.77	0.80
Uniformly	1	0.40	0.45	0.52[b]	0.87[b]	0.83	0.83	0.92[b]	0.93[b]	0.84	0.91[b]	0.92[b]	0.92[b]	0.82	0.80	0.85	0.88
Maximize	2	0.38	0.41	0.46	0.81	0.82	0.82	0.90[a]	0.90	0.83	0.89[b]	0.90	0.90	0.74	0.71	0.77	0.81
	10	0.36	0.38	0.43	0.75[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.86[a]	0.89[a]	0.89[a]	0.73	0.68	0.75	0.79
	20	0.37	0.38	0.42[a]	0.74[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.86[a]	0.89[a]	0.88[a]	0.74	0.69	0.76	0.79
	50	0.36	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.89[a]	0.88[a]	0.74	0.69	0.76	0.79
	100	0.37	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.75	0.69	0.76	0.80
	0	0.37	0.38	0.42[a]	0.73[a]	0.82	0.82	0.89[a]	0.89[a]	0.83	0.85[a]	0.88[a]	0.88[a]	0.77	0.70	0.77	0.80
Standard Error		0.073				0.015				0.023				0.100			

[1]Random = individuals were randomly assigned to pools; Minimize = individuals were pooled so that phenotypic variation within pools was minimized; Uniformly maximize = individuals were pooled so that phenotypic variation within pools was uniformly maximized

[2]1 = individually genotyped and phenotyped; 2 = pool size of 2; 10 = pool size of 10; 20 = pool size of 20; 50 = pool size of 50; 100 = pool size of 100; 0 = data from generation 15 did not enter the evaluation
[3]Sires born in generations 14, 13, 12, or 11
[4]Gen11 = individuals up to and including those born in generation 11 were genotyped
[5]Gen12 = individuals up to and including those born in generation 12 were genotyped
[6]Gen13 = individuals up to and including those born in generation 13 were genotyped
[7]Gen14 = individuals up to and including those born in generation 14 were genotyped
[a]Within a column and pooling strategy, the least-squares means difference with a pool size of one is significant
[b]Within a column and pooling strategy, the least-squares means difference with when no information from generation 15 is included is significant

Figure 3.1. Correlation of average phenotype and average true breeding value (TBV) in pools. Pools resulting from different pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools) and pool sizes

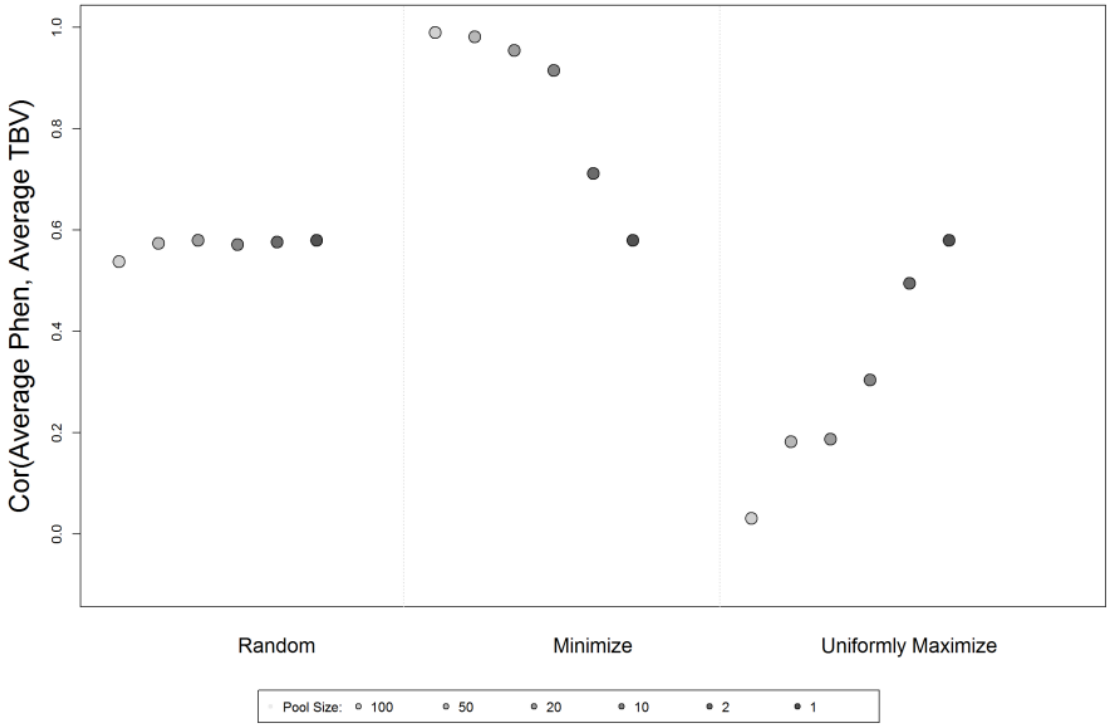


Figure 3.2. Average relationships of individuals across pools. Pools resulting from different pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools) and pool sizes

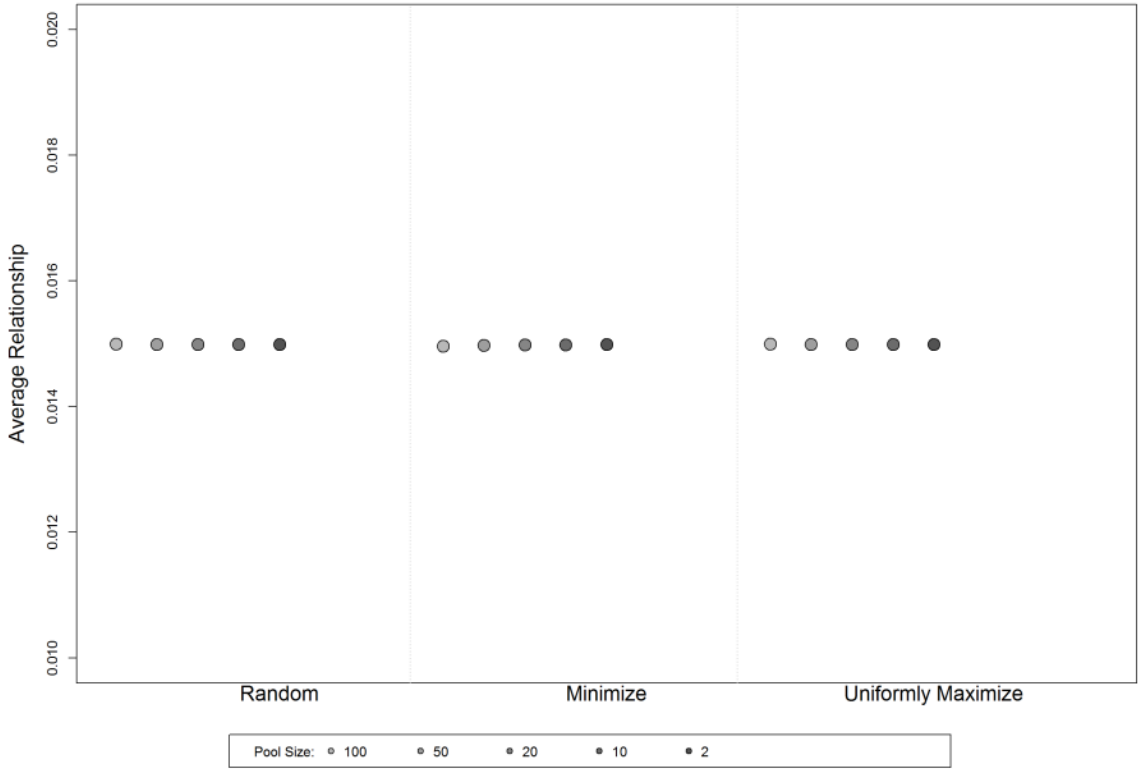


Figure 3.3. Average relationships of individuals within pools. Pools resulting from different pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools) and pool sizes

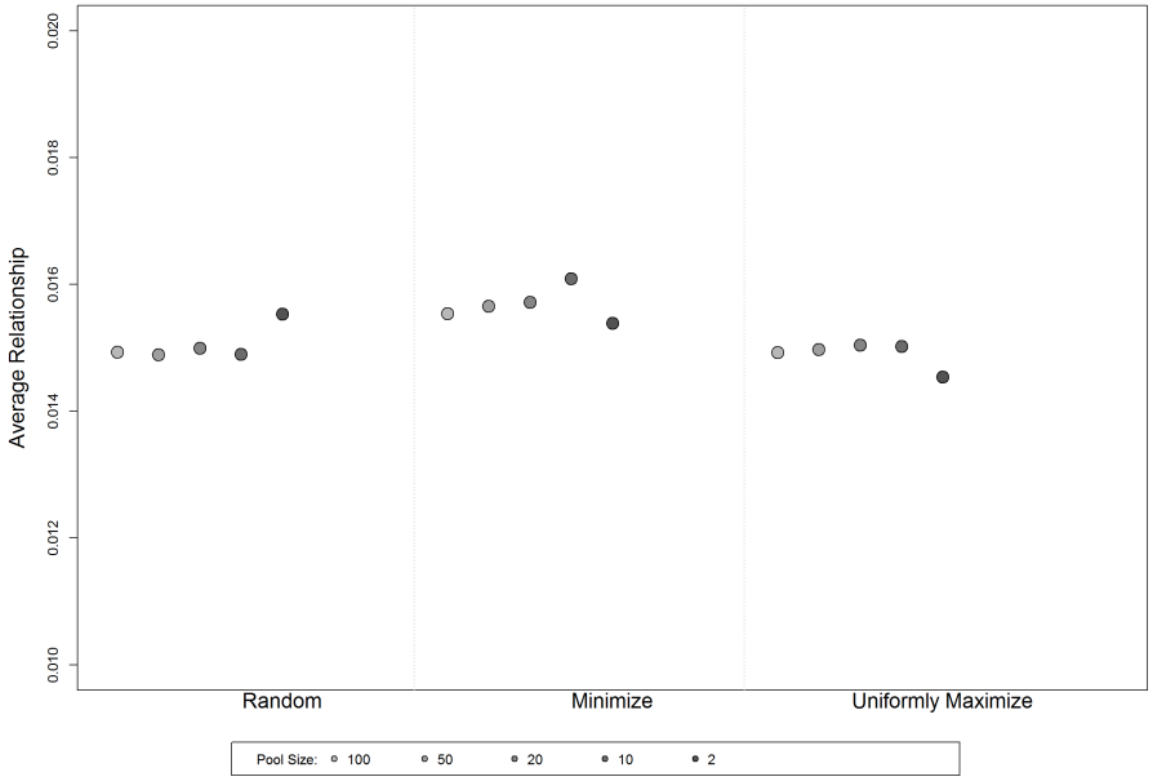


Figure 3.4. Estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value (TBV) and predicted EBV). Presented by generation of birth resulting from different generational gaps in genotyping (Gen11 = individuals up to and including those born in generation 11 were genotyped; Gen12 = individuals up to and including those born in generation 12 were genotyped; Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools), and pool sizes with error bars along x-axis

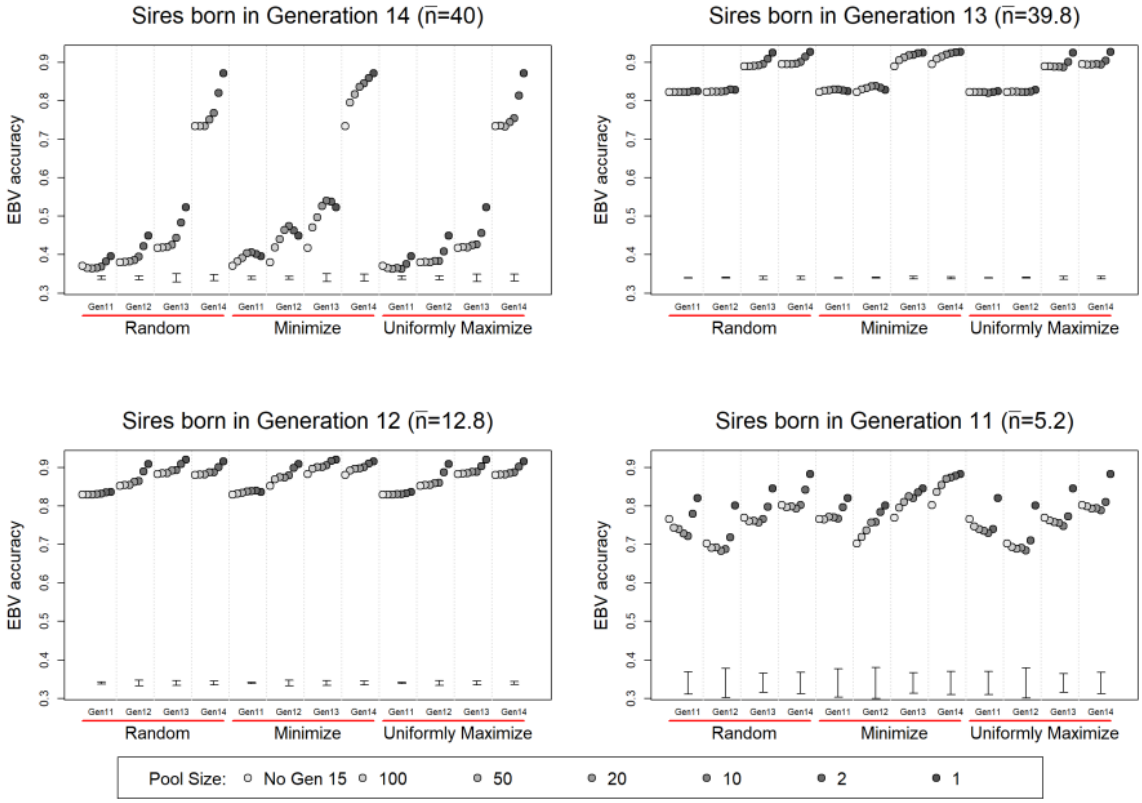


Figure 3.5. Estimated breeding value (EBV) accuracies of dams (estimated as the correlation between true breeding value (TBV) and predicted EBV). Presented by generation of birth resulting from different generational gaps in genotyping (Gen11 = individuals up to and including those born in generation 11 were genotyped; Gen12 = individuals up to and including those born in generation 12 were genotyped; Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools), and pool sizes with error bars along x-axis

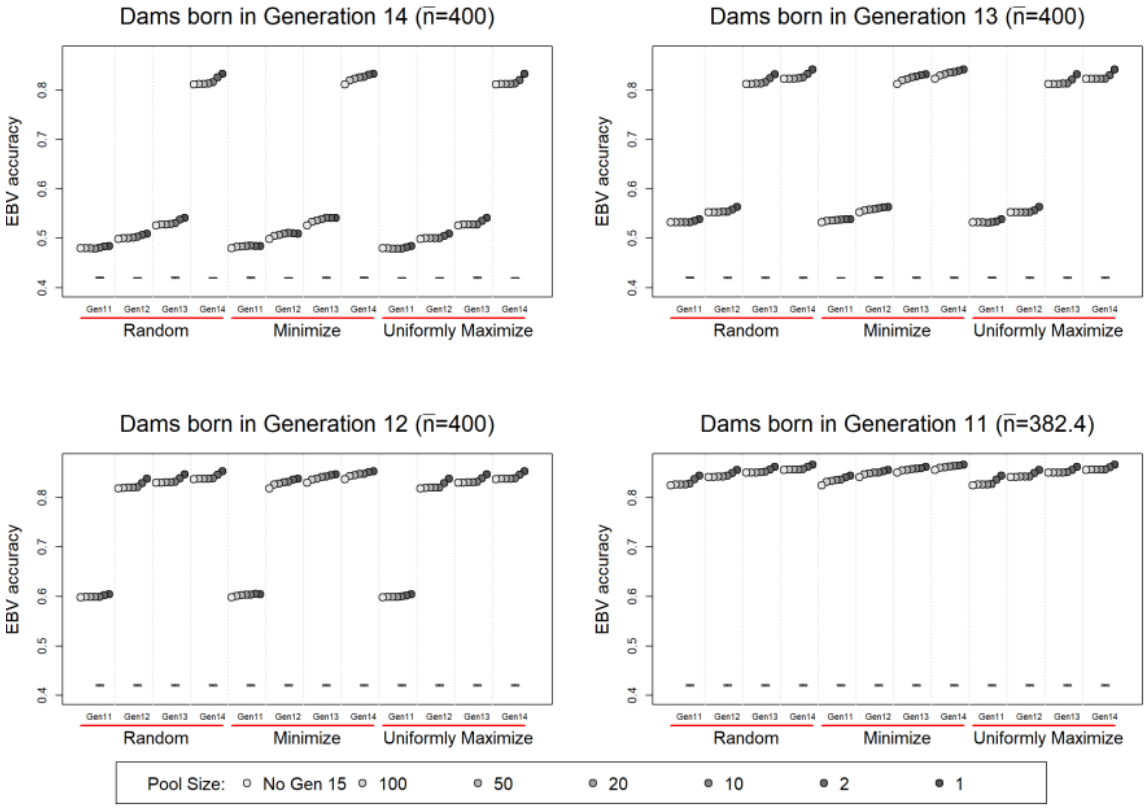
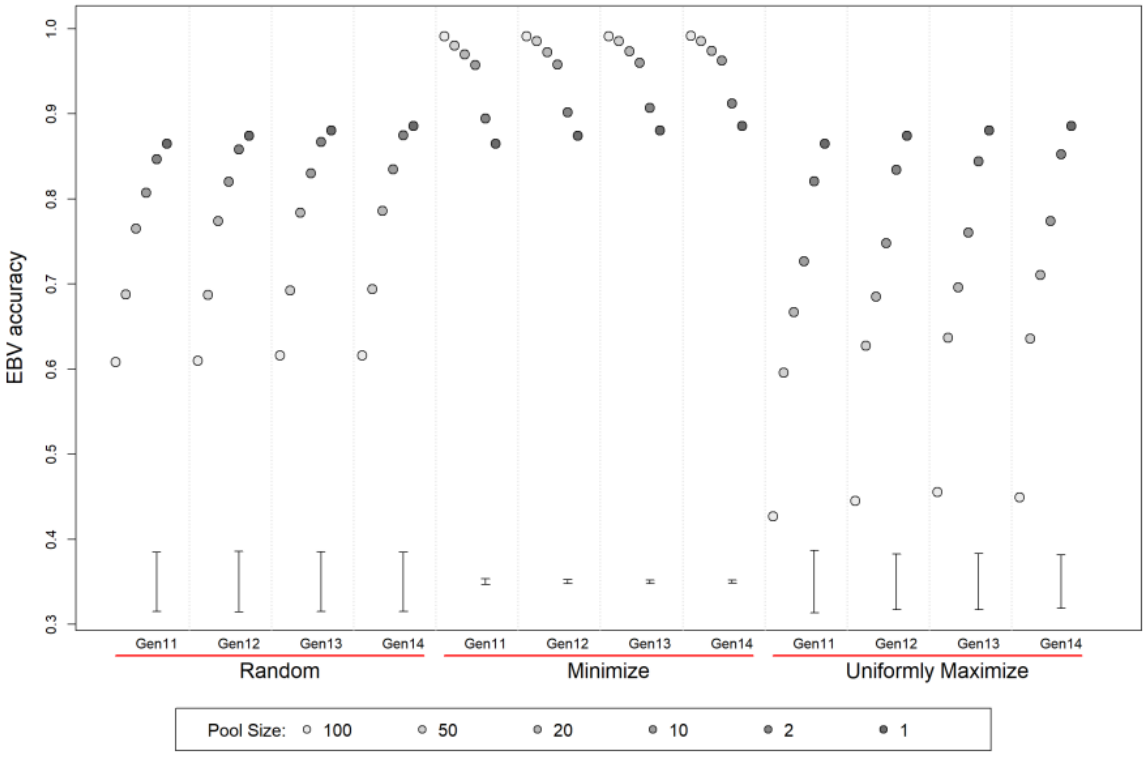


Figure 3.6. Estimated breeding value (EBV) accuracies of pools (estimated as the correlation between the average true breeding value (TBV) of the individuals within the pool and predicted EBV of the pool). Pools resulting from different generational gaps in genotyping (Gen11 = individuals up to and including those born in generation 11 were genotyped; Gen12 = individuals up to and including those born in generation 12 were genotyped; Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools; Uniformly Maximize = uniformly maximize phenotypic variation within pools), and pool sizes with error bars along x-axis



Chapter 4

USING POOLED DATA FOR GENOMIC PREDICTION IN A BIVARIATE FRAMEWORK WITH MISSING DATA

4.1 Abstract

Estimated breeding values (EBV) for economically relevant traits (ERT) are often informed with indicator traits from nucleus animals. Pooling data can enable the use of true ERT from commercial animals within genetic evaluations. Two moderately heritable traits (h^2=0.4) with varying genetic correlations (0.1, 0.4, and 0.7), genotypes mimicking a 50K SNP chip, and pedigree data were simulated for a cattle population consisting of 15 generations (n = 32,000) with random selection. The last generation of individuals, generation 15, was subjected to pooling (n=2,000). Missing records were induced in two ways. With sequential culling, all records for Trait 1 were recorded while only the top 25%, 50%, 75%, or 100% of individuals with Trait 1 records had a Trait 2 record. Random missing records were induced by randomly selecting 80%, 90%, or 100% of individuals to have records for each trait separately. Gaps in genotyping were also explored whereby genotyping occurred through generation 13 or 14. Pools of 1, 20, 50, and 100 animals were constructed randomly or by minimizing phenotypic variation within pools. Results were also compared to scenarios where data from generation 15 did not enter the evaluation. The EBV were estimated using a bivariate single-step genomic best linear unbiased prediction model and EBV accuracies (estimated as the correlation between EBV and true breeding value) were calculated for each trait. The effects of gaps in genotyping, pool size, pooling strategy, genetic correlation, missing value scenario,

and percentage of records available on EBV accuracy were evaluated. Pools of 20 animals constructed by minimizing phenotypic variation generally led to accuracies that were not different than using individual progeny data. Gaps in genotyping led to significantly different EBV accuracies ($p < 0.05$) for sires and dams born in the generation nearest the commercial animals that comprised the pools. As more records were recorded, EBV accuracies of both Trait 1 and Trait 2 increased. Trait 2 EBV accuracies approached Trait 1 EBV accuracies as less animals were culled due to sequential culling. Pooling of any size generally led to larger accuracies than no information from generation 15 regardless of the way missing records arose, the percentage of records available, or the genetic correlation. Given the results from this research, pooling data to aid in the use of commercial ERT in genetic evaluations can be utilized in multivariate cases with varying relationships between the traits and in the presence of systematic and randomly missing phenotypes.

4.2 Introduction

Most of the data included in beef cattle genetic evaluations in the US is recorded within the nucleus (seedstock) segment; however, economically relevant traits (ERT) are, by definition, only observed at the commercial level. Records (phenotypes) are routinely collected within the commercial level but the pedigree relationships needed to connect these records to seedstock animals are often missing due to lack of recording, group mating, or the information does not follow the animals as they move through the industry (Bell et al., 2017). These relationships could be estimated using genomics, but that would require all commercial animals with a phenotype to be individually genotyped. This level

of genotyping would not be economical. Nevertheless, inclusion of commercial data has enormous potential to increase the response to selection for traits that are economically important to the beef industry including feedlot performance, reproductive longevity, disease resistance, and carcass merit. An optimal solution would be to collect the true ERT from commercial herds and estimate relationships between commercial animals and seedstock animals in an economical manner for use in routine genetic evaluations.

Genome-wide association studies (GWAS) in conjunction with pooling has shown to reduce the cost of genotyping (Sham et al., 2002) by grouping together animals with similar observations and then genotyping a pooled DNA sample from those groups (Darvasi and Soller, 1994). Many studies have used pooled DNA for GWAS to identify quantitative trait loci (QTL) in humans (e.g. general cognitive ability in children (Fisher et al., 1999) and colorectal and prostate cancer in a Polish population (Gaj et al., 2012)) and livestock (e.g. low reproductive cattle with the presence of SNP mapped to the Y chromosome (McDaneld et al., 2012), fertility in Holstein cattle (Huang et al., 2010), and somatic cell score in Valdostana Red Pied cattle (Strillacci et al., 2014)).

Pooling has also been investigated for its utility in genetic prediction. Work has been done with simulation - e.g. Sonesson et al. (2010) simulated an aquiculture population whereas Alexandre et al. (2019) and Baller et al. (2020) simulated cattle populations. Pooled data in prediction has also seen use in real data sets – e.g. Henshall et al. (2012) and Reverter et al. (2016) used Brahman Tropical composite cattle, Bell et al. (2017) used Merino sheep, and Alexandre et al. (2020) used in silico Angus data. Most research has focused on the usefulness of pooling on a single trait. Alexandre et al. (2019) extended this concept to two traits, where pools were constructed on one trait or a

combination of two traits using genomic best linear unbiased prediction (GBLUP), and genomic EBV (GEBV) were estimated with univariate models.

To our knowledge, previous studies have not attempted to quantify how pooling separately on the traits affects the EBV accuracy of each trait or combined all information from the two traits in a bivariate model. Additionally, this study was designed to evaluate how the genetic correlation between the two traits can affect EBV accuracy as well as the impact of missing records. The objectives of this study were to evaluate factors that could impact the usefulness of pooling data for genetic prediction in a bivariate context. Consequently, the factors of pooling size, pooling strategy, generational gaps of genotyping, genetic correlation between two traits, how missing values arise, and the percentage of available records were evaluated within a single-step GLBLUP framework to determine how these factors impact EBV accuracy.

4.3 Materials and Methods

Animal care and use committee approval was not required for this research as all data were simulated.

4.3.1 Simulation

Five replicates of a simulation mimicking a purebred beef cattle population were carried out using Geno-Diver (Howard et al., 2017). Following Baller et al. (2019, 2020), each replicate contained a different founder genome comprised of 29 chromosomes each with a length of 87 Mb, which was determined as the average length of chromosomes

using the NCBI *Bos Taurus* 2009 assembly. Markers that represented a 50K SNP panel were randomly distributed across the genome; the location of 1,724 markers per chromosome and the quantitative trait loci (QTL) were drawn randomly from a uniform distribution with the parameters of 0 and the length of the chromosome. It was assumed the QTL occurred once per 3 Mb, resulting in 29 QTL per chromosome. Expanding on the simulations of Baller et al. (2019, 2020), two traits were simulated, each with a heritability of 0.4 resulting from phenotypic, additive, and dominance variances set to 1, 0.4, and 0, respectively. Three different genetic correlations between the phenotypes were simulated for each of the 5 replicates representing low genetic correlation (0.1), moderate genetic correlation (0.4), and high genetic correlation (0.7). The founder genomes were generated by the Markovian Coalescence Simulator (MaCS) program (Chen et al., 2009). Following Baller et al. (2019, 2020) founder genomes were generated to contain a large amount of short-range LD and the effective population size of the founder generation was set to 70. Founder animals consisted of 100 sires and 2,000 dams that were randomly mated for five generations and were randomly replaced, which were used to establish the pedigree. An additional ten generations were simulated where animals were mated randomly with the caveat that animals with a relationship of 0.125 or greater were not mated together. The last 10 generations were randomly selected, with replacement rates of 0.4 and 0.2 for sires and dams, respectively. Animals were also culled when they had been in the population as a parent for 12 generations. Each mating resulted in one progeny; thus, each sire had 20 progeny per generation while each dam only had 1. The final population consisted of a total of 15 generations (n=32,000).

4.3.2 Missing Records

In industry, missing records can manifest in many ways, two of which were simulated in this study – sequential culling and randomly missing records. Selection occurs at various points in an animal's lifetime. Some animals are culled based on a previously recorded trait(s) and do not have the opportunity to express traits later in life. To simulate this, all individuals had an observable Trait 1 phenotype. The animals with the highest 75%, 50%, or 25% Trait 1 phenotype had an observable Trait 2 phenotype recorded..

Missing records can also occurrandomly simply due to missed observations in the field. To simulate this scenario, three different percentages were considered – 100%, 90%, or 80% of records were available (0%, 10%, or 20% or records were missing, respectively). The randomly missing records were determined for each trait independently, but with the same percentage of missing records – leading to 100% of Trait 1 and 100% of Trait 2 available, 90% of Trait 1 and 90% of Trait 2 available, or 80% of Trait 1 and 80% of Trait 2 available. Even though animals were randomly chosen, the same random animals were chosen within each replicate for consistency of comparison; for example, the same 80% of animals were chosen to have records retained within each replicate. Independently, the same 90% of animals were chosen to have records retained within each replicated.

4.3.3 Pooling

The individuals born in generation 15 (n=2000) were assigned to pools. Two sets of pools were constructed: the first set were constructed based on Trait 1 records and the

second set were based on Trait 2 records. Baller et al. (2020) recommended pool sizes of 2, 10, 20, or 50 while Kuehn et al. (2018) recommended pool sizes of 20 as a minimum. Consequently, pool sizes of 20, 50, and 100 were simulated to illustrate a gradient from a recommended minimum to larger values. Pool sizes of 20, 50, and 100 individuals resulted in a total of 100 pools based on Trait 1 and 100 pools based on Trait 2 (total = 200), 40 pools based on Trait 1 and 40 pools based on Trait 2 (total = 80), or 20 pools based on Trait 1 and 20 pools based on Trait 2 (total = 40), respectively. Pool assignments were determined in two different ways: 1) randomly or 2) minimizing the phenotypic variation (of Trait 1 or Trait 2) within a pool. Random pools were determined by randomly assigning individuals to a pool for Trait 1 and then again randomly assigned a pool for Trait 2. For example, for pool size of 20, an animal would first be randomly assigned to one pool from Pool 1 to Pool 100 for inclusion of its Trait 1 record, and then be randomly assigned to one pool from Pool 101 to 200 for inclusion of its Trait 2 record. To construct pools to minimize phenotypic variation within pools, individuals were first ranked based on their phenotypic record for Trait 1 and then grouped together dependent on the pool size. This process was then repeated for Trait 2. For example, with a pool size of 20, the animals with the smallest 20 phenotypes for Trait 1 were included in Pool 1 and the smallest 20 phenotypes for Trait 2 were included in Pool 201. Pools that were constructed based on Trait 1 were assumed to have a missing record for Trait 2 and vice versa unless if the individuals making up the pools for Trait 1 and Trait 2 were identical in which case the pool had a record for both traits. Individuals could only be included in one pool per trait per “scenario”, where “scenario” is defined as a combination of missing record strategy, pooling strategy, percentage of missing records, and generation in which

genotyping stopped, but could be found in two pools if both traits were recorded. Pool size was consistent within each scenario.

The phenotypic record for a given pool was determined as the average phenotype of the individuals contributing to that pool. Genotypes of the pools were average genotype calls across all SNP of the individuals that made up the pool, and ranged from 0 to 2, as described by Baller et al. (2020). It was assumed all genotypes were known without error and there was also no error introduced by pool formation leading to no additional residual error due to the process of pooling DNA samples or genotyping.

Pedigree ties between the commercial and seedstock animals are known to exist but they are often not recorded. Thus, following Baller et al. (2020), the pedigree of the animals in generation 15 were assumed unknown. The only ties between the pooled commercial animals and the seedstock population were estimated by genomic relationships. Missing records for animals in generation 15 followed the same scenarios as with the earlier generations: sequential culling and randomly missing records.

To provide a comparison of extreme cases, scenarios were considered where animals from generation 15 entered the evaluation individually (pool size of 1) and when the animals from generation 15 did not enter the evaluation at all (No gen 15). For pool size of 1, each animal in generation 15 had an opportunity to have an individual record for each trait dependent on whether or not their phenotypes were used for pooling and to have their individual genotype enter into the evaluation. For the case of missing records, some animals were not pooled at all; for consistency of comparing across scenarios, only the individuals that did appear in a pool were considered for pool size of 1. In this case

the genotype calls of these individuals were entered into the evaluation as the traditional "0", "1", or "2".

4.3.4 Missing generation of genotypes

In some cases, parents could potentially not be phenotyped because of the way missing records can arise. Even if the parents did not have a recorded phenotype, they were assumed to be genotyped. As with Baller et al. (2020), generational gaps in genotyping were induced between the seedstock and commercial animals because the cost of genotyping in real populations can be prohibitive. Therefore, the genotypes of animals above the pooled individuals were masked. Two scenarios were considered: 1) animals up to and including those born in generation 13 were genotyped (Gen13) and 2) animals up to and including those born in generation 14 were genotyped (Gen14). Baller et al. (2020) explored additional scenarios where more generations had genotypes masked but they led to similar results as Gen13. All animals in generations 6-14 were included in the pedigree regardless of the genotyping scenario. Additionally, founder animals' may be missing or were not genotyped. Therefore, only animals in generations 0-5 that appeared in a three-generation pedigree of the pooled animals were included in the pedigree and it was assumed these animals were not genotyped. All other animals in generations 0-5 were excluded from the analysis.

4.3.5 Analysis

A bivariate animal model utilizing single-step GBLUP was used to estimate EBV. Single-step GBLUP combines genomic and pedigree information into one kinship matrix

called **H** (Aguilar et al., 2010; Christensen and Lund, 2010). The model used when only individual observations were available (pool sizes of 1 and when generation 15 did not enter the evaluation) was: $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & 0 \\ 0 & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$ where y_i is a vector of individual phenotypic observations for the ith trait; $\mathbf{X}_i$ was a known incidence matrix relating the observations to the fixed effects for the ith trait; b_i was a vector of fixed effects for the ith trait; $\mathbf{Z}_i$ was a known incidence matrix relating observations to the random additive genetic effects for the ith trait; u_i was a vector of random additive genetic effects for the ith trait; and e_i was a vector of random residuals for the ith trait. The only fixed effect included in the model for either trait was the intercept. It was assumed that $\text{var}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{G} \otimes \mathbf{H}$ and $\text{var}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{R} \otimes \mathbf{I}$ where **G** is a 2x2 matrix containing the variance components for the additive effects and **R** is a diagonal matrix containing the variances for the residual effects. The details of the construction of the inverse of the kinship matrix **H** ($\mathbf{H^{-1}}$) was described previously by Baller et al. (2020).

The underlying model introduced by Baller et al. (2020) was extended to a bivariate case. However, it was assumed the individual observations, both genotypes and phenotypes for Traits 1 and 2, of animals in generation 15 were unknown. Thus, the final prediction model used was $\begin{bmatrix} y_1^* \\ y_2^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* & 0 \\ 0 & \mathbf{X}_2^* \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1^* & 0 \\ 0 & \mathbf{Z}_2^* \end{bmatrix} \begin{bmatrix} u_1^* \\ u_2^* \end{bmatrix} + \begin{bmatrix} e_1^* \\ e_2^* \end{bmatrix}$ where y_i^* is a vector of individual and pooled phenotypic observations for the ith trait; $\mathbf{X}_i$ was a known incidence matrix relating the individual and pooled observations to the fixed effects for the ith trait; b_i was the same vector of fixed effects for the ith trait as above (containing only the intercept); $\mathbf{Z}_i$ was a known incidence matrix relating individual and pooled observations to the random additive genetic effects for the ith trait; u_i was a vector of

random additive genetic effects for individuals and pools for the ith trait; and e_i was a vector of random residuals for individuals and pools for the ith trait. It was assumed that $\text{var}\begin{bmatrix} u_1^* \\ u_2^* \end{bmatrix} = \mathbf{G} \otimes \mathbf{H}^*$ and $\text{var}\begin{bmatrix} e_1^* \\ e_2^* \end{bmatrix} = \mathbf{R} \otimes \text{diag}(\frac{1}{q})$ where again **G** and **R** are 2x2 matrices containing the variance components for the additive and residual effects, respectively. The variance of the residuals is the Kronecker product of **R** and a diagonal matrix with elements $\frac{1}{q}$, where q is the pool size, because the phenotypes in y_i are heterogeneous in information content – the phenotypes for animals in generations 0-14 are individual phenotypes whereas the phenotypes for pools are averages of animals from generation 15. The inverse of $\mathbf{H}^*$ was constructed the same as **H** except that the allelic frequencies were estimated from individuals and pools. Pool constructions as well as the computation of inverses of **H** and $\mathbf{H}^*$ were carried out in R (R Core Team, 2017). Breeding values were estimated in the ASReml v4.1 software (Gilmore et al., 2009) using the preconditioned conjugate gradients (PCG) method.

The accuracy of EBV for sires and dams were estimated as the correlation between the true breeding values (TBV) and the EBV. The accuracies were estimated separately for sires and dams, generation in which they were born (11, 12, 13, or 14), and for each trait (Trait 1 and Trait 2). The accuracy of the pools was estimated as the correlation between the average TBV of the animals that made up the pool and the EBV. An observation (EBV accuracy of a sire or dam born within a particular generation, replicate, missing record strategy, pooling strategy, percentage of missing records, and generation in which genotyping stopped – considered a "scenario") was deemed an outlier if it was identified in both an interquartile range (IQR) test within a replicate and an IQR test within a pool size. All data from the "scenarios" deemed as outliers were

removed. A generalized linear model was used to test for the association of factors with being an outlier. A variable identifying an observation as an outlier or not, y, was distributed as a Binomial with parameters $N_{ijklmno}$ and $\pi_{ijklmno}$ and the link function, $\eta_{ijklmno}$ where:

$$\eta_{ijklmno} = \log\left(\frac{\pi_{ijklmno}}{1-\pi_{ijklmno}}\right) = \eta + \tau_i + \beta_j + \gamma_k + \delta_l + \rho(\delta)_{lm} + b_o$$

η was the overall mean; τ was the effect of generational gap; β was the effect of pooling strategy; γ was the effect of pool size; δ was the effect of the way missing values arise; $\rho(\delta)$ was the effect of percentage of available records nested within the way missing values arise; and b was the effect of replicate. A total of 8 outliers were identified and removed. Outliers were only identified within accuracies for Trait 1. The only significant factor associated with the outliers was pooling strategy in which minimizing phenotypic variation had a higher effect than random assignment.

In the presence of outliers, medians are more robust than means; thus, final plotted accuracies are median values across the five replicates. However, to determine the significance of effects on the EBV accuracy, Analysis of Variance tests were performed with the following model:

$$y_{ijklmno} = \mu + \tau_i + \beta_j + \gamma_k + \delta_l + \rho(\delta)_{lm} + \alpha\beta_{ij} + \alpha\gamma_{ij} + \beta\gamma_{jk} + \alpha\delta_{il} + \alpha\rho(\delta)_{ilm} + \beta\delta_{il} + \beta\rho(\delta)_{ilm} + \gamma\delta_{il} + \gamma\rho(\delta)_{ilm} + b_n + e_{ijklmno}$$

where y was the EBV accuracy of sires/dams born in generations 11, 12, 13 or 14 or pools for Trait 1 or Trait 2 with outliers removed; μ was the overall mean; τ was the effect of generational gap; β was the effect of pooling strategy; γ was the effect of pool size; δ was the effect of the way missing values arise; $\rho(\delta)$ was the effect of percentage

of available records nested within the way missing values arise; b was the random effect of replicate; and e was the random residual. The model was restricted to only two-way interactions. It was assumed b and e were distributed normally with a mean of zero and variance of σ_b^2 and σ_e^2, respectively. Significance was determined at $\alpha = 0.05$.

4.3.6 Expectations of pooled genomic relationships

Baller et al. (2020) assumed individuals were only included in one pool, but with the extensions provided in this research, individuals can now be included in more than one pool – a pool based on its Trait 1 phenotype and a separate pool based on its Trait 2 phenotype. Because of this modification, a slight generalization in the expectations of the pooled genomic relationships between the pools presented by Baller et al. (2020) is needed to account for the possibility of shared individuals among pools. Let the matrix $\mathbf{G}_{22}^{0}$ represent the relationships between individuals in generation 15. Similarly, let $\mathbf{G}_{22}^{\mathrm{p}}$ represent the relationships between the pools. The expected genomic relationship matrix $\mathbf{G}_{22}^{\mathrm{p}}$is a function of $\mathbf{G}_{22}^{0}$ and follows:

1. $\{\mathbf{G}_{22}^{\mathrm{p}}\}_{kk'} = \left(\frac{1}{\mathrm{q}}\mathrm{I}_{\mathrm{k}}'\right)\{\mathbf{G}_{22}^{0}\}_{kk'}\left(\frac{1}{\mathrm{q}}\mathrm{I}_{\mathrm{k}'}\right)$ where $\{\mathbf{G}_{22}^{\mathrm{p}}\}_{kk'}$ is the *kk'* element of $\mathbf{G}_{22}^{\mathrm{p}}$ corresponding to pools *k* and *k'*, $\{\mathbf{G}_{22}^{0}\}_{kk'}$ is the *kk'* submatrix of $\mathbf{G}_{22}^{0}$ corresponding to individuals in pools *k* and *k'*, and I_{k}' and $\mathrm{I}_{k'}$ are indicator vectors for pools *k* and *k*' with elements 1 if the individual is in the pool and 0 if the individual is not in the pool.

Assume all individuals in generation 15 are unrelated. From the expectations above it can be seen that for pools of individuals, the diagonal elements of $\mathbf{G}_{22}^{\mathrm{p}}$ are equal to $\frac{1}{\mathrm{q}}$ and

the off-diagonals of $\mathbf{G}_{22}^{\mathrm{p}}$ are proportional to $\frac{\mathrm{m}}{\mathrm{q}^2}$ where m is the number of individuals in common between two pools. Thus, the off-diagonals of $\mathbf{G}_{22}^{\mathrm{p}}$ between pools that were based off of the same trait are expected to be zero as they share no common individuals, but are expected to be proportional to $\frac{1}{q^2}$ if one animal is in common between pools based on different traits, proportional to $\frac{2}{q^2}$ if two animals are in common, and so on. If the individuals in generation 15 are related, as is the case in this simulation and likely with real data, the diagonal elements of $\mathbf{G}_{22}^{\mathrm{P}}$ are expected to be greater than $\frac{1}{\mathrm{q}}$ and the off diagonal elements of $\mathbf{G}_{22}^{\mathrm{P}}$ between pools based on different traits will be greater than $\frac{m}{q}$ as the individuals in the pools become more related.

4.4 Results and Discussion

4.4.1 Pooling

Figure 4.1 depicts the correlation between the average phenotype and average TBV of the pools. Regardless of genetic correlation, the way in which missing values arise, the percentage of available records, or the trait considered, pool sizes of 20, 50, and 100 led to larger correlations of average phenotype and TBV compared to pool sizes of 1; this agrees with Baller et al. (2020). Previously, Baller et al. (2020) observed pools constructed randomly led to approximately similar correlations between average phenotype and TBV regardless of pool size. In the current study, this was not observed. No identifiable pattern in regards to pool sizes were observed with random pooling. However, the range of correlations between average phenotype and TBV were larger for sequential culling than for random missing records.

The average relationships within a pool and across pools were approximately equal regardless of pool size. The comparison across pools was only considered within the trait the pools were designed for. Regardless of how missing values arise, the average relationships within a pool and between pools were approximately the same for Traits 1 and 2 when pools were formed to minimize phenotypic variation. However, when pools were formed randomly, the average relationships of Trait 2 were typically higher than those of Trait 1, both within and across pools. The difference between the average relationships of pools based on Trait 1 and 2 becomes larger as the percentage of available records becomes smaller. The average relationships within pools and across pools within the trait the pools were designed for were lower than those observed by Baller et al. (2020). This could be an artifact of selection – Baller et al. (2020) simulated a population whereby selective replacement based on EBV was practiced whereas the current simulation employed random selection.

When considering the average relationships of individuals between pools based on Traits 1 and 2, it is important to note again that the same individuals were used for pooling across all pool sizes and pooling strategies. Additionally, within the way missing records arise and the percentage of individuals available, the individuals were always the same for consistency. Regardless of genetic correlation, the average relationship of individuals between pools based on Traits 1 and 2 increased as the percentage of records available increased when missing records arose randomly. This was caused by the fact that it was very unlikely the same animals would randomly have missing records for both traits, thus the greater difference in animals as the percentage of missing records increased. The average relationship of individuals between pools based on Traits 1 and 2

also increased as the percentage of records available increased with sequential culling and a genetic correlation of 0.7. This increase in relationship is expected as it is more likely related animals were retained during sequential culling when the genetic correlation is high. With a genetic correlation of 0.4 and sequential culling, the relationships between pools based on different traits were approximately the same regardless of the percentage of records available, except for when 25% of Trait 2 records were available, which led to lower average relationships. With a genetic correlation of 0.1, sequential culling, and across all percentages of available records, the relationships between pools based on different traits were approximately equal.

4.4.2 EBV accuracies of sires and dams

Figures 4.2 and 4.3 depict the median EBV accuracies of sires for sequential culling and randomly missing records, respectively, depending on genetic correlation, generation the sires were born in, pooling strategy, percentage of missing records, and the generation in which genotyping stopped. Results of dams are not shown as they follow the same patterns as the sires. Although the same patterns exist with the sires and dams, two key differences do exist. First, the median EBV accuracies of dams were numerically lower than those of the sires. Additionally, the difference between EBV accuracy when pool sizes of 1 were used and when generation 15 did not enter the evaluation at all was smaller for dams than sires. Both of these were due to the fact that dams only had one progeny per generation while sires had 20.

Pooling data has been implemented in practice. Reverter et al. (2016) used pooling within Brahman cattle for pregnancy and lactation status using GBLUP.

Estimations of GEBV for fertility were obtained for bulls that were not sires of the cattle that were pooled. Bell et al. (2017) used pooling within Merino sheep using dag scores also using GBLUP to attain estimates of GEBV. The accuracies of GEBV resulting from pooled data were not compared to a baseline of GEBV resulting from individual data, and so it is not known if the loss of accuracy in prediction due to pooling is significant or not. Alexandre et al. (2020) used Angus data in silico to compare the GEBV accuracies from pooling compared to GEBV accuracies of individual data. Nevertheless, the work of Reverter et al. (2016), Bell et al. (2017), and Alexandre et al. (2020) demonstrate the potential use of pooled data in genetic evaluations.

4.4.3 Generational gap of genotyping

For sires and dams born in generation 14, the EBV accuracies of both traits were lower when genotyping stopped at generation 13 than when genotyping occurred through generation 14 by 0.140 and 0.136 for sires and dams, respectively. This is because if genotyping stopped at generation 13, animals born in generation 14 were not genotyped. Large decreases in EBV accuracy were not found in sires or dams born in generations 13 or earlier dependent on when genotyping stopped because the animals born in these generations were always genotyped. Baller et al. (2020) also noted that EBV accuracies of sires and dams by year of birth were highest when the genotyping occurred through or past the generation considered. Therefore, larger EBV accuracies are a result of connectedness arising from genomic relationships rather than pedigree relationships (Baller et al., 2020). Using single-step GBLUP in a simulated data set, the accuracy of GEBV increased as more genotyped individuals were used (Lourenco et al., 2017).

4.4.4 Pooling strategy and size

When pools were constructed randomly, the EBV accuracy resulting from any pool size or when generation 15 did not enter the evaluation was significantly lower than that from a pool size of 1. When pools were constructed to minimize phenotypic variation, more interesting comparisons were apparent. Ideally, for pooling to be an acceptable approach to include commercial data into evaluations, EBV accuracies of pools would be significantly different than those from when generation 15 did not enter the evaluation and not different from a pool size of 1. This occurred for pool sizes of 20, 50, and 100 for sires born in generation 14 for Trait 1, pool size of 20 for dams born in generation 14 for Trait 1, and pool size of 20 for sires born in generations 13 and 14 for Trait 2. A less optimal situation would be where the EBV accuracies as a result from pooling were still significantly higher than when generation 15 did not enter the evaluation but also significantly lower than pool sizes of 1. This occurred with pool sizes of 20, 50, and 100 for sires born in generation 13 for Trait 1 and pool sizes of 50 and 100 for sires born in generation 14 for Trait 2. These comparisons may be statistically significant; however, numerically, the largest pairwise difference was 0.03 as they were averaged over generation in which genotyping stopped, genetic correlation, the way in which missing records arose, and the percentage of missing records nested within how the missing records arose (results not shown).

Previously, Baller et al. (2020) constructed pools to uniformly maximize phenotypic variation within pools, but it was determined this strategy resulted in comparable results to random allocation to pools and did not see improvement in EBV

accuracy above those from minimizing phenotypic variation within pools. Baller et al. (2020) concluded that when pools were constructed by minimizing phenotypic variation, pool sizes of 2, 10, 20, or 50 did not lead to EBV accuracies different from when individual progeny data were used. In a simulation of two traits, Alexandre et al. (2019) investigated pooling strategies based on trait 1, trait 2, a combination of both, or randomly to estimated GEBV. In contrast to the current study, pools were not reformed for individual traits, nor was a bivariate model used. Accuracies of GEBV of sires, estimated as the correlation of GEBV and TBV within a trait, were greatest when pools were constructed on the trait itself and lowest when pools were constructed randomly. Alexandre et al. (2020) investigated the use of pooling using Angus data in silico using three traits. The genomic EBV were again calculated using univariate models. Accuracy of GEBV were calculated as the correlation between the sire's GEBV with pooled progeny data and the sire's GEBV using individual progeny data. Pooling strategies employed by Alexandre et al. (2020) were 1) random pooling and 2) by phenotype – which is equivalent to minimizing phenotypic variation within pools in the current study. All three traits were not recorded across all animals which hindered the calculation of GEBV accuracy for one trait when the pools were constructed based on another trait. Regardless, they also found pooling by trait led to larger GEBV accuracies than pooling randomly. Alexandre et al. (2019) suggested pool sizes of 10 in order to compromise the loss in GEBV accuracy and cost saving of pooling; Alexandre et al. (2020) suggested this could be extended to pool sizes greater than 10. Pool sizes of 1, 2, 5, 10, 15, 20, and 25 were investigated; even pool sizes of 25 did not lead to unreasonable losses of GEBV accuracies compared to individual data. In a study investigating the efficiency of

estimated genomic relationships of pools to the animals that make up the pools and to other potentially related individuals, Kuehn et al. (2018) suggested pools of at least 20 to lessen pool construction error.

4.4.5 Missing records

Table 4.1 contains the least-squares means of the percentage of records available nested within how the missing records arose. For Trait 1, the EBV accuracies resulting from different rates of sequential culling were not significantly different for sires or dams born in generations 13 or 14, meaning that the EBV accuracies of Trait 1 did not increase or decrease as more animals were sequentially culled. This is not surprising as 100 percent of Trait 1 records entered the evaluation for these scenarios. However, when looking at the EBV accuracies of Trait 2 resulting from different rates of sequential culling, all pairwise comparisons were significant. This means that as more animals were sequentially culled, the EBV accuracies of Trait 2 decreased significantly. When records were randomly missing, pairwise comparisons were significant, meaning that as the percentage of available records increased, so did the EBV accuracies. Even though these comparisons were statistically significant, the numerical increase in EBV accuracy were small, typically only by 0.1 from 80% to 90% available records or 90% to 100% available records. It is important to note that these least-squares means were averaged over pool sizes, pooling strategy, genetic correlation, and the generation in which genotyping stopped. Overall, as more records were available, the EBV accuracies of the traits increased.

The effect of the population size on EBV accuracy has been well documented in literature. When GEBV are estimated by SNP effect estimates summed across an individual's genome, the more phenotypic records that are available to estimate the SNP effects, the more accurate genomic selection will be (Hayes et al., 2009). Daetwyler et al. (2010) also found that larger numbers of individuals in the training set led to larger EBV accuracies when both GBLUP and BayesB were used. Abdollahi-Arpanahi et al. (2015) reported that as the training population (the number of individuals with genotypes and phenotypes) increased, the correlation between predicted and observed values increased.

Guo et al. (2014) studied the difference in the reliabilities of GEBV, measured as the squared correlation between GEBV and TBV, of two traits using all available data or assuming 90% of the EBV for the first trait was not used for genomic selection or 90% of the EBV for second trait was not used for genomic selection. The GEBV were estimated using GBLUP where the response variables were traditional EBV. The first trait had a heritability of 0.3 while the second trait had a heritability of 0.05 and the genetic correlation was 0.5. When there were missing records for the first trait, the reliability of GEBV decreased by 0.258 as compared to when both traits were recorded on all animals. When there were missing records for the second trait, the reliability of GEBV decreased by 0.171 as compared to when both traits were recorded on all animals.

The interaction of the generation in which genotyping stopped and the percentage of missing records nested within how the missing records arose was significant for EBV accuracies of Trait 2 for sires born in generation 14 and also for the EBV accuracies of both traits for dams born in generation 14 (results not shown). The largest numerical differences resulted from comparisons made between whether genotyping stopped at

generation 13 or 14, which is not surprising given the significant effect of this factor on EBV accuracy which was previously discussed.

When pools were constructed in order to minimize phenotypic variation, pools of any size generally led to larger accuracies than when data from generation 15 did not enter the evaluation. This was regardless of how the missing values arose or the percentage of available records. These are encouraging results suggesting that missing values do not affect the usefulness of pooling.

4.4.6 Genetic correlation

The interaction of the generation in which genotyping stopped and the genetic correlation between the two traits was significant for sires and dams born in generation 14 for both traits. Again, the largest numerical differences arose from comparisons of when genotyping stopped at generation 13 and 14. The interaction between the genetic correlation and the way in which the missing records arose was significant for some trait, sire/dam, and generations in which they were born combinations. Although this interaction was statistically significant, numerically the differences were not large, usually ranging from 0.01 to 0.03 (results not shown). The largest difference (0.05) was observed for the EBV accuracy of Trait 2 for sires born in generation 13 when sequential culling was initiated and comparing across genetic correlations of 0.4 and 0.7. Jia and Jannink (2012) investigated the effect genetic correlation had on the prediction accuracy of two traits with multi-trait genomic selection within simulation. One trait had a heritability of 0.1 while the other had a heritability of 0.8. As the genetic correlation increased, the prediction accuracy of the lowly heritable trait increased; however, the

highly heritable trait saw no increase in prediction accuracy even as the genetic correlation increased between 0.1 and 0.9. In the current study, the effect of genetic correlation on EBV accuracy did not lead to large numerical differences given the moderate heritability of the traits.

Across all genetic correlations, the generations in which the sires and dams were born in, and Traits 1 and 2, the EBV accuracy consistently decreased by 0.01 when randomly missing records decreased from 100% to 90% and then again from 90% to 80%. Thus, randomly missing records did not make a large impact on EBV accuracy across the studied genetic correlations. Additionally, when considering Trait 1 for sires and dams during sequential culling, the differences in EBV accuracy was generally in the range of 0.01 regardless of the percentage of animals culled and genetic correlation. Therefore, sequential culling did not have an impact on the EBV accuracies for Trait 1. The differences in EBV accuracies for Trait 2 considering no culling to 25% of Trait 2 recorded was the smallest (0.06) for sires born in generation 14 and genetic correlation of 0.7. All other differences in EBV accuracy for sires and dams across the genetic correlations was approximately 0.12. In general, the EBV accuracies of Trait 2 when considering sequential culling increased as the percentage of data increased, regardless of genetic correlation. Consequently, as more records were available due to less sequential culling, the EBV accuracies of Trait 2 approached the EBV accuracies of Trait 1.

It would be expected that the EBV accuracies would be approximately equal across different genetic correlations and sires/dams, especially when considering Trait 1 EBV accuracy during sequential culling or missing records when all records are available (100%). This is because regardless of genetic correlation, both Trait 1 and 2 were

simulated to have a heritability of 0.4. However, when considering sires born in generations 14 when genotyping was through generation 14 and for the genetic correlations of 0.1 and 0.4, the EBV accuracies of Traits 1 and 2 were not the same. This was likely due to a larger TBV variance for Trait 2.

4.4.7 EBV accuracy of pools

Even though pools were constructed by trait, all pools received EBV for both traits. Figure 4.4 depicts the median EBV accuracies of the pools that were determined by Trait 1 and Figure 4.5 depicts the median EBV accuracies of the pools that were determined by Trait 2. Significant interactions were quite varied depending on the which trait was observed and which trait the pools were determined by. For example, when considering pools for Trait 1 and the EBV accuracy of Trait 1, significant interactions only included pool size by pooling strategy and genetic correlation by the percentage of available records nested within how they the missing records arose. However, when considering pools for Trait 1 and the EBV accuracy of Trait 2, nearly all possible interactions were significant. When considering pools for Trait 2 and the EBV accuracy of either trait, nearly all interactions involving pool size and pooling strategy were significant.

A few conclusions can be drawn about the EBV accuracies of the pools. As long as the pools were constructed to minimize phenotypic variation, the EBV accuracy of the pools was generally highest for pool sizes of 100 and lowest for pool sizes of 1 for the trait in which the pools were made for. This is consistent with Baller et al. (2020). When the genetic correlation between the traits was high (0.7), the same pattern was true for the

correlated trait. In fact, the EBV accuracy were almost as high for the correlated trait as the EBV accuracies the pools made for. As the genetic correlation decreased to 0.4, the EBV accuracy of the correlated trait began to decrease, especially compared to the EBV accuracy of the trait the pools were made for (results not shown). The EBV accuracy of any pool size was generally larger than pool size of 1. When considering the genetic correlation of 0.1, the EBV accuracies of pools for the alternate trait resulting from any pool size was approximately the same. When considering sequential culling and a genetic correlation of 0.1, the EBV accuracies of pools of 100, 50, and 20 were less than the accuracy from a pool size of 1. When considering pools formed randomly, the EBV accuracies of pools generally increased as pool size decreased; this is also consistent with Baller et al. (2020). This was true for both traits regardless of which trait the pools were made for.

4.5 Conclusions

The results presented in this paper demonstrate the usefulness of pooled data in genetic evaluations that employ a bivariate model using single-step GBLUP across a range of genetic correlations and scenarios in which missing values can arise. Similar to the univariate case, when pools were constructed to minimize phenotypic variation, pool sizes of at least 20 could be used to attain EBV accuracies not significantly different than those attained from individual data. Larger pool sizes (50 and 100) also led to improvement of EBV accuracies for sires born the generation directly before pooling was initiated. When 100% of the phenotypes were available, sires and dams had the highest EBV accuracies. As the percentage of phenotypes decreased due to randomly missing

records, the EBV accuracies of the sires and dams also decreased, but the numerical differences were not large (0.01). Thus, the percentage of randomly missing records investigated in this study did not practically impact the EBV accuracies, regardless of genetic correlation. Additionally, the accuracy of Trait 1 did not decrease with sequential culling because the number of phenotypes available stayed the same for Trait 1 across all sequential culling scenarios. The largest impact of missing records was seen with sequential culling and Trait 2. As the percentage of Trait 2 phenotypes decreased due to the sequential culling, the EBV accuracy of Trait 2 also decreased, regardless of genetic correlation. Consequently, as more records were available due to less sequential culling, the EBV accuracies of Trait 2 approached the EBV accuracies of Trait 1. When considering pooling by minimizing phenotypic variation and a genetic correlation of 0.7, the EBV accuracy of pools was almost as high for the correlated trait as the EBV accuracies the pools were made for. As the genetic correlation decreased, the EBV accuracy of the correlated trait decreased, especially compared to the EBV accuracy of the trait the pools were made for. The results herein provide encouraging conclusions that as long as pools are made to minimize phenotypic variation, pooling can be used across a variety of genetic correlations and ways in which missing values arise to garner the use of commercial ERT within genetic evaluations.

4.6 Acknowledgements

The authors would like to acknowledge the Holland Computing Center at the University of Nebraska-Lincoln for their assistance and use of computational resources.

4.7 Literature Cited

Assessment of bagging GBLUP for whole-genome prediction of broiler chicken traits. J. Anim. Breed. and Genet. 132:218-228. doi:10.1111/jbg.12131.

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. doi:10.3168/jds.2009-2730.

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769. doi:10.1093/jas/skz344.

Alexandre, P. A., A. Reverter, S. A. Lehnert, L. R. Porto-Neto, and S. Dominik. 2020. In silico validation of pooled genotyping strategies for genomic evaluation in Angus cattle. J. Anim. Sci. 98. doi:10.1093/jas/skaa170.

Baller, J. L., S. D. Kachman, L. A. Kuehn, M. L. Spangler. 2020. Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J. Anim. Sci. 98. doi:10.1093/jas/skaa184.

Baller, J. L., J. T. Howard, S. D. Kachman, and M. L. Spangler. 2019. The impact of clustering methods for cross-validation , choice of phenotypes , and genotyping strategies on the accuracy of genomic predictions. J. Anim. Sci. 97:1534–1549. doi:10.1093/jas/skz055.

Bell, A. M., J. M. Henshall, L. R. P. Neto, S. Dominik, R. Mcculloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7. doi:10.1186/s12711-017-0303-8.

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.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2. doi:10.1186/1297-9686-42-2.

Daetwyler, H. D., R. Pong-wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 185:1021–1031. doi:10.1534/genetics.110.116855.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics. 138:1365–1373. doi:10.1007/bf00222881.

Fisher, P. J., D. Turic, N. M. Williams, P. Mcguffin, P. Asherson, D. Ball, I. Craig, T. Eley, L. Hill, K. Chorney, M. J. Chorney, C. P. Benbow, D. Lubinski, R. Plomin, and M. J. Owen. 1999. DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. Hum. Mol. Genet. 8:915–922. doi:10.1093/hmg/8.5.915.

Gaj, P., N. Maryan, E. E. Hennig, J. K. Ledwon, A. Paziewska, A. Majewska, J. Karczmarski, M. Nesteruk, J. Wolski, A. A. Antoniewicz, K. Przytulski, A. Rutkowski, A. Teumer, G. Homuth, T. Starzynska, J. Regula, and J. Ostrowski. 2012. Pooled sample-based GWAS : A cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population. PLoS One. 7. doi:10.1371/journal.pone.0035307.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, S. J. Welham, and R. Thompson. 2015. ASReml User Guide Release 4.1 Functional Specification. VSN International. Hemel Hempstead, United Kingdom. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-4.1-Functional-Specification.pdf.

Guo, G., F. Zhao, Y. Wang, Y. Zhang, L. Du, G. Su. 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet. 15. doi:10.1186/1471-2156-15-30.

Henshall, J. M., R. J. Hawken, S. Dominik, and W. Barendse. 2012. Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. Genet. Sel. Evol. 44:1–13. doi:10.1186/1297-9686-44-12.

Howard, J. T., F. Tiezzi, J. E. Pryce, and C. Maltecca. 2017. Geno-Diver: 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.

Huang, W., B. W. Kirkpatrick, G. J. M. Rosa, and H. Khatib. 2010. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Anim. Genet. 41:570–578. doi:10.1111/j.1365-2052.2010.02046.x.

Jia, Y. and J. L. Jannink. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 192:1513–1522. doi:10.1534/genetics.112.144246.

Kuehn, L. A., T. G. McDaneld, J. W. Keele. 2018 Quantification of genomic relationship from DNA pooled samples. In: Proceedings of the World Congress on Genetics Applied to Livestock Production; February 12 to 16; Auckland, New Zealand. http://www.wcgalp.org/proceedings/2018/quantification-genomic-relationship-dna-pooled-samples. Accessed 11 June 2020.

Lourenco, D. A. L., S. Tsuruta, B. O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J. K. Bertrand, T. S. Amen, L. Wang, D. W. Moser, and I. Misztal. 2015. Genetic

evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653–2662. doi:10.2527/jas2014-8836.

McDaneld, T. G., L . A. Kuehn, M. G. Thomas, W. M. Snelling, T. S. Sonstegard, L. K. Matukumalli, T. P. L. Smith, E. J. Pollak, and J. W. Keele. 2012.Y are you not pregnant: Identification of Y chromosome segments in female cattle with decreased reproductive efficiency. J. Anim. Sci. 90:2142-2151. doi:10.2527/jas.2011-4536.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available https://www.R-project.org/.

Reverter, A., L. R. Porto-Neto, M. R. S. Fortes, R. Mcculloch, R. E. Lyons, S. Moore, D. Nicol, J. Henshall, and S. A. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree 1. J. Anim. Sci. 94:4096–4108. doi:10.2527/jas2016-0675.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: A tool for large-scale association studies. Nat. Rev. Genet. 3:862–871. doi:10.1038/nrg930.

Sonesson, A. K., T. H. E. Meuwissen, and M. E. Goddard. 2010. The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. Genet. Sel. Evol. 42:1–9. doi:10.1186/1297-9686-42-41.

Strillacci, M. G., E. Frigo, F. Schiavini, A. B. Samoré, F. Canavesi, M. Vevey, M. C. Cozzi, M. Soller, E. Lipkin, and A. Bagnato. 2014. Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. BMC Genet. 15. doi:10.1186/s12863-014-0106-7.

Table 4.1. Least-squares means estimates of EBV accuracies due to the percent of missing records nested within how the missing records arose

		Trait 1[3]				Trait 2[4]			
		Sire		Dam		Sire		Dam	
Missing Records[1]	Percent Available[2]	14[5]	13[6]	14	13	14	13	14	13
Random Missing	80%	0.84[a]	0.93[a]	0.82[a]	0.90[a]	0.84[a]	0.93[a]	0.82[a]	0.90[a]
	90%	0.85[b]	0.93[a]	0.83[b]	0.90[b]	0.84[a]	0.94[ab]	0.83[b]	0.91[b]
	100%	0.86[b]	0.94[b]	0.84[c]	0.91[c]	0.85[b]	0.94[b]	0.84[c]	0.91[c]
Sequential Culling	25%	0.85[a]	0.94[a]	0.84[a]	0.91[a]	0.75[a]	0.84[a]	0.73[a]	0.81[a]
	50%	0.85[a]	0.94[a]	0.84[ab]	0.91[a]	0.80[b]	0.90[b]	0.79[b]	0.87[b]
	75%	0.85[a]	0.94[a]	0.84[ab]	0.91[a]	0.83[c]	0.93[c]	0.82[c]	0.90[c]
	100%	0.86[a]	0.94[a]	0.84[b]	0.91[a]	0.85[d]	0.94[d]	0.84[d]	0.91[d]
Std. Error		0.007	0.004	0.005	0.001	0.005	0.016	0.006	0.005

[1]Random Missing = Missing records occur randomly; Sequential Culling = Missing records occur because of sequential culling
[2]80% = 80% of Trait 1 and Trait 2 records are available; 90% = 90% of Trait 1 and Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available; 25% = 100% of Trait 1 records and 25% of Trait 2 records are available; 50% = 100% of Trait 1 records and 50% of Trait 2 records are available; %75 = 100% of Trait 1 records and 75% of Trait 2 records are available
[3]EBV accuracy of Trait 1
[4]EBV accuracy of Trait 2
[5]Sires or dams born in generation 14
[6]Sires or dams born in generation 13
[abcd] Within a column and missing records scenario, least-squares means with the same letter are not significantly different $\alpha = 0.05$.

Figure 4.1. Correlation of average phenotype and average true breeding value (TBV) in pools. Pools resulting from different genetic correlations, how missing records occur (Random Missing = Missing records occur randomly; Sequential Culling = missing records occur because of sequential culling), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools), percentage of available records (80% = 80% of Trait 1 and Trait 2 records are available; 90% = 90% of Trait 1 and Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available; 25% = 100% of Trait 1 records and 25% of Trait 2 records are available; 50%= 100% of Trait 1 records and 50% of Trait 2 records are available; 75% = 100% of Trait 1 records and 75% of Trait 2 records are available), and pool sizes

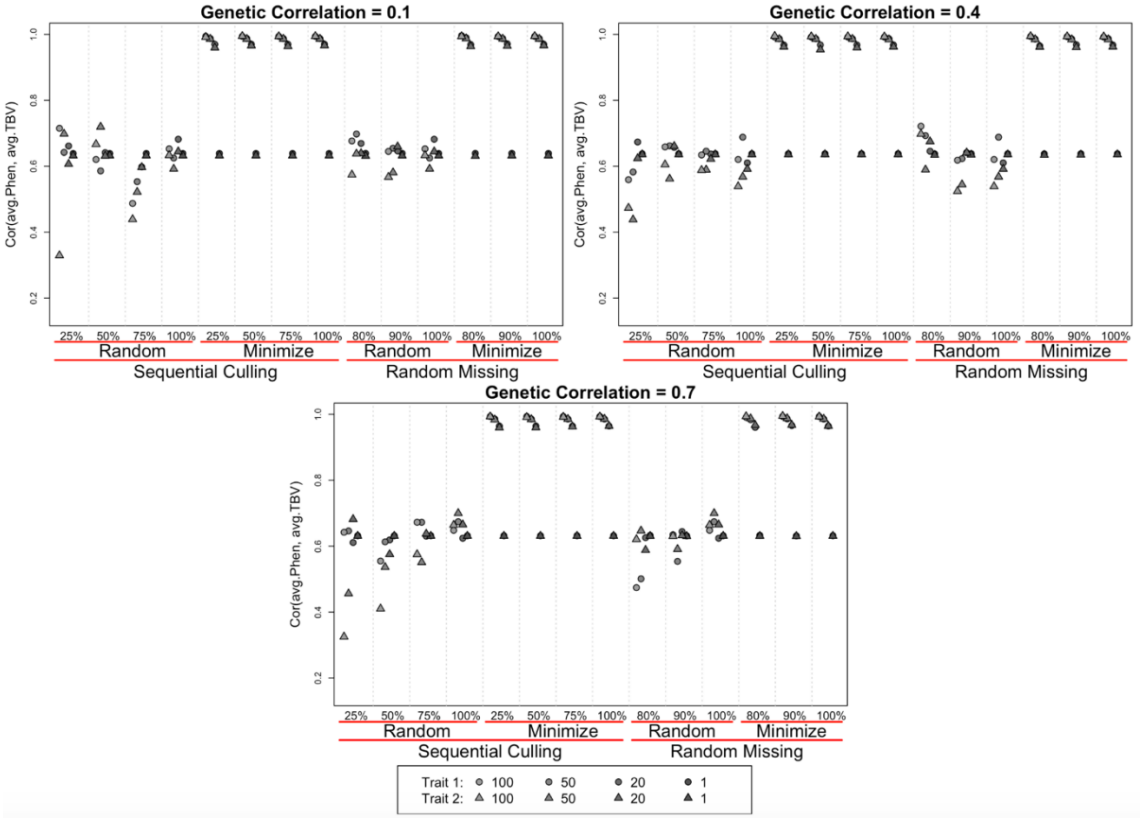


Figure 4.2. Use of sequential culling leading to estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value (TBV) and EBV). Presented by generation of birth resulting different genetic correlations, pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools), percent of available records (25% = 100% of Trait 1 records and 25% of Trait 2 records are available; 50%= 100% of Trait 1 records and 50% of Trait 2 records are available; 75% = 100% of Trait 1 records and 75% of Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available), different generational gaps in genotyping (Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped) and pool sizes with ranges along x-axis

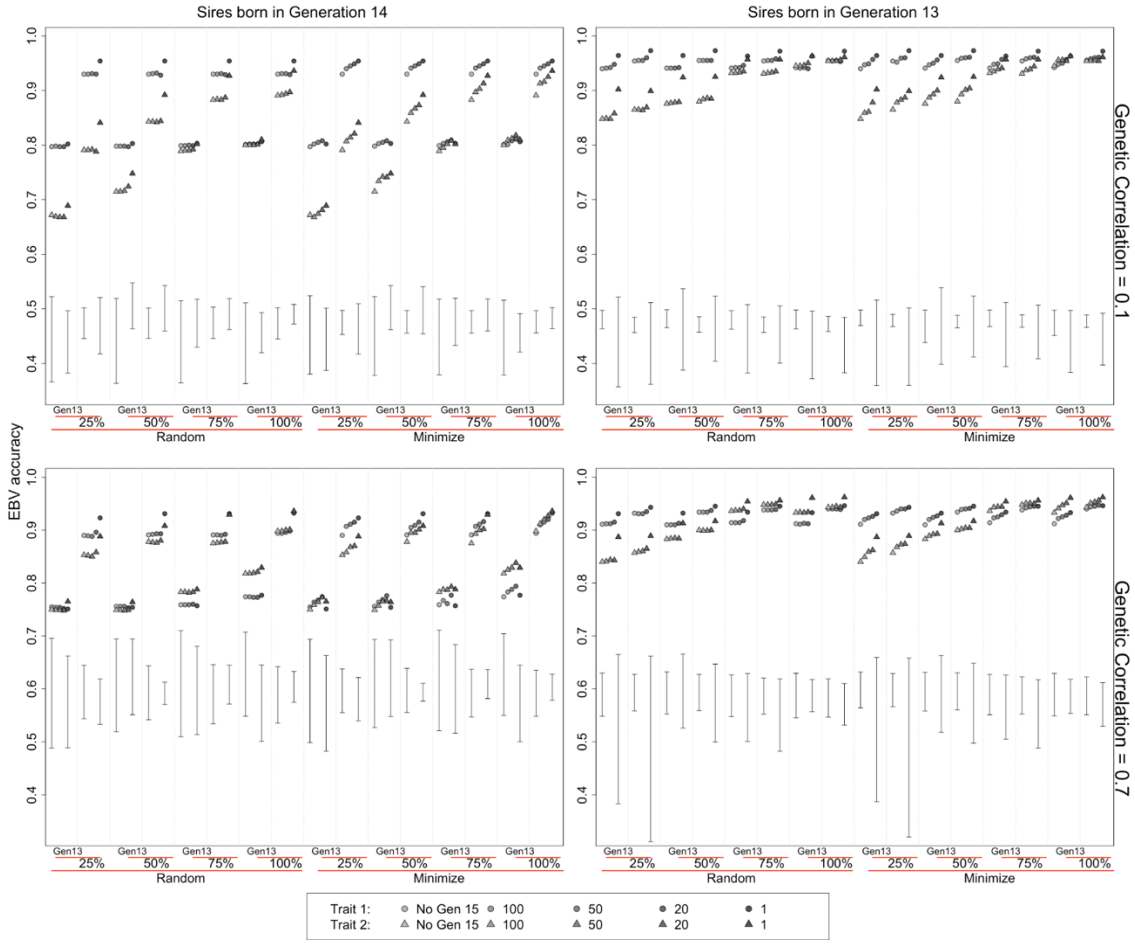


Figure 4.3. Use of randomly missing records leading to estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value (TBV) and EBV). Presented by generation of birth resulting from different genetic correlations, pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools), percent of available records (80% = 80% of Trait 1 and Trait 2 records are available; 90% = 90% of Trait 1 and Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available), different generational gaps in genotyping (Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped) and pool sizes with ranges along x-axis

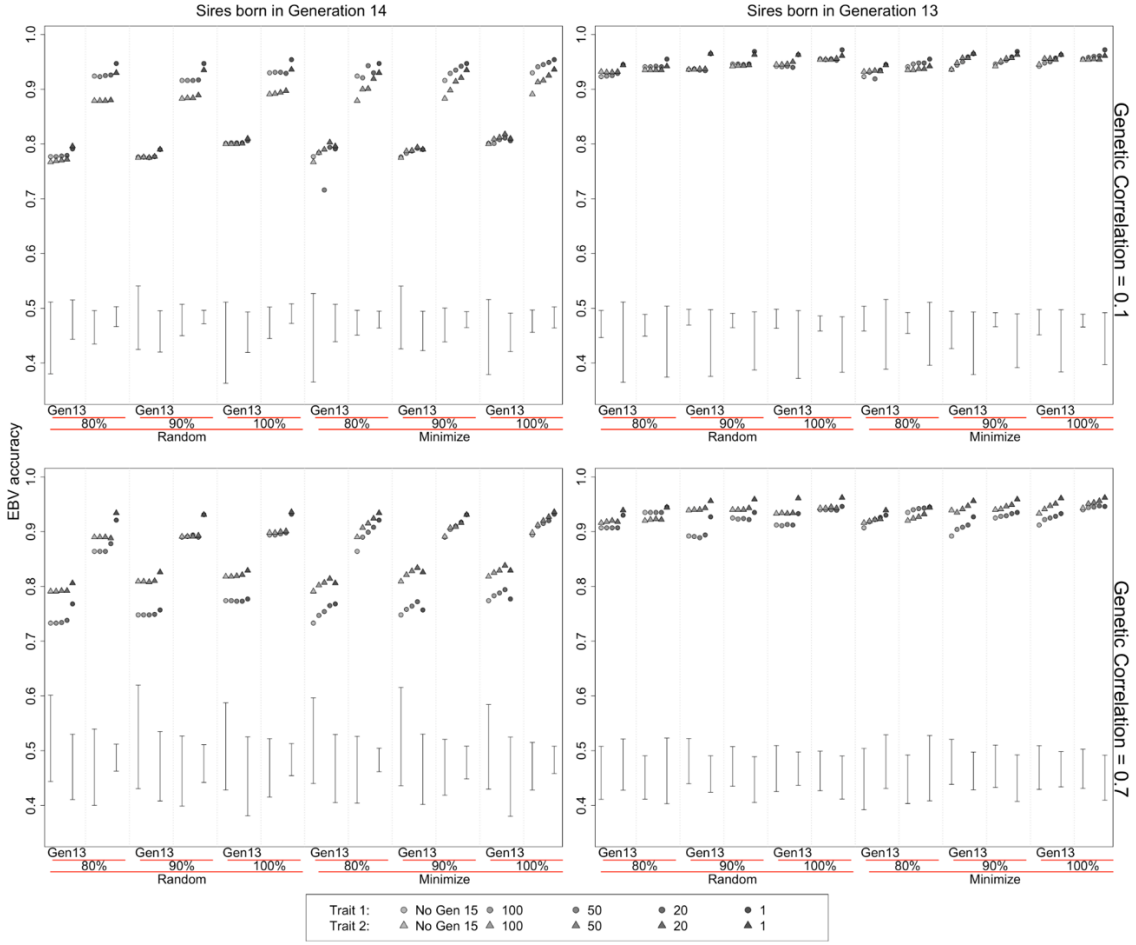


Figure 4.4. Trait 1 pools' estimated breeding value (EBV) accuracies (estimated as the correlation between the average true breeding value (TBV) of the individuals within the pool and EBV of the pool). Pools resulting from different genetic correlations, how missing records occur (Random Missing = Missing records occur randomly; Sequential Culling = missing records occur because of sequential culling), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools), percent of available records (80% = 80% of Trait 1 and Trait 2 records are available; 90% = 90% of Trait 1 and Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available), different generational gaps in genotyping (Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped) and pool sizes with ranges along x-axis

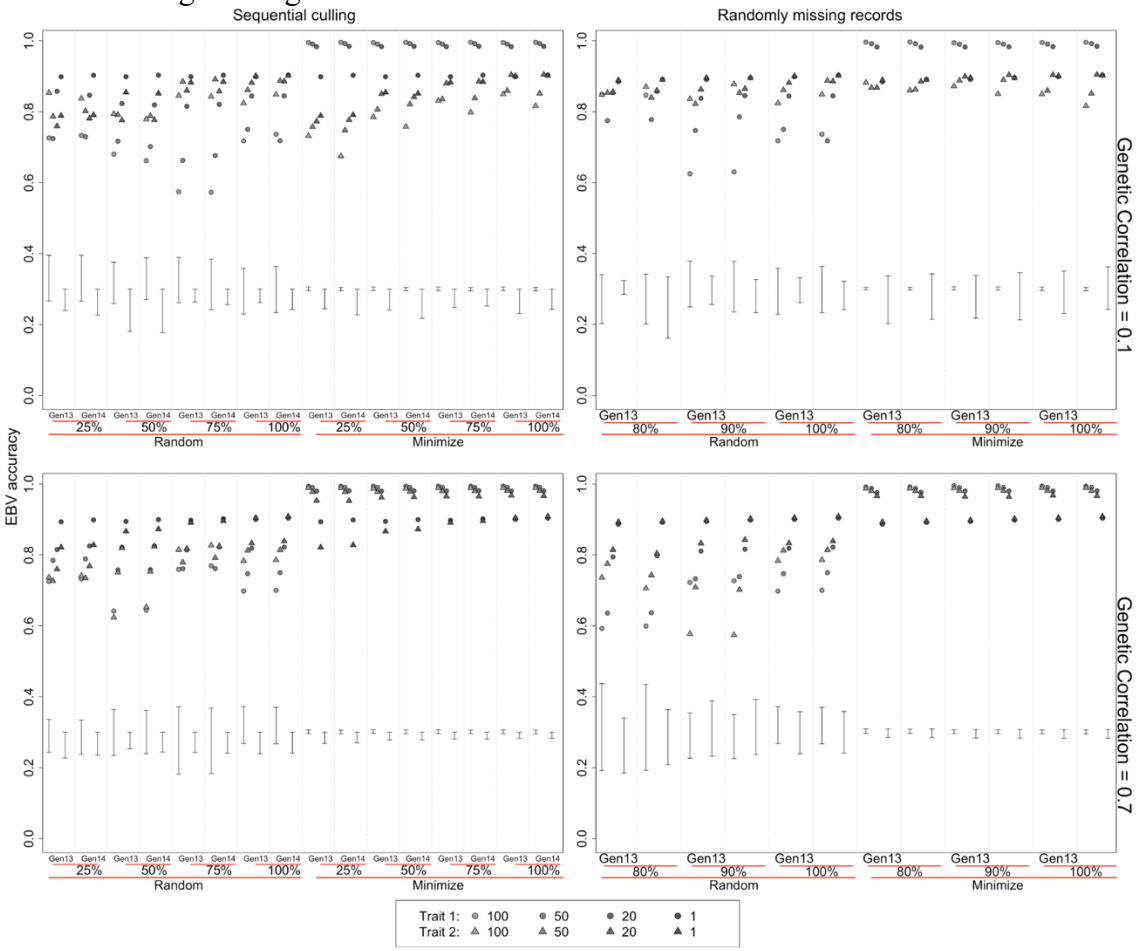
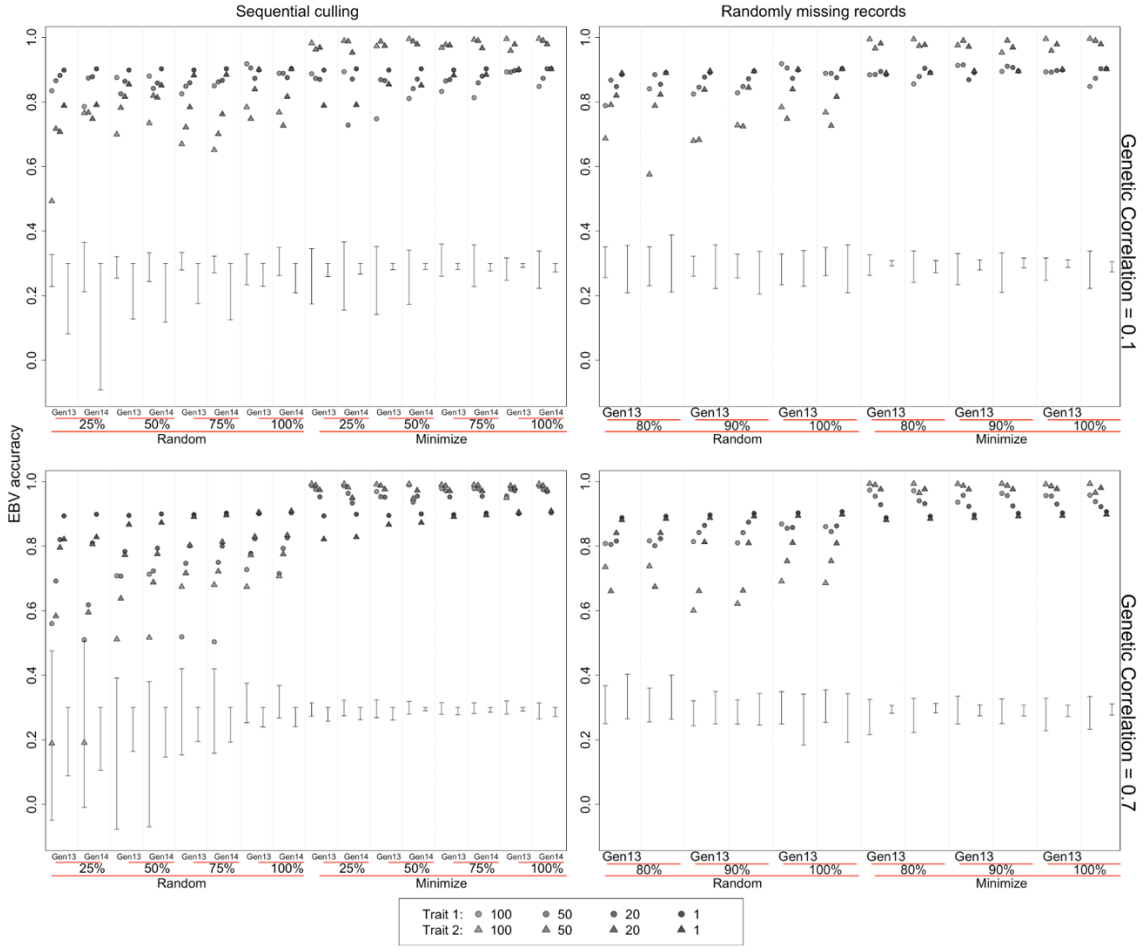


Figure 4.5. Trait 2 pools' estimated breeding value (EBV) accuracies (estimated as the correlation between the average true breeding value (TBV) of the individuals within the pool and predicted EBV of the pool). Pools resulting from different genetic correlations, how missing records occur (Random Missing = Missing records occur randomly; Sequential Culling = missing records occur because of sequential culling), pooling strategies (Random = randomly allocated to pools; Minimize = minimize phenotypic variation within pools), percent of available records (80% = 80% of Trait 1 and Trait 2 records are available; 90% = 90% of Trait 1 and Trait 2 records are available; 100% = 100% of Trait 1 and Trait 2 records are available), different generational gaps in genotyping (Gen13 = individuals up to and including those born in generation 13 were genotyped; Gen14 = individuals up to and including those born in generation 14 were genotyped) and pool sizes with ranges along x-axis



Chapter 5

SYNTHESIS

One of the most important objectives of animal breeding is to accelerate the rate of genetic change within livestock populations. Accuracy of selection is a major contributor to this rate of change. In the second chapter, we showed deregressed expected progeny differences (DEPD) and random genotyping led to the largest estimated accuracies of molecular breeding values (MBV). Using cross-validation, random clustering also led to the largest estimated accuracies, while clustering by k-means and k-medoids led to the lowest accuracies. However, within simulation, no clustering method was associated with more or less bias. Cross-validation is used to assess the predictive ability of a model with data not used to train it. The use of k-means clustering was deliberately aimed to maximize the relationships within clusters and minimize relationships across clusters (Saatchi et al., 2011). This too was the aim of all other clustering methods, including principal component (PC) analysis, on the numerator relationship matrix (Boddhireddy et al., 2014). However, the estimate of bias for clustering methods was likely related to the ability of the methods to minimize and maximize relationships within and across clusters. Moving forward, a better estimate of bias associated with clustering methods could be assessed in simulations representing admixed instead of purebred populations. This would allow the relationships to be minimized within clusters, but more importantly, the relationships across clusters would be further maximized in comparison to a single-breed simulation.

Even though a majority of genetic evaluations in the beef industry now employ single-step methods (e.g. Misztal and Lourenco, 2018; Golden et al., 2018), the work of Chapter 2 was relevant at the time it was written. Cross-validation is also still used to validate causal variants.

Chapters 3 and 4 explored estimated breeding value (EBV) accuracies of sires and dams using single-step genomic best linear unbiased prediction (GBLUP) and the use of pooling data to potentially integrate economically relevant traits (ERT) from the commercial industry. We showed in both univariate and bivariate cases that pooling to minimize phenotypic variation within pool sizes of at least 20 could be used to achieve accuracies not significantly different from those attained from individual data. Additionally, in the bivariate case, it was shown pooling could be used across a variety of scenarios in which missing values arise and a range of genetic correlations between the two traits of interest. Previous work by Reverter et al. (2016) and Bell et al. (2017) demonstrated the use of pooling in real livestock populations. However, the resulting genomic EBV (GEBV) could not be compared to GEBV attained from individual data; therefore, it is not known how large of a loss of accuracy exists in real populations due to pooling. Alexander et al. (2020) quantified this loss using in silico Angus data. Further research of pooling compared to individual data in the use of genetic evaluations should be conducted to fully validate this methodology.

The simulations within Chapters 3 and 4 mimicked a purebred cattle population where animals were genotyped with a single SNP chip (50k SNP) and phenotyped for quantitative traits. Further research could include threshold traits, such as disease or temperament. It may be reasonable to assume pools of commercial animals and

sires/dams within seedstock are genotyped at different densities. In this case, pools may be genotyped with a low-density chip to drive genotyping costs even further down, while sires/dams in the seedstock herds are genotyped with high-density chips, or even sequenced. This type of scenario requires further research into the imputation of pooling allele frequencies (PAF). Lastly, genetic evaluations of cattle can include purebred individuals as well as crossbred individuals. Therefore, further research could include the simulation and validation of pooling within admixed populations. In the case of admixed populations, it would be possible to estimate breed fractions but not heterosis. Regardless, the results herein provide promising results for the use of pooling commercial data within genetic evaluations. Moving forward, it is reasonable to assume the National Cattle Evaluation (NCE) could incorporate pooled commercial data in order to include true ERT phenotypes to increase the accuracy of seedstock animals.

5.1 Literature Cited

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769. doi:10.1093/jas/skz344.

Bell, A. M., J. M. Henshall, L. R. P. Neto, S. Dominik, R. Mcculloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7. doi:10.1186/s12711-017-0303-8.

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 1. J. Anim. Sci. 92:485–497. doi:10.2527/jas2013-6757.

Golden, B. L., M. L. Spangler, W. M. Snelling, and D. J. Garrick. 2018. Current single-step national beef cattle evaluation models used by the American Hereford Association and International Genetic Solutions, computational aspects, and implications of marker selection. Proc. 11th Genetic Prediction Workshop. Kansas City, MO. p. 14-22.

Misztal, I. and D. Lourenco. 2018. Current research in unweighted and weighted ssGBLUP. Proc. 11th Genetic Prediction Workshop. Kansas City, MO. p. 1-13.

Reverter, A., L. R. Porto-Neto, M. R. S. Fortes, R. Mcculloch, R. E. Lyons, S. Moore, D. Nicol, J. Henshall, and S. A. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree 1. J. Anim. Sci. 94:4096–4108. doi:10.2527/jas2016-0675.

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. M. 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 K-means clustering for cross-validation. Genet. Sel. Evol. 43:40. doi:10.1186/1297-9686-43-40.