5-21-2022

Using pooled data for genomic prediction in a bivariate framework with missing data

Johnna L. Baller

Stephen D. Kachman

Larry A. Kuehn

Matthew L. Spangler

Received: 16 December 2021 | Accepted: 21 May 2022

DOI: 10.1111/jbg.12727

ORIGINAL ARTICLE

Journal of Animal Breeding and Genetics WILEY

Using pooled data for genomic prediction in a bivariate framework with missing data

Johnna L. Baller[1] | Stephen D. Kachman[2] | Larry A. Kuehn[3] | Matthew L. Spangler[1]

[1]Department of Animal Science, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[2]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[3]USDA, ARS, U.S. Meat Animal Research Center, Clay Center, Nebraska, USA

Correspondence
Matthew L. Spangler, Department of Animal Science, University of Nebraska, Lincoln, NE 68583, USA.
Email: mspangler2@unl.edu

Abstract

Pooling samples to derive group genotypes can enable the economically efficient use of commercial animals within genetic evaluations. To test a multivariate framework for genetic evaluations using pooled data, simulation was used to mimic a beef cattle population including two moderately heritable traits with varying genetic correlations, genotypes and pedigree data. There were 15 generations ($n = 32{,}000$; random selection and mating), and the last generation was subjected to genotyping through pooling. Missing records were induced in two ways: (a) sequential culling and (b) random missing records. 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. The EBV was estimated using a bivariate single-step genomic best linear unbiased prediction model. 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 ($p < .05$) for sires and dams born in the generation nearest the pools. 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. Pooling to aid in the use of commercial data 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.

KEYWORDS

beef cattle, bivariate models, DNA pooling, genomic prediction

1 | INTRODUCTION

Most of the data included in beef cattle genetic evaluations in the US are recorded within the nucleus (seedstock) segment; however, often economically relevant traits (ERT) are only observed at the commercial level. Records (phenotypes) are routinely collected at the commercial level but the pedigree relationships needed to connect these records to seedstock animals are often missing due to the 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 all commercial animals with a phenotype would need to be individually genotyped. This level of genotyping would not be economical. Nevertheless, the 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 have been 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 & 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 have 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) was estimated with univariate models.

Choosing animals to pool together in practice might best be facilitated at random, perhaps in part to ensure similar environmental effects or simply for ease of implementation. However, using real data and in silico, there are examples where pools have been constructed attempting to minimize phenotypic variation (Alexandre et al., 2020; Bell et al., 2017; Henshall et al., 2012; Reverter et al., 2016). Differences in pool construction and the impact on genomic prediction have been reported in simulation studies involving one trait (Baller et al., 2020) and two traits (Alexandre et al., 2019), both of which concluded minimizing phenotypic variation within the pools led to the highest accuracies as compared to other pool construction strategies.

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. 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.

2 | MATERIALS AND METHODS

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

2.1 | 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 five replicates representing low genetic correlation (0.1), moderate genetic correlation (0.4) and high genetic correlation (0.7). The QTL effects were generated by sampling from three independent gamma distributions, then the samples were combined to generate the additive effects of Trait 1 and 2 (Howard et al., 2018). 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 10 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).

2.2 | 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. Missing records were simulated across the whole population, not just the last generation where pooling occurred. 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 process, 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 occur randomly 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% of 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 replicate.

2.3 | Pooling

The individuals born in generation 15 (n = 2,000) were assigned to pools. Two sets of pools were independently constructed: the first set was constructed based on Trait 1 records, and the second set was 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. In the case where there were no missing records, pool sizes of 20, 50 and 100 individuals resulted in 200 pools (100 based on Trait 1 and 100 based on Trait 2), 80 pools or 40 pools, respectively. In the case where there were missing records for a trait, the number of pools based on that trait would be proportionally less. Pool assignments were determined in two different ways: (a) randomly or (b) minimizing the phenotypic variation within a pool. Random pools were formed by randomly assigning individuals to a pool based on Trait 1 and to a pool based on Trait 2. For example, for a pool size of 20 and no missing records, an animal would be randomly assigned to two pools, one pool from the 100 pools based on Trait 1 and one pool from the 100 pools based on Trait 2. To construct pools to minimize phenotypic variation within pools, individuals with records for Trait 1 were first ranked based on their phenotypic record for Trait 1 and then grouped together depending on the pool size. This process was then repeated for individuals with a record for Trait 2. For example, with a pool size of 20 and no missing records, 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 101. Pools based on Trait 1 had a phenotypic record for Trait 1 and a missing record for Trait 2 and vice versa. Individuals could only be included in one pool per trait per scenario, where the 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 pool based on a trait was the average phenotype for that trait 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 was 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 entered 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 a 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."

2.4 | Missing generation of genotypes

All parents were assumed to be genotyped even if they did not have a recorded phenotype because of randomly missing records. 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: (a) animals up to and including those born in generation 13 were genotyped (Gen13) and (b) 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.

2.5 | 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 & 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; 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; 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

$$\text{var}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \mathbf{G} \bigotimes \mathbf{H} \text{ and var}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \mathbf{R} \bigotimes \mathbf{I}$$

where **G** is a 2×2 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** ($\mathbf{H}^{-1}$) were 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 genotypes, pedigrees and phenotypes of animals in generation 15 were unknown, but the individual phenotypes of Traits 1 and 2 contributed to the pool means (i.e. individual data were unobserved, but pool means were observed). 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 the ith trait for both individuals and pools; and e_i was a vector of random residuals for individuals and pools based on the ith trait. It was assumed that

$$\text{var}\begin{bmatrix} u_1^* \\ u_2^* \end{bmatrix} = \mathbf{G} \bigotimes \mathbf{H}^* \text{ and var}\begin{bmatrix} e_1^* \\ e_2^* \end{bmatrix} = \mathbf{R} \bigotimes \text{diag}\left(\frac{1}{q}\right)$$

where again $\mathbf{G}$ is a 2×2 matrix containing the variance components for the additive effects, $\mathbf{H}^*$ is a kinship matrix relating individual animals and pools of animals, and $\mathbf{R}$ is a diagonal matrix containing the variances for the residual effects. 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 variance of the residuals is

$$\text{var}\begin{bmatrix} e_1^* \\ e_2^* \end{bmatrix} = \text{diag}\left(\sigma_{ei}^2 / q_{ij}\right),$$

where σ_{ei}^2 is the residual variance for the ith trait and q_{ij} is 1 for an individual record and the pool size for a pooled record. For simplicity, the variance structure for the residuals used in the model assumes that animals are randomly assigned to pools. When pools were formed to minimize the phenotypic variance the assumption of random assignment does not hold, but the variance structure is one that would be used in practice. The inverse of $\mathbf{H}^*$ was constructed the same as $\mathbf{H}$ except that the allelic frequencies were estimated from individuals and pools. Pool constructions and the computation of inverses of $\mathbf{H}$ and $\mathbf{H}^*$ were carried out in R (R Core Team, 2017). Breeding values were estimated in the ASReml v4.1 software (Gilmour et al., 2015) using the preconditioned conjugate gradients (PCG) method.

The accuracy of EBV for sires and dams was estimated as the correlation between the true breeding values (TBV) and the EBV. The accuracies were estimated separately for sires and dams, the 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 final simulated set) 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. The IQR test identifies an observation as an outlier if the observation is either more than $Q_3 + (1.5 \times \text{IQR})$ or less than $Q_1 - (1.5 \times \text{IQR})$, where Q_1 and Q_3 are the first and third quantiles, respectively. All data from a final simulated set with at least one outlier were excluded from the analysis.

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 after excluding all observations from a final simulated set with at least one outlier 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 the 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 that 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 $\alpha = .05$.

2.6 | 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 $\mathbf{G}_{22}^0$ represents the relationships between individuals in generation 15. Similarly, let $\mathbf{G}_{22}^\text{p}$ represents the relationships between the pools. The expected genomic relationship matrix $\mathbf{G}_{22}^\text{p}$ is a function of $\mathbf{G}_{22}^0$ and follows:

$\left\{\mathbf{G}_{22}^\text{p}\right\}_{kk'} = \left(\frac{1}{q} I_k'\right)\left\{\mathbf{G}_{22}^0\right\}_{kk'}\left(\frac{1}{q} I_{k'}\right)$ where $\left\{\mathbf{G}_{22}^\text{p}\right\}_{kk'}$ is the kk' element of $\mathbf{G}_{22}^\text{p}$ corresponding to pools k and k', $\left\{\mathbf{G}_{22}^0\right\}_{kk'}$ is the kk' submatrix of $\mathbf{G}_{22}^0$ corresponding to individuals in pools k and k', and I_k' and $I_{k'}$ 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.

Assuming all individuals in generation 15 are unrelated. From the expectations above it can be seen that for pools of individuals, the diagonal elements of $\mathbf{G}^{p}_{22}$ are equal to $\frac{1}{q}$ and the off-diagonals of $\mathbf{G}^{p}_{22}$ are proportional to $\frac{m}{q^2}$ where m is the number of individuals in common between two pools. Thus, the off-diagonals of $\mathbf{G}^{p}_{22}$ 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 $\mathbf{G}^{P}_{22}$ are expected to be greater than $\frac{1}{q}$ and the off-diagonal elements of $\mathbf{G}^{P}_{22}$ between pools based on different traits will be greater than $\frac{m}{q}$ as the individuals in the pools become more related.

3 | RESULTS AND DISCUSSION

3.1 | Pooling

Figure 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 with 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 was observed with random pooling. However, the range of correlations between average phenotype and TBV was 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 result could be an artefact 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 pooled across traits, it is important to note again

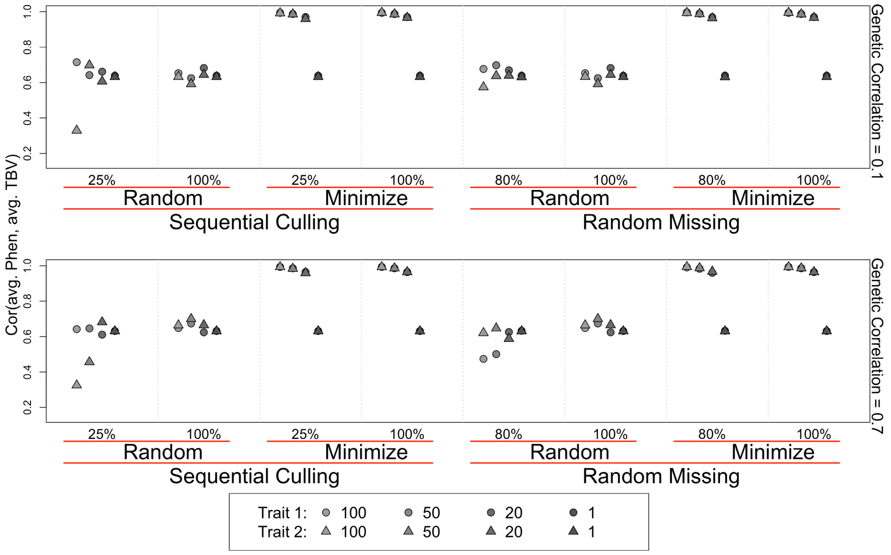

FIGURE 1 Correlation of average phenotype and average true breeding value (TBV) in pools. Pools resulting from different genetic correlations, how missing records occur (random missing = missing records occur randomly; sequential culling = missing records occur because of sequential culling), pooling 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, 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) and pool sizes [Colour figure can be viewed at wileyonlinelibrary.com]

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 increase was due to more animals being included for both traits with more records as it was very unlikely the same animals would randomly have missing records for both traits. 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.

3.2 | EBV accuracies of sires and dams

Figures 2 and 3 depict the median EBV accuracies of sires born in generation 14 for sequential culling and randomly missing records, respectively, depending on genetic correlation, pooling strategy, percentage of missing records and when genotyping stopped at generation 14. Results of dams are not shown as they follow the same patterns as the sires. Although the same patterns are present 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.

3.3 | 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. 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 (results not shown). Baller et al. (2020) also noted that EBV accuracies of sires and dams by the generation 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., 2015).

3.4 | 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 result occurred for sires born in generation 14 for Trait 1 across all pool sizes and was also true for dams born in generation 14 only when pool sizes were of size 20 for Trait 1. For Trait 2, this result occurred for sires born in generations 13 and 14. Significant differences in pool size were likely different for Trait 1 compared with Trait 2 because missing records, especially for sequential culling, were induced for Trait 2. Differences between sires and dams regarding significant differences in pool sizes were likely due to the amount of information available due to the number of progeny each sex had. A less optimal situation would be where the EBV accuracies as a result of 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 (data not shown). Thus, with that small numeric difference, the decreased cost of pooling may still be much more economical in its effect on accuracy than individual genotyping.

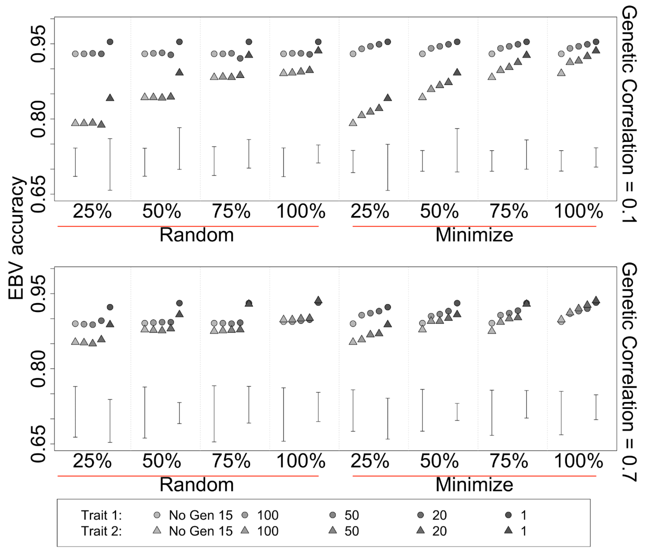

FIGURE 2 Use of sequential culling leading to estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value [TBV] and EBV). Presented sires born in generation 14 with accuracies resulting from different genetic correlations, pooling strategies (random = 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) and pool sizes with ranges in accuracy along the *x*-axis [Colour figure can be viewed at wileyonlinelibrary.com]

Reverter et al. (2016) used pooling within Brahman cattle for pregnancy and lactation status using GBLUP. Cattle were pooled based on results from a pregnancy test in pools of 15–28 individuals. 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 sheep were pooled by sex and dag score category with pool sizes of 33 to 40 individuals. The accuracies of GEBV resulting from pooled data from Bell et al. (2017) or Reverter et al. (2016) were not compared with a baseline of GEBV resulting from individual data, and so, it is not known if the loss of accuracy in prediction due to pooling was significant or not, warranting validation of pooling with simulation.

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 estimate 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 was again calculated using univariate models. Accuracy of GEBV was 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 (a) random pooling and (b) 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

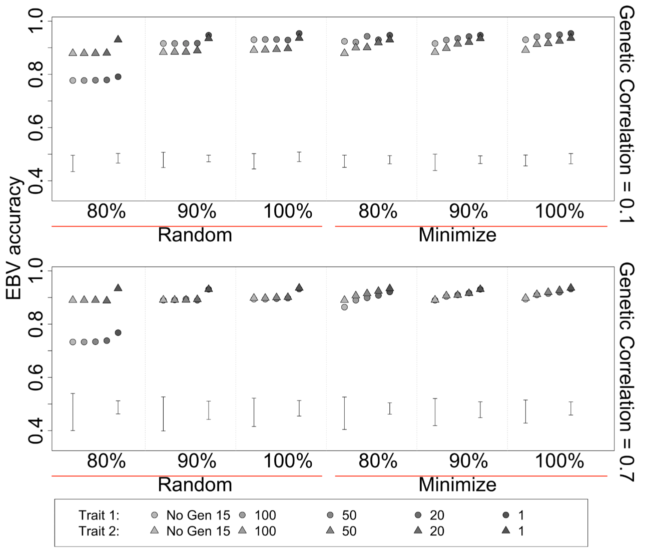

FIGURE 3 Use of randomly missing records leading to estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value [TBV] and EBV). Presented sires born in generation 14 with accuracies resulting from different genetic correlations, pooling strategies (random = 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) and pool sizes with ranges in accuracy along the *x*-axis [Colour figure can be viewed at wileyonlinelibrary.com]

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 with 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.

3.5 | Missing records

Table 1 contains the least-squares EBV accuracy means by the percentage of records available nested within how the missing records arose. As expected, the accuracy of Trait 1 EBV for sires and dams was not impacted by sequential culling given all animals had a Trait 1 phenotype recorded. However, sequential selection impacted Trait 2 EBV accuracy as all pairwise comparisons of percentage of missing records within how the missing records arose were significant. When records were randomly missing, pairwise comparisons of percentage of missing records within how the missing records arose 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 was 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.

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 the second trait was not used for genomic selection. The GEBV was 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 interactions of pool size and pooling strategy with the percentage of missing records nested within how the missing records arose were not significant. This result signifies that the impact of pool size and pooling strategy is not dependent on the percentage of missing records

TABLE 1 Least-squares mean estimates of EBV accuracies due to the percent of missing records nested within how the missing records arose

		Trait 1[§]				Trait 2[¶]			
		Sire		Dam		Sire		Dam	
Missing records[†]	Percent available[‡]	14[††]	13[‡‡]	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

Note: [a,b,c,d]Within a column and missing record scenario, least-square means with the same letter are not significantly different $\alpha = .05$.

[†] Random missing = missing records occur randomly; sequential culling = missing records occur because of sequential culling.

[‡] 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.

[§] EBV accuracy of Trait 1.

[¶] EBV accuracy of Trait 2.

[††] Sires or dams born in generation 14.

[‡‡] Sires or dams born in generation 13.

nested within how the missing records arose, rather they are consistent across those investigated herein. 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 (data 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 missing records on EBV accuracy. Furthermore, the only sources of progeny information for parental animals born in generation 14 were pooled data whereas earlier generations (i.e. generation 13) benefited from offspring with individual records in addition to descendants contained within the pools.

Regardless of how the missing values arose or the percentage of available records, 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. These are encouraging results suggesting that missing values do not affect the usefulness of pooling.

3.6 | Genetic correlation

The interactions of pool size and pooling strategy with genetic correlation were not significant. This result again signifies that the impact of pool size and pooling strategy are not dependent on genetic correlation, rather they are consistent across the genetic correlations investigated herein. 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 generations 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 generation of birth combinations. Although this interaction was statistically significant, numerically the differences were not large, usually ranging from 0.01 to 0.03 (data 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 the 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 the percentage of records available decreased randomly from 100% to 90% and then again from 90% to 80%. Thus, randomly missing records did not have a large impact on EBV accuracy across the studied genetic correlations. Additionally, the accuracy of Trait 1 EBV for sires and dams was negligibly impacted by sequential culling, the differences in EBV accuracy were generally in the range of 0.01 regardless of the percentage of animals culled and the genetic correlation. 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 were approximately 0.12. In general, the EBV accuracies of Trait 2 when considering sequential culling increased as the percentage of culled 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.

3.7 | EBV accuracy of pools

Even though pools were constructed by trait, all pools received EBV for both traits. Figure 4 depicts the median EBV accuracies of the pools that were determined by Trait 1 and Figure 5 depicts the median EBV accuracies of the pools that were determined by Trait 2. Significant interactions were quite varied depending on if observing the trait in which the pools were made or the correlated trait. 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 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

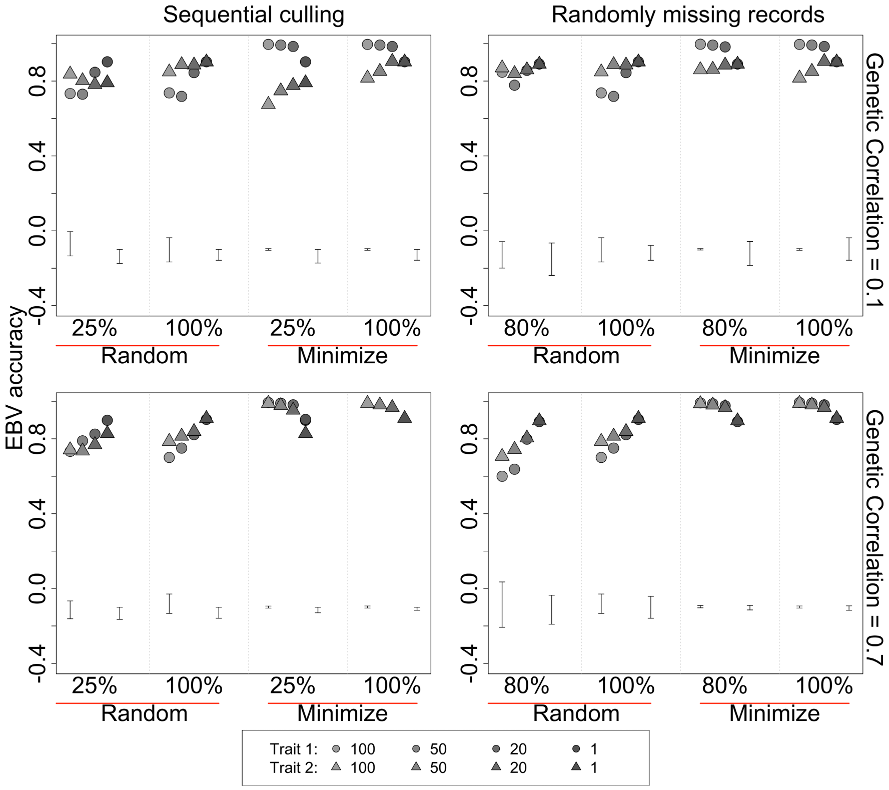

FIGURE 4 Trait 1 pools' estimated breeding value (EBV) accuracies (estimated as the correlation between the average true breeding value [TBV] of the individuals within the pool and EBV of the pool). Pools resulting from different genetic correlations, how missing records occur (random missing = missing 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), individuals up to and including those born in generation 14 were genotyped (Gen14) and pool sizes with ranges in accuracy along the *x*-axis [Colour figure can be viewed at wileyonlinelibrary.com]

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 was 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 with the EBV accuracy of the trait the pools were made for (data not shown). The EBV accuracy of any pool size was generally larger than the 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 were approximately the same. When considering sequential culling and a genetic correlation of 0.1, the EBV accuracies of the correlated trait resulting from 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, which is also consistent with Baller et al. (2020). This is expected given that when pools are formed randomly and pool size increases the variation among pools decreases. This pattern was observed for both traits regardless of which trait the pools were made for.

4 | CONCLUSIONS

The results presented herein 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. There were no significant interactions of pool size or pooling strategy with either percentage of missing records nested within how the missing records arose or genetic correlation, suggesting the robustness of pooling recommendations in the

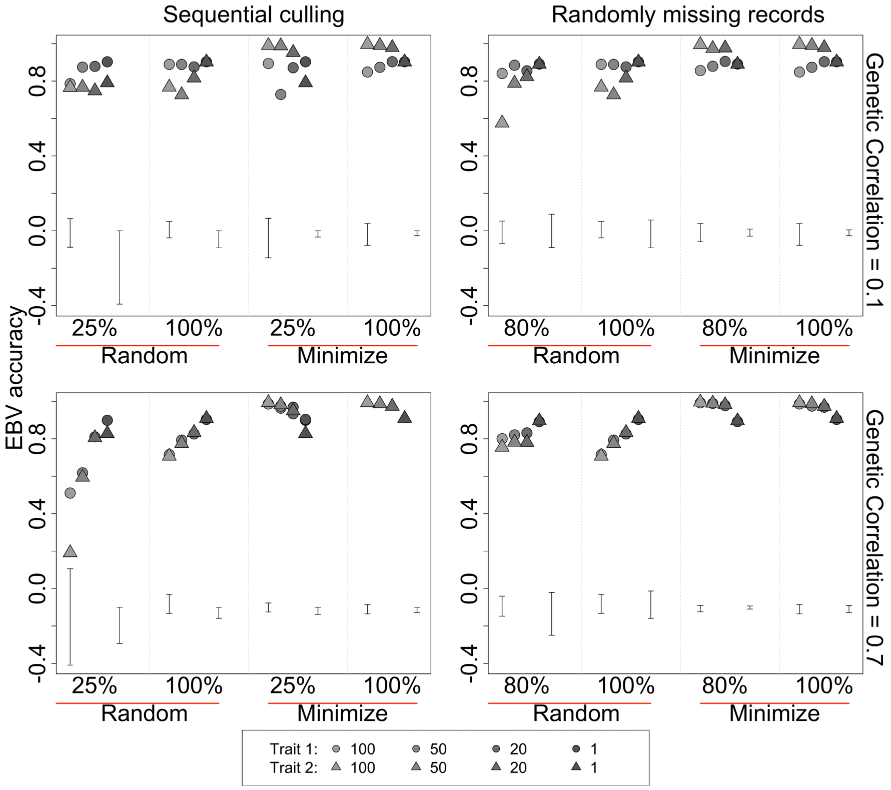

FIGURE 5 Trait 2 pools' estimated breeding value (EBV) accuracies (estimated as the correlation between the average true breeding value (TBV) of the individuals within the pool and predicted EBV of the pool). Pools resulting from different genetic correlations, how missing records occur (random missing = 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), individuals up to and including those born in generation 14 were genotyped (Gen14) and pool sizes with ranges in accuracy along the *x*-axis [Colour figure can be viewed at wileyonlinelibrary.com]

bivariate case or when missing values are present. 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 with 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.

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. This project is based on research that was partially supported by the Nebraska Agricultural Experiment Station with funding from the Hatch Act (Accession Number 1011203) through the USDA National Institute of Food and Agriculture.

CONFLICT OF INTEREST

The authors declare no conflicts of interest to objectively present this research. Mention of a trade name, proprietary product or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable. USDA is an equal opportunity provider and employer.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Matthew L. Spangler https://orcid.org/0000-0001-5184-501X

REFERENCES

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

Alexandre, P. A., Porto-Neto, L. R., Karaman, E., Lehnert, S. A., & Reverter, A. (2019). Pooled genotyping strategies for the rapid construction of genomic reference populations. *Journal of Animal Science*, *97*, 4761–4769. https://doi.org/10.1093/jas/skz344

Alexandre, P. A., Reverter, A., Lehnert, S. A., Porto-Neto, L. R., & Dominik, S. (2020). In silico validation of pooled genotyping strategies for genomic evaluation in Angus cattle. *Journal of Animal Science*, *98*(6), 1–5. https://doi.org/10.1093/jas/skaa170

Baller, J. L., Howard, J. T., Kachman, S. D., & Spangler, M. L. (2019). The impact of clustering methods for cross-validation, choice of phenotypes, and genotyping strategies on the accuracy of genomic predictions. *Journal of Animal Science*, *97*, 1534–1549. https://doi.org/10.1093/jas/skz055

Baller, J. L., Kachman, S. D., Kuehn, L. A., & Spangler, M. L. (2020). Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. *Journal of Animal Science*, *98*, 1–12. https://doi.org/10.1093/jas/skaa184

Bell, A. M., Henshall, J. M., Neto, L. R. P., Dominik, S., Mcculloch, R., Kijas, J., & Lehnert, S. A. (2017). Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. *Genetics Selection Evolution*, *49*, 1–7. https://doi.org/10.1186/s12711-017-0303-8

Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome Research*, *19*, 136–142. https://doi.org/10.1101/gr.083634.108

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, *42*, 2. https://doi.org/10.1186/1297-9686-42-2

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

Fisher, P. J., Turic, D., Williams, N. M., Mcguffin, P., Asherson, P., Ball, D., Craig, I., Eley, T., Hill, L., Chorney, K., Chorney, M. J., Benbow, C. P., Lubinski, D., Plomin, R., & Owen, M. J. (1999). DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. *Human Molecular Genetics*, *8*, 915–922. https://doi.org/10.1093/hmg/8.5.915

Gaj, P., Maryan, N., Hennig, E. E., Ledwon, J. K., Paziewska, A., Majewska, A., Karczmarski, J., Nesteruk, M., Wolski, J., Antoniewicz, A. A., Przytulski, K., Rutkowski, A., Teumer, A., Homuth, G., Starzynska, T., Regula, J., & Ostrowski, J. (2012). Pooled sample-based GWAS: A cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population. *PLoS One*, *7*, e35307. https://doi.org/10.1371/journal.pone.0035307

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J., & Thompson, R. (2015). *ASReml user guide release 4.1 functional specification*. VSN International. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/2018/02/ASReml-4.1-Functional-Specification.pdf

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

Henshall, J. M., Hawken, R. J., Dominik, S., & Barendse, W. (2012). Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples. *Genetics Selection Evolution*, *44*, 1–13. https://doi.org/10.1186/1297-9686-44-12

Howard, J. T., C. Maltecca, F. Tiezzi, J. E. Pryce, and M. L. Spangler. (2018). *Geno-Diver (V3) genetic simulation toolkit*. https://github.com/jeremyhoward/Geno-Diver/blob/master/Geno-Diver%20Manual.pdf

Howard, J. T., Tiezzi, F., Pryce, J. E., & Maltecca, C. (2017). Geno-Diver: A combined coalescence and forward-in-time simulator for populations undergoing selection for complex traits. *Journal of Animal Breeding and Genetics*, *134*, 553–563. https://doi.org/10.1111/jbg.12277

Huang, W., Kirkpatrick, B. W., Rosa, G. J. M., & Khatib, H. (2010). A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Animal Genetics*, *41*, 570–578. https://doi.org/10.1111/j.1365-2052.2010.02046.x

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

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 to 16; Auckland, New Zealand*. http://www.wcgalp.org/proceedings/2018/quantification-genomic-relationship-dna-pooled-samples

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., & Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *Journal of Animal Science*, *93*, 2653–2662. https://doi.org/10.2527/jas2014-8836

McDaneld, T. G., Kuehn, L. A., Thomas, M. G., Snelling, W. M., Sonstegard, T. S., Matukumalli, L. K., Smith, T. P. L., Pollak, E. J., & Keele, J. W. (2012). Y are you not pregnant: Identification of Y chromosome segments in female cattle with decreased reproductive efficiency. *Journal of Animal Science*, *90*, 2142–2151. https://doi.org/10.2527/jas.2011-4536

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reverter, A., Porto-Neto, L. R., Fortes, M. R. S., Mcculloch, R., Lyons, R. E., Moore, S., Nicol, D., Henshall, J., & Lehnert, S. A. (2016). Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. *Journal of Animal Science*, *94*, 4096–4108. https://doi.org/10.2527/jas2016-0675

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., & Owen, M. (2002). DNA pooling: A tool for large-scale association studies. *Nature Reviews. Genetics*, *3*, 862–871. https://doi.org/10.1038/nrg930

Sonesson, A. K., Meuwissen, T. H. E., & Goddard, M. E. (2010). The use of communal rearing of families and DNA pooling in aquaculture genomic selection schemes. *Genetics, Selection, Evolution*, *42*, 1–9. https://doi.org/10.1186/1297-9686-42-41

Strillacci, M. G., Frigo, E., Schiavini, F., Samoré, A. B., Canavesi, F., Vevey, M., Cozzi, M. C., Soller, M., Lipkin, E., & Bagnato, A. (2014). Genome-wide association study for somatic cell score in Valdostana Red Pied cattle breed using pooled DNA. *BMC Genetics*, *15*, 106. https://doi.org/10.1186/s12863-014-0106-7

How to cite this article: Baller, J. L., Kachman, S. D., Kuehn, L. A., & Spangler, M. L. (2022). Using pooled data for genomic prediction in a bivariate framework with missing data. *Journal of Animal Breeding and Genetics*, *139*, 489–501. https://doi.org/10.1111/jbg.12727