Using Pooling to Capture Commercial Data for Inclusion in Genetic Evaluations

Johnna L. Baller
Stephen D. Kachman
Larry A. Kuehn
Matthew L. Spangler

Summary with Implications

Economically relevant traits are those that directly impact commercial-level profit, and as such can only be measured at the commercial level. To capture and use these phenotypes in genetic evaluations, quantifiable relationships that connect routinely collected phenotypes from commercial animals to selection candidates in the seedstock sector are needed. Unfortunately, these relationships are largely unknown. Using pooled genotyping (pooling), relationships between commercial and seedstock animals can be established at a reduced cost. In return, the accuracy of expected progeny differences (EPD) of the seedstock selection candidates are increased and estimated breeding values (EBV) for the pools of commercial animals can be used for management. Seedstock animals with prior low accuracy, those that did not have progeny in genetic evaluations, benefit the most from this strategy. Generally speaking, a pool of any size is better than no information from commercial animals. However, some pool formations are better than others. Pooling in order to minimize phenotypic variation using pool sizes of 10 or greater in order to optimize EPD/EBV accuracy and cost is recommended.

Introduction

Although genetic change in economically relevant traits (ERT) that directly impact profit at the commercial level is the goal, genetic evaluations primarily utilize phenotypes collected within the seedstock sector of the beef industry. Thus, the EPD produced are for indicator traits. However, millions of ERT are collected annually within the commercial segments of the beef industry (feedlots, packing plants, commercial cow/calf herds). This information is rarely included in genetic evaluations due to the inability to connect commercial animals and seedstock selection candidates through known pedigrees. Relationships do exist between these groups of animals, but pedigree information is often unknown or incomplete. Relationships could be resurrected with genomics. However, it would require all commercial animals with records to be genotyped in order to estimate the relationships, which would be costly. An optimal solution would be to collect the ERT from commercial animals and estimate relationships between commercial animals and seedstock animals in an economical manner for use in genetic evaluations. Pooling data, genotypes and phenotypes, has been used to reduce the cost of genotyping while allowing for the inclusion of phenotypes that are typically only observed at the commercial level in genetic evaluations. Therefore, 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 seedstock and commercial animals on the resulting accuracy of EBV of parents and pools using simulation.

Procedure

A beef cattle population consisting of 15 generations (n=32,000) was simulated to have a phenotype with a heritability of 0.4, similar to most growth and carcass traits, and the markers mimicked those from a 50k single nucleotide polymorphism (SNP) panel. Individuals from generation 15 were considered commercial animals and included in pools. In practice, a pool represents a group of animals whose DNA has been equally combined and genotyped as a single sample and whose phenotype is the mean of the animals included in the pool. As simulated, the observed genotype and phenotype of the pools were mean values of the individuals that made up the group. Pool sizes included 2, 10, 20, 50, or 100 individuals, resulting in 1,000, 200, 100, 40, or 20 pools, respectively. Additional scenarios were included where individuals from generation 15 were individually genotyped and phenotyped and where the progeny information did not enter the evaluation at all (as if the commercial progeny did not have any information recorded). Pool assignments were determined in three ways: 1) randomly, 2) minimizing phenotypic variation within pools which led to individuals with similar phenotypes being grouped together, and 3) uniformly maximizing phenotypic variation within pools which led to the least variation across pools. Generational gaps in genotyping were induced by masking the genotypes of individuals born in generations 11 through 14 given, in practice, not all seedstock ancestors are genotyped. 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). Estimated breeding values were generated from a single-step genomic best linear unbiased prediction model. This model combines relationships derived from both genomics and traditional pedigrees into a single relationship matrix which allows for estimation of EBV in one step. The accuracy of the EBV of sires/dams born in generations 11, 12, 13 or 14 and the pools were assessed as the correlation of the EBV with true breeding values. As the accuracy becomes closer to 1, the EBV are better predictors of the true genetic merit of the animals/pools. The simulations were replicated 5 times; results were averaged over the 5 replicates.

Results

Figure 1 depicts the EBV accuracies of sires by generation of birth that resulted from different generational gaps in geno-

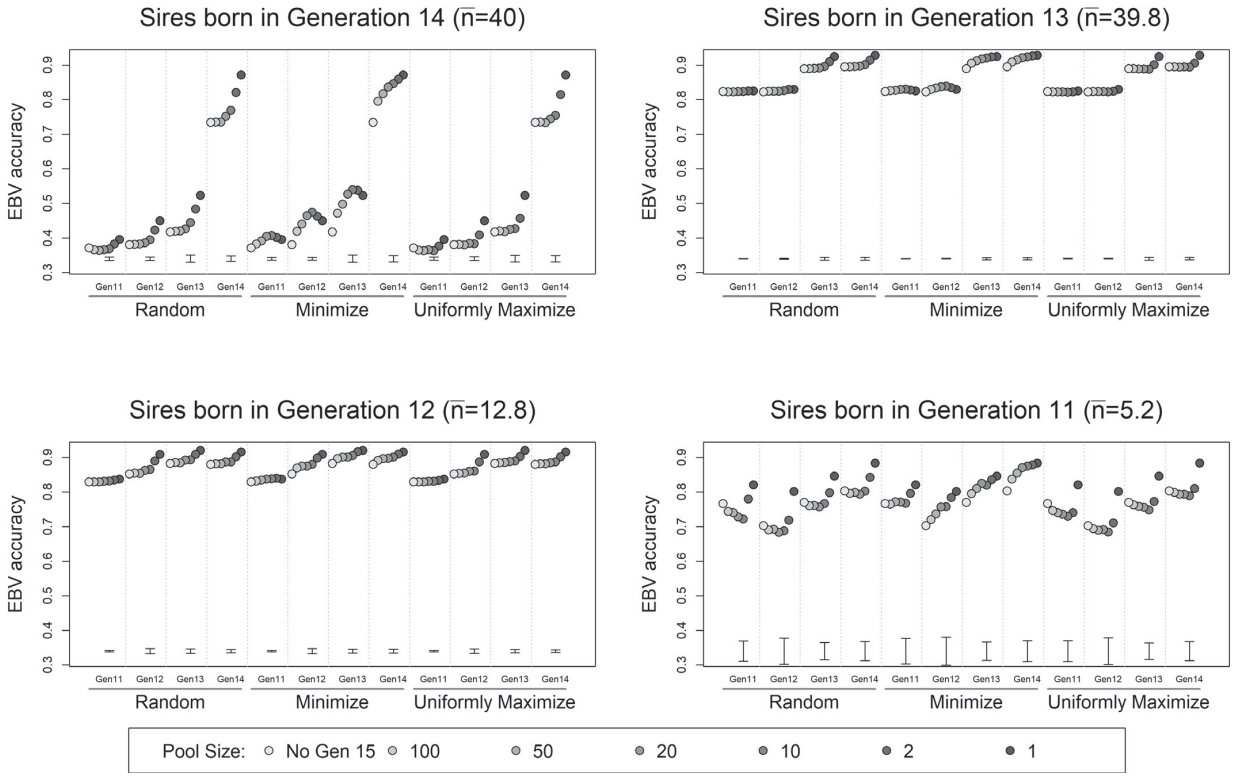

Figure 1. Estimated breeding value (EBV) accuracies of sires (estimated as the correlation between true breeding value and EBV) by generation of birth that resulted from different generational gaps in genotyping (Gen11 = individuals up to and including those born in generation 11 were genotyped; Gen12 = 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 (No Gen 15 = progeny records from generation 15 did not enter the evaluation) with error bars along x-axis

typing, pooling strategies, and pool sizes; accuracies of dams and grand dams/sires are not shown.

Pooling strategy

Random assignment and uniformly maximizing phenotypic variation within pools led to similar results. Minimizing phenotypic variation within pools led to larger EBV accuracies than the other two scenarios. The largest differences 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. Therefore, the ways in which pools are constructed does impact the accuracies of prediction.

Pooling size

Pool size also had a considerable impact on EBV accuracy. When pools were formed by allocating animals at random or by uniformly maximizing variation, EBV accuracy was reduced compared to having individual data with the exception of pool sizes of 2. Overall, even though there was a reduction in EBV accuracy resulting from pooling compared to individual data, the reduction was not statistically significant when pools were designed to minimize phenotypic variation.

EBV accuracy of pools

Including pools in the evaluation results in EBV for the pools themselves. The EBV accuracy of pools were significantly impacted by pool size and the interaction between pool size and pooling strategy. Accuracy of EBV of the pools decreased as pool size increased when pools were formed by randomly allocating animals or when animals were assigned to pools to uniformly maximize phenotypic variation. The opposite trend was observed when pools were formed by minimizing phenotypic variation, pool sizes of 100 led to the largest EBV accuracies. This result is because the average phenotype of the pools more closely reflected the average true breeding value of the pool as the pool size increased.

Generational gaps in genotyping

The EBV accuracies of sires and dams because of pooling were generally higher than if no data from generation 15 entered the evaluation. In other words, some

information from commercial progeny, even if the records are pooled, is better than no information from the commercial progeny. This was consistent whether the sires or dams in question were genotyped or were not. However, EBV accuracies for sires/dams were larger if the sires/dams in a particular generation were genotyped compared to if they were not genotyped. The largest increase in EBV accuracy resulting from the sire/dam being genotyped was observed with sires and dams born in generation 14. The increase in EBV accuracy from when sires were and were not genotyped was not as large for sires born in generations 11, 12 or 13 because EBV accuracy of those sires were already relatively high due to additional progeny that entered the evaluation individually. Dams, on the other hand, had larger 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. Thus, additional information had a large impact.

Conclusions

The accuracies presented from this simulation represent the theoretical maximum EBV accuracies; realized EBV accuracies resulting from pooling may be less due to lab and genotyping errors. However, the results presented herein show the potential use of pooling data at the commercial level for use in genetic evaluations in an economical manner.

Pooled phenotypes and genotypes can be a potential solution to economically include millions of commercial phenotypes that are currently not able to be used in genetic evaluations. Of the three pooling scenarios simulated, pooling in order to minimize phenotypic variation within pools, meaning to group phenotypically similar individuals together, led to the largest EBV accuracies of sires, dams, and of the pool themselves. When pools were constructed this way, pool sizes of 2, 10, 20, or 50 did not generally lead to differences in EBV compared to when progeny were individually genotyped and phenotyped. These EBV accuracies herein represent a theoretical maximum as in practice, it would likely not be possible to minimize phenotypic variation across contemporary groups. The EBV accuracies in practice will likely fall between those of random pooling and minimizing phenotypic variation. Sires with prior low EBV accuracy – those who do not have progeny that enter the genetic evaluation individually- benefit the most from pooling data in terms of increasing EBV accuracy. Overall, all seedstock animals benefit by utilizing commercial progeny with true ERT recorded. The EBV for the pools could be used to inform future management or marketing decisions.

Johnna L. Baller, graduate student, University of Nebraska–Lincoln

Stephen D. Kachman, Professor, Statistics, University of Nebraska–Lincoln

Larry A. Kuehn, Meat Animal Research Center, Clay Center, NE

Matthew L. Spangler, Professor, Animal Science, University of Nebraska–Lincoln