
BEEF IMPROVEMENT FEDERATION

# 10TH GENETIC PREDICTION WORKSHOP

## *Incorporation of Genomics into National Cattle Evaluation*

December 12-13, 2013
Kansas City, Missouri

# Genetic Prediction Workshop
# December 12-13, 2013

## *Incorporation of Genomics into National Cattle Evaluation*

**Conference Organizer:**
**Robert L. Weaber, Kansas State University**

**Conference Program/Proceedings Editor:**
**R. Mark Thallman, US Meat Animal Research Center**

**Technical Support:**
**Lois Schreiner, Kansas State University**

# Preface

The Beef Improvement Federations publishes Guidelines for use in genetic improvement of beef cattle. These Guidelines have provided for uniformity in methods of measuring traits, recording and analyzing data and estimating breeding value of animals which have provided the basis for significant genetic improvement in beef cattle in the U.S. and Canada for more than forty years. The procedures recommended, developed in committee meetings and workshops and approved by the Beef Improvement Federation Board of Directors, have provided procedures employed in beef cattle genetic improvement programs by member organizations of the Beef Improvement Federation in the United States and Canada. The procedures have also been adopted and used in beef cattle genetic improvement programs in many other countries around the world, especially Europe, Africa, and Australia. This, the 10th Genetic Prediction Workshop was organized and cosponsored by the Beef Improvement Federation and Regional Technical Committee NCERA-225 comprised of scientists at land grant Universities, the USDA, breed associations and other organizations that support and conduct beef cattle genetic evaluations in the U.S. and Canada. The primary purpose of this workshop is to share experiences and ideas regarding the incorporation of genomics into national cattle evaluation. Much progress has been made in the last few years, but there is much more to be learned and done. These proceedings may serve as background material for the next edition of the Guidelines for Uniform Beef Improvement Procedures to be published by the Beef Improvement Federation.

# TABLE OF CONTENTS

# BIF Genetic Prediction Workshop

**NCERA-225 (Regional Committee on Implementation and Strategies for National Beef Cattle Genetic Evaluation)**

*Holiday Inn KCI Airport and KCI Expo Center*
*Kansas City, Missouri*

## *Joint Meeting Schedule of Events*

**Wednesday, December 11, 2013**

7:00 p.m. NCERA-225 Business meeting-Truman Room-Hotel

**Thursday, December 12, 2013**

7:00 a.m. Breakfast Buffet for NCERA-225 Registrants - Pershing Room-Hotel

8:00 a.m. – 12:00 Noon NCERA-225 Station Reports - Salon EF-Expo Center

11 a.m. – 1:00 p.m. Registration for BIF Genetic Prediction Workshop - Foyer-Expo Center

12:00 Noon – 1:00 p.m. Lunch Buffet for BIF Registrants - Salon ABC-Expo Center

1:00 – 5:00 p.m. BIF Genetic Prediction Workshop - Ambassador-Expo Center

1:00 p.m. Introduction - *Dr. Mark Thallman, US Meat Animal Research Center*

1:15 p.m. Strengths and Weaknesses of Methods of Incorporating Genomics into Genetic Evaluation - *Dr. Matt Spangler, University of Nebraska-Lincoln*

2:00 p.m. AAA Experience with Incorporating Genomics into Genetic Evaluation – *Dr. Sally Northcutt, American Angus Association*

2:30 p.m. AHA Experience with Incorporating Genomics into Genetic Evaluation – *Mr. Jack Ward, American Hereford Association*

3:00 p.m. Break

3:30 p.m. ASA Experience with Incorporating Genomics into Genetic Evaluation – *Dr. Lauren Hyde, American Simmental Association*

4:00 p.m. Proposed Enhancements to the Across-Breed EPD System – *Dr. Larry Kuehn, US Meat Animal Research Center*

5:30 - 6:30 p.m. Social (Cash Bar) - Salon ABC-Expo Center

6:30 p.m. Dinner Buffet for BIF GPW Registrants - Salon ABC-Expo Center

## Friday, December 13, 2013

7:00 a.m. Breakfast Buffet for BIF Registrants – *Salon ABC-Expo Center*

8:00 a.m. – 12:00 -Noon BIF Genetic Prediction Workshop continues – Ambassador-Expo Center

8:00 a.m. Introduction for Second Day – *Dr. Bob Weaber, Kansas State University*

8:10 a.m. Experience in Developing Breed-specific Predictions for GE-EPDs – *Dr. Dorian Garrick, Iowa State University*

9:00 a.m. Individual Reliabilities of Molecular Breeding Values – *Dr. Steve Kachman, University of Nebraska Lincoln*

9:45 a.m. Break

10:15 a.m. Bayesian Regression as an Alternative Implementation of Genomic-Enhanced Genetic Evaluation – *Dr. Rohan Fernando, Iowa State University*

11:15 a.m. Conference Wrap-up and Discussion – *Dr. Mark Thallman, US Meat Animal Research Center*

12:00 Noon Adjourn

# STRENGTHS AND WEAKNESSES OF METHODS OF INCORPORATING GENOMICS INTO GENETIC EVALUATION

*Dr. Matt Spangler*

*University of Nebraska-Lincoln*

## Historical Summary

From an historical point of view, there have been considerable changes in the arena of beef cattle genomics. Changes include new genotyping platforms, the incorporation of genomic information into genetic selection decisions, the usefulness of it in predicting genetic merit, and our understanding of how best to utilize it. When genomic information was first integrated into National Cattle Evaluation (NCE) by the American Angus Association in 2009, the paradigm was completely different than it is today. At that time, the identification of animals in training populations was largely unknown, and thus the relationship between them and the target population was also unknown. Furthermore, only molecular scores were returned for use in NCE, the actual genotypes were held by the commercial genotyping company.

In short time, the global understanding of key issues began to penetrate our industry. Retraining, or recalibration, became a necessity and the beef industry understood that the efficacy of genomic predictors were not robust over several generations. The issue of robustness was also very clear across breeds, and the use of genomic predictors trained in Angus could not be used with any beneficial degree of accuracy in a closely related breed like Red Angus (Kachman et al., 2013). Consequently, for breeds to capitalize on the benefits of augmenting traditional EPD with genomic information, they must first make an initial investment in developing a training population. Generally speaking, breed associations were advised to genotype a minimum of 1,000 animals that were preferably moderate to high accuracy. The choice of animals in the initial training population was mostly ad-hoc.

Genomic assays, or SNP panels, also changed. The initial "backbone" of genomic prediction was the Ilumina BovineSNP50 (50K) assay. The Ilumina High-Density (HD) assay that included approximately 770,000 SNP was later released but did not penetrate the commercial market likely due to the increased cost and early research results that showed little predictive advantage of the HD assay over the 50K. More recently, an 80K product (GGP-HD; GeneSeek) and a reduced assay (GGP-LD; GeneSeek) have been released. The 80K product has taken the place of the 50K assay in some settings and the LD assay has allowed for imputation to denser content at a lower cost. Along with changes in panel density, there has been a considerable evolution in the entity that performs training. Initially, two primary companies performed this service and marketed the resulting genomic predictions: Pfizer Animal Genetics (now Zoetis) and Merial Igenity (now owned by GeneSeek). Other breed associations desired to own the intellectual property behind the genomic predictions they used and turned to the

National Beef Cattle Evaluation Consortium (NBCEC) lead by Dorian Garrick at Iowa State to perform the exercise of developing prediction equations.

**Inclusion of Genomic Predictors in National Cattle Evaluation**

As different breed associations began including genomic information into their NCE, the nuisances related to methodology for doing so increased. The method used by AAA was first proposed by Kachman (2008) and used by MacNeil et al. (2010) in their prototype evaluation. This became known as the "correlated trait approach" and assumed that the linear combination of SNP (Molecular Breeding Value; MBV) could be fitted as a correlated indicator trait in existing multiple-trait models. A primary benefit of this was the familiarity of the concept to breed associations. It also allowed for genomic information to influence the predictions of animals in the pedigree that were not genotyped. An initial pitfall was the model complexity associated with fitting multiple MBV for the same trait as a result of working with multiple commercial companies. As other breeds began to include genomic information into their NCE, "new" methods of doing so began to appear. It is important to note that the choice of inclusion method was arguably based on the genetic service provider (entity that conducted NCE) and not through model comparison. The majority of breeds that followed implemented a blending (indexing) approach whereby the MBV and EPD were indexed together to produce a genomically enhanced EPD. Initially this was done post evaluation and consequently only impacted the prediction of the genotyped animal. This created the largest difference between blending and the correlated trait approach. Currently, some breeds (e.g. Hereford) are moving towards blending such that ungenotyped animals will be influenced as well. Yet another variation on the theme was the American Simmental's approach of considering MBV as an external source of information much the same way they considered a non-Simmental parent EPD as an external piece of information in their Hybrid genetic evaluation. This again illustrated that the choice of inclusion method was conditional upon the differences in the genetic evaluation platform used by each breed. A primary benefit of this line of thinking was that it allowed for variable accuracy of MBV. It was intuitive to think that MBV did not predict the genetic merit of every animal in the population with the same degree of accuracy, mostly due to their relationship to the training population. This assumption is a flaw of the current implementation of the correlated trait approach and other blending methods. These incorporation methods could be altered to accommodate variable MBV accuracy, but advancements in both methodology and software are required. Initially ASA attempted to weight MBV proportional to a metric of reliability that was estimated from the posterior distribution of molecular scores via GenSel software. Unfortunately, this "reliability" was not the appropriate metric to use as it took on values that exceeded the bounds of reliability. Although a failed initial attempt to tackle issues that were known to exist, it arguably spurred the genomics community to start thinking of how to accommodate variable accuracy of genomic predictors.

All of these methods are essentially variations on the same two-step theme; train MBV and then fit them into NCE. The single step approach (Legarra et al., 2009) that is being utilized by other species has never been adopted by the beef industry. Initially this would not have been possible since the actual genotypes were not accessible to breed associations. All of the methods currently being used by beef breed associations have flaws. We know that bias exists in the genomic predictors. Now that the animals contained in training populations are known, these sources of bias can be estimated and accounted for.

To date several beef breed associations publish genomically enhanced EPD, and this list is likely to grow. Although the American Angus Association was the first to do this, organizations such as the American Hereford Association, American Simmental Association, Red Angus Association of America, American Gelbvieh Association, North American Limousin Foundation, and American Brahman Breeders (tenderness only) have at least initial prototype evaluations that include genomics. Even though these breeds have made tremendous progress, there is still a need to remind breeders that the fundamentals of genetic selection have not changed. Recording of phenotypes has always been critical and in an era of genomics it remains equally important. It is clear that retraining will be necessary overtime, and dense recording of phenotypes is needed to enable this. Although genomic predictors have enabled increased accuracy for young animals, phenotypes on progeny are needed to advance the accuracy of EPD past what is possible by genomics alone.

**Future Goals**

A needed, and logical, short term goal is to better understand the sources of bias that can arise due to the incorporation of genomic information into NCE. These can be caused by varying degrees of relationship between animals in the training set and those being predicted. Bias can also be created by using a selected subset of animals in training that are not representative of the general population. Admittedly, as the number of genotyped animals within a population increases, the impact of these concerns will begin to decrease. However, we are far from being at that point. These concerns might be even greater for "novel" traits, or those that are sparsely recorded. When phenotypes are expensive to collect, it is less likely that they will be chosen at random from a population. Although there are currently major efforts underway relative to developing genomic predictors for feed efficiency and susceptibility to Bovine Respiratory Disease (BRD), these efforts will serve as a starting point for breeds to work from. Ideally, some degree of strategy would be employed to build upon grant funded training sets to further the development of genomic predictors for these types of traits.

A longer-term goal centers on a greater understanding of the inter-workings of NCE. Issues related to the incorporation of genomic information into NCE and understanding and acceptance of the resulting enhanced EPD illustrate fundamental misunderstandings related to key concepts of genetic prediction. From a breed association perspective, genomics and the hurdles that have accompanied it might allow for (or force) dramatic changes in NCE platforms. In many cases, this would be a much needed overhaul from platforms that were built decades ago. From a breeder perspective, the true understanding of accuracy and possible change are needed to fully grasp the impact of genomics. Finally, the current status of genomics in beef cattle represents a strong partnership between breed associations and universities working together to enhance NCE. However, a succession plan must be in place such that at some point in the future breed associations develop the capability to move this process in house. This might, indirectly, force some degree of consolidation within the breed association community.

**Literature Cited**

Kachman, S. 2008. Incorporation of marker scores into national cattle evaluations. *Proc. 9th Genetic Prediction Workshop, Kansas City, MO,* pp. 88-91.

Kachman, S.D., M. L. Spangler, G. L. Bennett, K. J. Hanford, L. A. Kuehn, E. J. Pollak, W. M. Snelling, R. M. Thallman, M. Saatchi, and D. J. Garrick. 2013. Comparison of within and across breed trained molecular breeding values in seven breeds of beef cattle. *Genetics Sel. Evol.* 45:30.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656-4663.

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.

# AAA Experience with Incorporating Genomics into Genetic Evaluation

*Sally L. Northcutt*

*Angus Genetics Inc., American Angus Association*

## Introduction

Genomic-enhanced expected progeny differences (EPDs) are computed on a weekly basis at the American Angus Association® (AAA). Angus Genetics Inc.® (AGI) is a subsidiary of the Association that facilitates the implementation of DNA technology and the infrastructure for the weekly national cattle evaluation (NCE). The AGI personnel receive results from two companies providing genotyping services. Experiences relative to implementation of genomic technology and breeder awareness have had a significant impact on the Association outreach and business plan.

## The process

The DNA samples are received from Angus breeders and anonymously barcoded to associate the AAA animal identity before forwarding samples to company laboratories for extraction and genotyping. Companies return SNP genotypes in A/B format to AGI, and breeders are encouraged to anticipate a four-week turnaround for results being incorporated into the weekly EPDs. Any genotypes, as well as phenotypic trait data, received by AGI prior to close of business each Tuesday are included in the NCE weekly release on Friday mornings. Remaining DNA samples are returned to AGI by the genomics companies after genotyping and are archived at the Association headquarters.

The AGI geneticists in conjunction with Association information systems personnel oversee receipt and application of SNP genotypes into NCE procedures. Initially, the received SNPs are directed through parentage validation procedures conducted in-house to review any parentage conflicts for corrections to the Association datacase. Specific SNP effects are used in a 50K prediction algorithm to generate molecular breeding values for NCE traits. AGI and Zoetis scientists maintain a long-term collaboration to facilitate research, calibration efforts, and derivation of SNP effects for the purposes of genomic-enhanced EPDs. Genotypes are warehoused at the Association for reference and future calibration efforts.

## Genomic impact on the EPDs

In the AAA genetic evaluations, the molecular breeding values are incorporated into the EPDs using a correlated trait approach. Through AGI research and development, a genetic relationship is calculated between the values obtained from the genomic test results and the phenotypic data at AAA.

Figure 1 presents the genetic correlations between the phenotypic records and the 50K molecular breeding value predictions. The genetic correlations effectively range from .60 to .70, except for milk and heifer pregnancy. December 2013 marks the third calibration release for the 50K prediction since February 2011.

Figure 1. Genetic correlations between 50K predictions (n=38,981) and the American Angus Association phenotypic database (December 2013).

| Trait | Genetic Correlation (SE) | |
|---|---|---|
| Calving Ease Direct | .63 | (.09) |
| Birth Weight | .66 | (.02) |
| Weaning Weight | .56 | (.02) |
| Yearling Weight (gain) | .67 | (.03) |
| Milk | .38 | (.03) |
| Yearling Height | .73 | (.02) |
| Yearling Scrotal | .77 | (.01) |
| Dry Matter Intake | .73 | (.02) |
| Docility | .65 | (.02) |
| Heifer Pregnancy | .49 | (.07) |
| Mature Weight | .71 | (.02) |
| Carcass Weight | .58 | (.02) |
| Marbling | .66 | (.03) |
| Ribeye Area | .68 | (.03) |
| Fat | .64 | (.03) |

**Available traits that include genomic results**

Breeders and users of Angus genetics are strongly encouraged to use EPDs as the genetic improvement tool of choice, because EPDs account for all the available information on an animal, such as individual measures, progeny data, pedigree and genomic results.

Several considerations regarding genomic results merit special mention. The multi-trait genetic evaluation for mature weight and height includes the genomic prediction for only mature weight, because there is a high correlation between the molecular breeding values for mature weight and height. Likewise, the calving ease NCE that includes calving score and birth weight phenotypes incorporates genomic results for calving ease direct only. The residual average daily gain (RADG) values provided in the weekly genetic evaluation include the genomic indicator for dry matter intake (DMI) rather than the genomic test result for residual feed intake (RFI).

The number of genotypes for use in the 50K prediction is expected to exceed 50,000 registered Angus animals by early 2014. Results are incorporated into at least 15 EPDs which are then components of the selection index suite. Angus producers have appreciated the simplicity of this approach to incorporating genomics into the Association's NCE.

# Incorporating Genomics into Genetic Evaluation, Hereford

*Jack Ward*

*American Hereford Association*

## Introduction

Over the course of the past few years, DNA testing has been available to cattle breeders to test for parentage, genetic abnormalities and most recently, quantitative traits of economical relevance. The American Hereford Association (AHA) released is first Genomic Enhanced EPD (GE-EPD) in the fall of 2012.

## Current Status of GE-EPD in US Hereford Cattle

The AHA is using a Hereford Specific prediction equation to generate genomic predictions used to enhance the EPD and increase the corresponding accuracies for its breeders. It is anticipated this will reduce generation intervals and increase rates of genetic improvement within the US Hereford population. The initial training and validation population was developed by 50K genotyping nearly 1,100 high accuracy Hereford sires. A 4-fold cross validation method was used so that all animals contributed to both training and validation, but validation was on animals not closely related to those in training. In spring 2013, a new training was done on nearly 3000 high accuracy sires and a 6-fold method was used for training and validation. The new correlations nearly doubled and the average across traits was about .52, which means we are accounting for about 27% of the genetic variation in US Hereford cattle.

## Incorporating Genomics in Genetic Evaluation

Currently, the AHA utilizes a 77K GGP-HD panel at GeneSeek and then utilizes Beagle software to impute to 50K to calculate MBVs. These MBVs are electronically delivered to AHA where MBVs are generated for non genotyped relatives of genotyped animals using regression based on pedigree relationships. The resultant MBVs and their corresponding accuracies are then used for post evaluation blending to calculate an EPD including pedigree, performance and genomic components.

## Challenges associated with Genomic Incorporation

The AHA utilizes a full multi-trait evaluation and currently includes information from Canada, Uruguay and Argentina (Pan American Evaluation). This has various issues. First, the computing time and power to run full multi-trait with the addition of genomic information has cost

and time constraints. Second, the AHA prediction equation works well in the US Hereford population, but does not produce predictions with the same accuracy in the other countries.

**Expectations for the future**

In the near future, GGP-HD genotypes from all countries will be used for training and validation in order to try and develop one set of predictions for all Hereford cattle in Pan Am. The post blending method will be replaced by an approach that directly uses SNP data in the evaluation. It is hoped this will be run on a very frequent basis so that interim EPD calculation are no longer required. In addition, sequencing of highly influential Hereford sires is currently being undertaken and that information will be utilized to try and discover the causal mutations for those genomic regions with large effects. This will apply to all the routine EPD traits as well as Feed Intake, Fertility, Tenderness and health.

**Conclusions**

This technology is still in an infancy stage and all parts including training, validation and incorporation will need to continue to grow to service the beef industry appropriately. It is important to be an early adopter in order to capitalize as the technology grows. It is clear that breeders must continue to collect phenotypes as the predictions work very well in the populations that they are discovered and as each generation passes, those predictions become less effective.

# ASA EXPERIENCE WITH INCORPORATING GENOMICS INTO GENETIC EVALUATION

*Lauren R. Hyde[1,2], Wade R. Shafer[1], Stephen C. McGuire[1], Mahdi Saatchi[3], and Dorian J. Garrick[3]*

*[1]American Simmental Association, Bozeman, MT*
*[2]Red Angus Association of America, Denton, TX*
*[3]Iowa State University, Ames*

## Background

At the annual meeting of the American Simmental Association (**ASA**) in January 2011, the Board of Trustees voted unanimously to fund and initiate the development of genomically enhanced expected progeny differences (**GE-EPD**). The project required a large number of DNA samples representing heavily used Simmental-influenced bulls. The association already had a repository of DNA samples on many of these bulls from donations made by Simmental breeders over several years. Bull studs (ABS, Accelerated Genetics, Genex and Select Sires) and cooperators in ASA's Carcass Merit Program also made significant contributions to the DNA repository. The project to develop GE-EPD was the result of a multi-year collaborative effort between the ASA and several research entities, including the USDA, the University of Illinois, the University of Missouri, Montana State University, Iowa State University, the National Beef Cattle Evaluation Consortium (**NBCEC**), and GeneSeek.

## Development of ASA's 50K Training Panel

Genotyping of DNA samples was completed in the summer of 2011 either at the University of Missouri in Columbia or at GeneSeek in Lincoln, NE. Most of the 2,703 samples were genotyped with the BovineSNP50 BeadChip (Illumina, San Diego, CA), but 264 samples were genotyped with the BovineHD BeadChip (Illumina, San Diego, CA) (Saatchi et al., 2012).

Scientists at Iowa State University, led by Dorian Garrick, executive director of the NBCEC, computed deregressed EBV (**DEBV**) (Garrick et al., 2009) and corresponding weighting factors from EPD and accuracies supplied by ASA on the genotyped animals, their sires and their dams. Application of the BayesC method produced estimates of SNP marker effects, which were used to derive direct genomic breeding values (**DGV**). The researchers evaluated the accuracy of the DGV using K-means clustering and a 5-fold cross-validation strategy (Saatchi et al., 2012).

The DGV and DEBV from all 5 validation sets were then fitted together in a weighted bivariate animal model to estimate (co)variance components. The estimated covariance between

DGV and DEBV was used to estimate genetic correlations that represent the accuracy of genomic predictions for each trait, and these ranged from .29 to .65 across 12 different traits (Table 1; Saatchi et al., 2012).

As shown in Table 1, genetic correlations for the ASA 50K panel were comparable to those of commercially available panels developed for the American Angus Association (Shafer, 2012). The results indicated that, for many traits, incorporation of DNA test results into ASA's multibreed genetic evaluation system would add a significant amount of information, improving accuracy of predicted genetic merit on young animals, and in turn providing the opportunity to increase rate of genetic change within the population.

Table 1. Genetic correlations between DGV and traits developed for two breed associations by three different providers

| Trait | AAA[1] | | ASA[2] |
|---|---|---|---|
| | Igenity (Neogen) | Pfizer (Zoetis) | NBCEC |
| Direct calving ease | 0.47 | 0.33 | 0.45 |
| Birth weight | 0.57 | 0.51 | 0.65 |
| Weaning weight | 0.45 | 0.52 | 0.52 |
| Yearling weight | 0.34 | 0.64 | 0.45 |
| Milk | 0.24 | 0.32 | 0.34 |
| Maternal calving ease | na | na | 0.32 |
| Stayability | na | na | 0.58 |
| Carcass weight | 0.54 | 0.48 | 0.59 |
| Marbling | 0.65 | 0.57 | 0.63 |
| Ribeye area | 0.58 | 0.60 | 0.59 |
| Backfat | 0.50 | 0.56 | 0.29 |
| Shear force | na | na | 0.53 |

[1]American Angus Association.
[2]American Simmental Association (including both Simmental and SimAngus).

**Business Model**

Following the successful development of ASA's 50K test, we established a unique business model to serve the needs of our membership. Rather than relying on commercial entities to manage genotypes and provide molecular breeding values (**MBV**), ASA sends samples for genotyping directly to GeneSeek, a Neogen company that is collaborating with NBCEC. The resulting raw genotypes can then be directly accessed by ASA, along with the corresponding MBV that are computed at GeneSeek after GeneSeek imputes the genotypes to represent those used in the prediction equations previously developed from the ASA training population. The genotypes are routinely shared with NBCEC for ongoing research including derivation of improved prediction equations. This model gives the ASA considerable flexibility (Spangler,

2012), including the opportunity to compute MBV in-house or to use the genotypes directly in genetic evaluation.

## Incorporation of MBV into Genetic Evaluation

In 2012 we developed the infrastructure required to incorporate DNA test results into ASA's multibreed international cattle evaluation (**MB-ICE**) and herdbook interim systems. In the genetic evaluation system, MBV were treated like external EPD as described by Quaas and Zhang (2006). In the interim system, we applied Kachman's (2012) blending method with breed effect included in the base adjustment B. When ASA starts running more frequent evaluations, the herdbook interim system will not be necessary and will be eliminated.

### *External EPD Approach*

The main advantage of the external EPD approach is that individual accuracies of MBV are taken into account, resulting in more accurate GE-EPD. Accuracies of MBV are largely dependent on how well an animal is connected through pedigree ties to the population of animals used to develop the test. The MBV on animals with several close relatives in the training population will be more informative (i.e., more accurate) than those on animals with few or no close relatives. Failure to account for differences in MBV accuracy will not properly weight the information provided by the MBV, resulting in less accurate or biased GE-EPD.

We first incorporated DNA in our fall 2012 MB-ICE, in which we used the external EPD approach to compute GE-EPD for calving ease, weight traits and carcass traits using MBV with reliabilities less than or equal to 0.50 on 1,263 animals. After release, we noted that some of these animals had higher accuracy values than expected given the amount of information on them. We attributed this to the method we use for approximating prediction error variances (**PEV**).

### *Approximation of Prediction Error Variance*

Actual PEV can be computed from the inverse of the coefficient matrix of Henderson's mixed model equations. However, for most beef cattle evaluations, there are millions of equations so it is computationally impossible to invert the coefficient matrix. Methods to approximate PEV have been developed, but none are perfect.

In the MB-ICE, PEV are approximated using a Taylor series of expansion of inverse matrices (Wang et al., 1995). During expansion some off-diagonal elements are ignored, resulting in artificially high accuracies on some animals (R. L. Weaber, Kansas State Univ., Manhattan, personal communication). Although ASA has developed and implemented a

procedure that tempers inflated accuracies, it is computationally expensive and does not lend itself well to routine genetic evaluation. For this reason, we used only the blending method to compute GE-EPD and corresponding accuracies for the spring and fall 2013 MB-ICE.

*Scaling of DGV*

After the fall 2013 MB-ICE was released to the public, it became apparent that animals with better (worse) than average non-enhanced EPD had even better (worse) GE-EPD. In theory, however, DNA test results should have an equal chance of increasing or decreasing an animal's EPD, and any changes should be independent of the previous prediction of genetic merit.

Mathematically, this means that the correlation between the original EPD and the change in EPD (i.e., the difference between the original EPD and the GE-EPD) should be 0. In addition, from Reverter et al. (1994), the expected value of the regression of the GE-EPD (more accurate EPD) on the original EPD (less accurate EPD) should be 1.

From analysis of birth weight EPD, within-fold correlations between original EPD and changes in EPD for the 5 cross-validation groups and a group representing animals genotyped since development of the initial set of prediction equations were -0.21, -0.20, 0.15, 0.17, 0.19 and 0.24. Corresponding regressions were 0.92, 0.91, 1.02, 1.04, 1.06 and 1.05. Although the regressions were generally close to 1, the correlations were not close to 0. Analysis of other traits revealed similar relationships.

These results can reflect inappropriate scaling of the DGV, or double counting of information in the DGV and EPD. Multiplicative rescaling factors were derived within-fold for all traits. Rescaling of the DGV moved the correlations closer to 0 and the regressions closer to 1. The ASA's interim system was reprogrammed to accommodate the rescaling factors and the correct handling of breed effects, and the association re-released blended EPD on over 5,000 Simmental-influenced animals on November 1, 2013.

Although the ASA has had some negative experiences incorporating genomics to its MB-ICE, the association maintains that implementation of any new technology can lead to unforeseen difficulties and firmly believes in the potential of genomics to advance beef cattle breeding.

**Literature Cited**

ASA. 2012. 50K testing for SimGenetic DNA-enhanced EPDs. http://www.simmental.org/site/userimages/ASA_DNA_50K.pdf. (Accessed 12 November 2013.)

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55-62.

Kachman, S. D. 2012. Derivation of a blended BV (white paper). Univ. of Nebraska, Lincoln.

Quaas, R. L., and Z. Zhang. 2006. Multiple-breed genetic evaluation in the US beef cattle context: methodology. Proc. 8th World Congr. Genet. Appl. Livest. Prod., Belo Horizonte, Brazil, CD-ROM Comm. 24:12-18.

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994. Technical note: detection of bias in genetic predictions. J. Anim. Sci. 72:34-37.

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. 44:38-47.

Shafer, W. R. 2012. The future is here: ASA's new DNA test for EPD enhancement. The Register 26(2):16-17.

Spangler, M. L. 2012. Marker-assisted EPD for other breeds: a changing paradigm. The Register 26(3):16-17.

Wang, C. S., R. L. Quaas, and E. J. Pollak. 1995. Approximating prediction error variance using Taylor series expansion. J. Dairy Sci. 78(Suppl. 1):248. (Abstr.)

# PROPOSED ENHANCEMENTS TO THE ACROSS-BREED EPD SYSTEM

*Larry Kuehn[1] and Mark Thallman[1]*

*[1]Roman L. Hruska U.S. Meat Animal Research Center, USDA-ARS, Clay Center, NE 68933*

## Introduction

Breed diversity is one of the most important resources available to livestock producers. While discussion of breeds and mating plans typically involves taking advantage of heterosis or hybrid vigor, additive variation between breeds also provide a valuable tool for genetic change. Unfortunately, although multibreed cattle evaluations are performed for some breeds, predictions of genetic merit of animals across all breeds are not available. Across-breed EPD adjustment factors serve as a substitute to multibreed evaluation.

The U.S. Meat Animal Research Center (USMARC), with the cooperation of 18 breed associations, has been publishing yearly across-breed EPD adjustment factors for over 20 years. These factors (Table 1 from 2013 report) are meant to provide commercial producers with a tool to compare the genetic merit of sires across breeds for traits commonly reported in national cattle evaluations. In order to produce these factors, breed averages are calculated in the USMARC Germplasm Evaluation Program and adjusted for the differences of the sire EPDs used in the program relative to the current breed averages (Table 2 for 2013). The resulting breed averages are then adjusted to the current average EPDs within each breed to produce additive adjustment factors.

The factors have been widely used by commercial producers to select bulls from multiple breeds of sires. The yearly update to breed differences is useful for seedstock breeders to measure advantages of their respective breeds and to highlight potential for complementary mating. While these aspects of the program have been positive, it is useful to examine limitations of the program and to identify areas that would benefit most from modifications. This report reviews the basic methodology behind the yearly update, and puts forward some ideas for improvement of the program.

## Current Methodology

The basic methodology for derivation of the across-breed EPD adjustment factors were described by Notter and Cundiff (1991) and Núñez-Dominguez et al. (1993). This work demonstrated the importance of adjusting estimates of breed differences using the EPDs of the sires sampled. Additionally, they proposed the use of the regression of animal performance records on sire EPDs as a scaling factor to adjust for scaling differences between the environment at USMARC and the average environment of national cattle evaluation herds. Further modifications have been incrementally implemented, generally in reports of across-breed

EPD factor updates at the annual Beef Improvement Federation meetings (most recent update was Kuehn and Thallman, 2013).

A summary of the methodology, quoted directly from Kuehn and Thallman (2013), used in the most recent update to the across-breed EPD adjustment factors follows:

***Derivation of breed differences from USMARC***

An animal model with breed effects represented as genetic groups was fitted to the Germplasm Evaluation Program data set (Arnold et al., 1992; Westell et al., 1988). In the analysis, all AI sires (sires used via artificial insemination) were assigned a genetic group according to their breed of origin. Due to lack of pedigree, dams mated to the AI sires and natural service bulls mated to $F_1$ females were also assigned to separate genetic groups (e.g.., Hereford dams were assigned to different genetic groups than Hereford AI sires). In order to be considered in the analysis, sires had to have an EPD for the trait of interest. All AI sires were considered unrelated for the analysis in order to adjust resulting genetic group effects by the average EPD of the sires. Standard fixed effects (including heterosis) and random effects were assumed for each set of traits.

Differences between resulting genetic group solutions for AI sire breeds were divided by two to represent the USMARC breed of sire effects. Resulting breed differences were adjusted to current breed EPD levels by accounting for the average EPD of the AI sires of progeny/grandprogeny, etc. with records.

***Estimation of the regression of progeny performance on sire EPD***

For all traits, regression coefficients of progeny performance on EPD of sire for each trait were calculated using an animal model with EPD sires excluded from the pedigree. Genetic groups were assigned in place of sires in their progeny pedigree records. Each sire EPD was 'dropped' down the pedigree and reduced by ½ depending on the number of generations each calf was removed from an EPD sire. In addition to regression coefficients for the EPDs of AI sires, models included the same fixed effects described previously. Pooled regression coefficients, and regression coefficients by sire breed were obtained. These regression coefficients are monitored as accuracy checks and for possible genetic by environment interactions. The pooled regression coefficients were used as described in the next section to adjust for differences in management at USMARC as compared to seedstock production (e.g., YWT of males at USMARC are primarily on a slaughter steer basis, while in seedstock field data, they are primarily on a breeding bull basis). For carcass traits, MAR, REA, and FAT, regressions were considered too variable and too far removed from 1.00. Therefore, the regressions were assumed to be 1.00 until more data is added to reduce the impact of sampling errors on prediction of these regressions. However, the resulting regressions are still summarized.

Records from the USMARC Germplasm Evaluation Project are not used in calculation of within-breed EPDs by the breed associations. This is critical to maintain the integrity of the regression coefficient. If USMARC records were included in the EPD calculations, the regressions would be biased upward.

***Adjustment of USMARC Solutions***

The calculations of across-breed adjustment factors rely on breed solutions from analysis of records at USMARC and on averages of within-breed EPDs from the breed associations. The basic calculations for all traits are as follows:

USMARC breed of sire solution (1/2 breed solution) for breed i (USMARC (i)) converted to an industry scale (divided by b) and adjusted for genetic trend (as if breed average bulls born in the base year had been used rather than the bulls actually sampled):

$$M_i = \text{USMARC (i)}/b + [\text{EPD(i)}_{YY} - \text{EPD(i)}_{\text{USMARC}}].$$

Breed Table Factor ($A_i$) to add to the EPD for a bull of breed i:

$$A_i = (M_i - M_x) - (\text{EPD(i)}_{YY} - \text{EPD(x)}_{YY}).$$

where,

USMARC(i) is solution for effect of sire breed i from analysis of USMARC data,

$\text{EPD(i)}_{YY}$ is the average within-breed 2013 EPD for breed i for animals born in the base year (YY, which is two years before the update; e.g., YY = 2011 for the 2013 update),

$\text{EPD(i)}_{\text{USMARC}}$ is the weighted (by total relationship of descendants with records at USMARC) average of 2013 EPD of bulls of breed i having descendants with records at USMARC,

b is the pooled coefficient of regression of progeny performance at USMARC on EPD of sire (for 2013: 1.17, 0.84, 1.01, and 1.14 BWT, WWT, YWT, and MILK, respectively; 1.00 was applied to MAR, REA, and FAT data),

i denotes sire breed i, and

x denotes the base breed, which is Angus in this report.

Average AI sire EPD ($\text{EPD(i)}_{\text{USMARC}}$) was weighted on the sires' influence on the USMARC data. The weighting factor was the sum of relationship coefficients between an individual sire and all progeny with performance data for the trait of interest relative to all other sires in that breed.

**Evaluation of current methodology and release of across breed predictions**

We believe that the across-breed EPD system provides the most accurate estimates of current breed differences for growth and carcass traits available. However, we believe it is prudent to take a step back and consider how the system would look if it were being newly developed today with current technology and to compare this with how effectively the goals and needs are being met by the current methodology and release cycle (annually at Beef Improvement Federation meeting).

The main current objectives for the across-breed EPD program:

1) Provide a tool to compare genetic merit of animals across breeds for traits predicted in national cattle evaluations
2) Provide a tool to deliver current breed differences on economically important traits.

While one could argue that these objectives are being accomplished, there are other goals/needs that would improve their efficacy.

1) Be able to provide the objective above continuously (adaptable to base shifts and sire EPD changes) rather than once per year
2) Allow producers to obtain EPD predictions adjusted to any breed (rather than only factors on an Angus base)
3) Provide a delivery system that can convert EPDs across breed when adjustments are not additive
4) Improve analysis to take advantage of the evolving structure of the Germplasm Evaluation Program
5) Account for heterosis potential based on producer dam breed composition
6) If feasible, include non-USMARC data to improve the quality of breed comparisons and consistency with current multibreed systems.

These goals primarily fall into categories of improvements that can be made in data processing and improvements in the delivery of across-breed predictions.

**Data processing**

When the across-breed EPD system was implemented, the Germplasm Evaluation Program consisted of several 6-8 breed evaluations over periods of 2-4 years (cycles). Sires from all breeds were crossed to common base cow breeds such that breed differences could be estimated as breed-of-sire effects (1/2 of breed differences) under a simple sire model and assuming a constant level of heterosis between all breeds.

Currently, the Germplasm Evaluation Program is continuously evaluating the 18 breeds that register the most cattle and conduct national cattle evaluations. As part of this process, the

cow herd is being graded up to purebred (>7/8) cows of each breed. This eventual goal has multiple benefits including the ability to estimate breed specific heterosis, evaluation of crossbreds with relatively current germplasm in both the dams and the sires, and continuous sampling of relevant industry germplasm. This system requires a more complicated model than that used for the original cycle-based germplasm evaluation.

Much of this improvement has already occurred through the use of an animal model with genetic groups. However, there are still adjustments to be made. While 18 different breeds are being evaluated, an 18 x 18 diallele design cannot be conducted quickly. As a result, crosses among prominent beef breeds are the initial focus. As data for estimating breed-specific heterosis is, therefore, limited and incremental, a random effects model for breed specific heterosis is being evaluated. Under this model, estimates of breed specific heterosis will be shrunk toward overall heterosis until there is enough information to show otherwise. Additional analysis improvements under consideration include weighting of newer data relative to older data from the germplasm evaluation to decrease reliance on genetic trend, autocorrelations of base cow genetic groups, and improving calculation of average AI EPD of sampled sires to account for information content of phenotypic records.

Multibreed evaluation is being conducted in at least one national cattle evaluation system. In 2012, a goal of this group was to put several of the breeds on the same base; that is, to make the EPDs of animals from each breed directly comparable without adjustment factors. In terms of adjustment factors from the across-breed EPD system, we would then expect each of the involved breeds to have the same adjustment factor for each trait (within error associated with breed difference estimates and sire sampling). Based on the results shown in Kuehn and Thallman (2013), this goal was achieved for some traits, but possibly not for others. While there is no way to determine which system is 'correct', attempting to resolve differences between the USMARC predicted differences and those used in national cattle evaluation for multibreed evaluation seems like an important goal to avoid confusing cattle producers. A full resolution of this issue would likely require the incorporation of data sets obtained outside of USMARC into the across-breed EPD program. It is important that breeds are fairly and contemporarily compared from data generated by known sires in order to incorporate other sources of data.

While considering data from other sources (resource or seedstock databases), it is important to point out that data from one location in south central Nebraska is likely not representative of the entire beef cattle industry and is likely not sufficient for evaluating breed differences for all conditions in the United States. We strongly believe the results from other research herds and projects, particularly those conducted in the southeastern United States, would complement the Germplasm Evaluation Program dataset by strengthening the overall power to accurately determine current breed differences and by making possible the estimation of breed differences that vary by environments (G x E interactions).

As mentioned previously, one of the limitations to the current across-breed EPD program is that it is only released once per year, often after spring bull sales. As data from USMARC only arrives at defined intervals, the derivation of breed solutions from USMARC data is not a limiting factor. The most important changes to the adjustment factors come about due to changes to national cattle evaluation bases. These can quickly be accommodated by the breed associations involved providing new breed means and EPDs for the sires sampled in the Germplasm Evaluation Program. However, there still needs to be an acceptable method to release these new factors quickly in a form that producers can reliably reference. These issues can be addressed through new approaches to delivery of across-breed predictions.

**Delivery of across-breed predictions**

Given the limitation of yearly updates, it seems most logical to consider a web-based approach that allows factors or general predictions to be updated as soon as information changes. Resources needed to provide a web-based system are not available at USMARC. However, we have discussed the possibility a partnership with extension faculty involved in the National Beef Cattle Evaluation Consortium (M. L. Spangler, personal communication). Not only would this web-based system provide the industry with real-time updates, but it could help us to achieve most of the remaining unfulfilled goals mentioned earlier.

Right now the across-breed system provides factors that adjust within-breed EPDs to an Angus base. The development of a web-based system would eliminate the need for producers to use (or at least see) adjustment factors. Instead, producers could enter the breed and EPDs of individual bulls they are interested in evaluating relative to bulls from other breeds and have EPDs returned on the scale of any breed they choose. Spreadsheet and query-based methods to perform this type of calculations are already available through several sources (e.g., M. M. Rolf, personal communication; www.gps-beef.com).

Moving the adjustment factors to the background on a website is useful for multiple reasons. Occasionally producers confuse the factors with breed differences. Eliminating a direct table of additive factors would prevent this confusion. Additionally, keeping factors hidden makes it possible to accommodate non-additive adjustments to EPDs.

Non-additive adjustments are most important when EPDs are reported on different scales. A hypothetical simple example may be if one breed reported weights in pounds and another reported in kilograms. A simple additive adjustment would not be sufficient to make EPDs comparable between these two breeds. The EPDs of one breed would also have to be multiplied (e.g. converted from kilograms to pounds by multiplying by 2.2) to put them on the same scale. Other trait complexes could be subject to scaling problems such as converting EPDs developed on an ultrasound basis to a carcass basis. Calving ease EPDs also likely need to be scaled in order to equitably compare them across breeds.

Calving ease EPDs are relatively unique in national cattle evaluation in that they are generally reported on a unitless scale because calving ease is reported on a subjective scale such that animals are scored as 1 for no difficulty, 2 for minor difficulty, 3 for major difficulty, and 4 for Caesarian births (BIF Guidelines, 2010; available at www.beefimprovement.org). Depending on the evaluation system, these scores are analyzed in multiple ways including a traditional linear model where scores are treated as normal, continuous random variables, or as a threshold character using a probit threshold model. If a probit threshold model is used, the resulting EPDs are generally reported as probabilities relative to an arbitrary assumed mean value for the incidence of calving difficulty. These resulting EPDs have the highest variation if the assumed mean is 50% incidence of calving difficulty, which is the mean used by some breeds. Other breeds consider this incidence far too high and use a mean more reflective of actual incidence of calving difficulty in their databases such as 15-20%, which results in lower variation in the EPDs. In order to put the EPDs on the same scale, we must account for differences due to both the mean and the variance of the EPDs. Additive factors (mean adjustment) are not sufficient. Background adjustment in a web-based system can be programmed easily and would minimize producer confusion.

Decision support tools have been a goal of the National Beef Cattle Evaluation Consortium for several years. Prototype tools in which producers enter parameters that characterize their current cow herds have been developed. Across-breed predictions could easily be incorporated into such a system. Additionally, adjustment for differences in heterosis due to breed composition of the cow herd could be included. Breed specific estimates of heterosis from the USMARC Germplasm Evaluation could also be included as part of a decision support program.

As a final advantage of a web-based delivery system, breed associations could choose to allow producers to scan their EPDs directly from the across-breed website. If a filter is applied, these EPDs could be provided relative to any breed the user chose. While this sharing of data is not at all required, it could increase the marketability of sires through comparisons across multiple breed databases.

**Conclusions**

Overall, the across-breed EPD adjustment factors have been a valuable tool to the industry. Changes to data processing and reporting of these factors could increase their utility and ensure they are available on a timely basis such as when producers are buying bulls. We are exploring the possibility of developing a web-based reporting system for across-breed predictions that would no longer require a table of additive adjustment factors. We welcome any comments or suggestions as we explore these possibilities.

**TABLE 1: 2013 ADJUSTMENT FACTORS TO ADD TO EPDs OF EIGHTEEN DIFFERENT BREEDS TO ESTIMATE ACROSS-BREED EPDs**

| Breed | Birth Wt. | Weaning Wt. | Yearling Wt. | Maternal Milk | Marbling Score[a] | Ribeye Area | Fat Thickness |
|---|---|---|---|---|---|---|---|
| Angus | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.000 |
| Hereford | 2.7 | -3.5 | -23.6 | -17.1 | -0.32 | -0.09 | -0.050 |
| Red Angus | 3.4 | -23.2 | -27.9 | -3.9 | -0.30 | -0.08 | -0.029 |
| Shorthorn | 5.8 | 11.3 | 38.8 | 20.2 | -0.16 | 0.21 | -0.142 |
| South Devon | 3.2 | -4.8 | -6.6 | -0.3 | 0.08 | 0.16 | -0.111 |
| Beefmaster | 6.3 | 35.7 | 29.5 | 9.9 | | | |
| Brahman | 11.0 | 42.8 | 5.9 | 23.2 | | | |
| Brangus | 4.5 | 14.6 | 6.0 | 5.8 | | | |
| Santa Gertrudis | 6.6 | 36.2 | 48.3 | 12.4 | -0.66 | -0.05 | -0.116 |
| Braunvieh | 1.9 | -21.6 | -42.3 | 0.1 | -0.67 | 0.22 | -0.102 |
| Charolais | 8.6 | 38.1 | 45.3 | 6.9 | -0.44 | 1.02 | -0.220 |
| Chiangus | 2.2 | -20.5 | -40.2 | 4.7 | -0.45 | 0.45 | -0.157 |
| Gelbvieh | 2.7 | -18.2 | -25.6 | 3.6 | -0.41 | 0.78 | -0.136 |
| Limousin | 3.8 | -1.8 | -35.9 | -8.7 | -0.71 | 1.09 | |
| Maine-Anjou | 4.2 | -15.3 | -36.7 | -6.8 | -0.84 | 0.95 | -0.229 |
| Salers | 1.8 | -4.8 | -19.5 | 2.2 | -0.10 | 0.79 | -0.207 |
| Simmental | 3.7 | -5.9 | -10.9 | -0.8 | -0.42 | 0.53 | -0.141 |
| Tarentaise | 1.7 | 30.3 | 20.3 | 24.1 | | | |

[a]Marbling score units: 4.00 = $Sl^{00}$; 5.00 = $Sm^{00}$

**TABLE 2: BREED OF SIRE MEANS FOR 2011 BORN ANIMALS UNDER CONDITIONS SIMILAR TO USMARC**

| Breed | Birth Wt. | Weaning Wt. | Yearling Wt. | Maternal Milk | Marbling Score[a] | Ribeye Area | Fat Thickness |
|---|---|---|---|---|---|---|---|
| Angus | 87.3 | 577.0 | 1045.3 | 565.3 | 6.09 | 13.12 | 0.611 |
| Hereford | 91.7 | 571.5 | 1009.7 | 543.2 | 5.36 | 12.87 | 0.552 |
| Red Angus | 88.1 | 561.5 | 1013.0 | 558.3 | 5.71 | 12.77 | 0.570 |
| Shorthorn | 93.7 | 556.5 | 1022.9 | 564.8 | 5.45 | 12.98 | 0.448 |
| South Devon | 91.4 | 566.0 | 1030.0 | 564.9 | 6.11 | 13.07 | 0.500 |
| Beefmaster | 92.1 | 575.6 | 1002.9 | 554.2 | | | |
| Brahman | 98.3 | 587.7 | 989.3 | 571.9 | | | |
| Brangus | 90.8 | 568.2 | 1008.4 | 559.3 | | | |
| Santa Gertrudis | 92.6 | 570.5 | 1013.9 | 555.4 | 4.96 | 12.66 | 0.487 |
| Braunvieh | 89.9 | 549.4 | 981.8 | 576.4 | 5.46 | 13.63 | 0.432 |
| Charolais | 94.7 | 592.4 | 1047.7 | 556.1 | 5.22 | 13.92 | 0.381 |
| Chiangus | 90.9 | 546.2 | 987.0 | 557.9 | 5.37 | 13.24 | 0.449 |
| Gelbvieh | 89.6 | 575.4 | 1027.1 | 571.4 | 5.26 | 13.78 | 0.422 |
| Limousin | 90.8 | 574.7 | 1007.7 | 555.7 | 4.90 | 14.33 | |
| Maine-Anjou | 91.8 | 554.1 | 1000.8 | 555.2 | 4.99 | 13.80 | 0.372 |
| Salers | 89.0 | 566.4 | 1019.5 | 564.0 | 5.73 | 13.52 | 0.394 |
| Simmental | 91.5 | 586.1 | 1038.8 | 564.4 | 5.29 | 13.82 | 0.402 |
| Tarentaise | 89.1 | 576.2 | 1008.2 | 567.0 | | | |

[a]Marbling score units: 4.00 = $SI^{00}$; 5.00 = $Sm^{00}$

**Literature Cited**

Arnold, J. W., J. K. Bertrand, and L. L. Benyshek. 1992. Animal model for genetic evaluation of multibreed data. J. Anim. Sci. 70:3322-3332.

Kuehn L. A., and R. M. Thallman. 2013. Across-breed EPD tables for the year 2013 adjusted to breed differences for birth year of 2011. Proc. Beef Improvement Federation 45th Annual Research Symposium and Annual Meeting, Oklahoma City, OK. June 12-15, 2013. pp. 114-141.

Notter, D. R., and L. V. Cundiff. 1991. Across-breed expected progeny differences: Use of within-breed expected progeny differences to adjust breed evaluations for sire sampling and genetic trend. J. Anim. Sci. 69:4763-4776.

Núñez-Dominguez, R., L. D. Van Vleck, and L. V. Cundiff. 1993. Breed comparisons for growth traits adjusted for within-breed genetic trend using expected progeny differences. J. Anim. Sci. 71:1419-1428.

Westell, R. A., R. L. Quaas, and L. D. Van Vleck. 1988. Genetic groups in an animal model. J. Dairy Sci. 71:1310-1318.

# PRACTICAL EXPERIENCES IN DEVELOPING BREED-SPECIFIC PREDICTIONS FOR GENOME-ENHANCED EPDs

*Dorian J Garrick and Mahdi Saatchi*[12]

[1]*Department of Animal Science, Iowa State University, Ames IA*
[2]*National Beef Cattle Evaluation Consortium*

## Introduction

Genome-enhanced EPDs (GE-EPDs) combine all available information on pedigree, performance and genotypes. Eventually, such analysis should be able to be undertaken in a single analysis. However, the transition to that ideal situation from traditional national cattle evaluation (NCE) using only pedigree and performance records to generate EPDs is not straightforward. It involves political, economic and technological considerations. Genomic prediction is currently an immature technology, unlike traditional NCE which has been progressively fine tuned over several decades. Development of GE-EPDs therefore usually begins with prototype systems to produce direct genomic values (DGVs) for purposes of validation. Following suitable validation, the DGVs are typically combined with EPDs to produce GE-EPDs in a second blending step. This paper outlines some of the practical experiences in developing DGVs, and in their blending to produce GE-EPDs, for a number of beef breed associations.

## Developing the Training Population

The development of DGV requires a training population of historical animals that have both genomic data (typically high-density SNP markers) and information related to their true breeding values (BV). That information might consist of NCE EPD or individual phenotypes on the genotyped animals or their relatives. The size of the training population is expected to influence the accuracy of prediction, with larger training populations of many thousands or tens of thousands being most desirable (Goddard and Hayes, 2009). Furthermore, the relatedness of the training population to the candidates that will obtain DGVs is important (Habier et al., 2007). The DGVs in immediate relatives of animals in training will be more accurate than the DGVs in distant or unrelated individuals (Saatchi et al., 2011). Accordingly, the individuals to be genotyped should be close relatives, such as immediate sires and grandsires.

However, there are many practical limitations in the creation of suitable training populations. Generally speaking it is expected that for a given size of training population, the accuracy of DGV will be greater for higher heritability traits. Progeny test EPDs are a bit like observing phenotypes on a high heritability trait, making widely-used sires good candidates to include in training. Young sires that are soon to be widely-used are also good candidates, as these will likely be closely related to many candidates targeted for GE-EPDs. However, they

will usually suffer from having lower accuracy EPDs at the time of training. In order to obtain high accuracy for some traits such as stayability or maternal calving ease, training sires may be quite distantly related to animals in the current calf crops.

Experience with improvements in the accuracy of DGV over time as training populations increase in size would suggest that a minimum of 1,000 genotyped individuals in a single breed is a good target for preliminary analyses. Genotyping costs have decreased considerably and are now $45-$75 per animal, resulting in an investment requirement of at least $50,000-$100,000 in order to get started. A more desirable training population of a few thousand individuals would cost nearer a quarter million dollars in genotyping costs alone. Funding for the development of training populations controlled by various breed associations has largely come from their own funds. Collection of marker information on sires that already have reliable EPDs will not increase their reliability. Conversely, collection of marker information on young animals with little phenotypic data will not contribute much to improving the accuracy of the training process. Subsidized genotyping using breed association or research funds has proven to be useful in achieving target training population size.

**Challenges with SNP Markers**

Most SNP markers being routinely used are bi-allelic, resulting in three possible genotypes. Making use of these markers require that both the loci and marker genotypes be uniquely recognized. Unfortunately, there have been practical problems both in locus recognition and in genotype calling. Loci are recognized by alpha-numeric names, such as ARS-BFGL-BAC-10172. Unfortunately, these names have sometimes changed between different versions of the same SNP chips, and between different chip densities. Loci can also be identified based on their chromosome and base pair location, but those co-ordinates can change between version of the genome reference. Genotypes can be identified in a number of ways, including the forward allele, backward allele, top allele, A/B allele etc. It is important that the same genotype is called uniformly on different SNP chips. Our preference has been to use the Illumina A/B calling system for that purpose, resulting in genotypes AA, AB or BB, which can be subsequently coded for analysis as 0, 1 or 2 copies of the B allele. The nature and format of the genotype files also sometimes varies, but these can be recreated according to the format required. Some small fraction of markers will have missing genotypes on any particular analysis of an animal. Some small fraction of marker loci may disappear from one version of a chip to the next. Genotypes missing on many animals can be deleted, while those missing on a subset of animals can be imputed. The simplest imputation is to replace missing genotypes with the average genotype (i.e. number of copies of the B allele) for that breed. More complex imputation can be undertaken using pedigree or non-pedigree methods, both of which can be used to impute genotypes from one chip density to another, provide there is a sufficiently large reference population available at the higher density.

## Challenges with Animal Identifiers

Different breed associations use different methods to represent animals – some use numeric registration numbers, whereas others use alphanumeric registration numbers. Breeders may also refer to animals by ear tag numbers or by name. Some breed associations allow animals of other breeds to be used, and these may or may not be given a new animal identifier. All these issues cause problems with genotypes. First, the same animal may be genotyped more than once, with different identifiers used each time. Second, different animals can have the same registration number, most commonly when numeric registration numbers are used.

Accordingly, we have adopted a modified version of the Interbull standard for identifying animals. The Interbull standard uses a 19 digit alphanumeric identifier. The first three digits represent breed (e.g. AAN = Aberdeen Angus, RAN = Red Angus). The second three digits represent country of first registration (e.g. USA, CAN). The third digit represents sex (M, F). The remaining 12 digits represent the registration identifier within the breed association and country, left padded with 0's.

There are particular problems when animals are dual-registered with different numbers in the breed associations in different countries. Many widely-used sires have offspring in both Canada an the US, and may have been genotyped in both countries using different identifiers. Ideally, the Interbull identifier should recognize the country of birth/first registration.

Many samples are submitted for genotyping without registration numbers. Often tag numbers or other means of identification such as barcodes or semen straw codes are used. We routinely cross reference every animal identifier from the value used in the genotype file to the value representing the Interbull standard. We have had to modify the Interbull standard as it is designed for purebred recording, whereas many of our animals are composite animals. Accordingly, we use the breed code to represent the breed association rather than breed. Further, we often obtain genotypes on animals that are not registered with any breed association, for example animals fed at Tri-County Steer Futurity (TCSF). In that case, we use the principal breed and also modify the 12-digit identifier to include a recognizable acronym like TCSF.

## Storing Genotypes

The most common form of Illumina bead studio final report file consists of one row for each SNP marker-animal combination. One approach is to load these records into a database. However, that can result in a huge number of database entries and this leads to slow queries with many animals and markers. Instead, we convert the files to an alternative ascii format with one row for each animal and a column for each SNP. We use a different file for each version of the SNP chip. It is useful to link these files to a database to keep track of which animals have been genotyped, when, and on what platform. We filter animals and markers in analyses by deleting rows and/or columns of the genotype matrix.

## Deregression

The breed association training data typically consists of widely-used sires with accurate EPDs, at least for commonly-recorded traits such as birth, weaning and yearling weight. We deregress these data to form DEBV and generate corresponding weights for each animal, following the approach of Garrick et al., 2009. That deregression procedure removes the parent average effect to avoid double counting when half- or full-sibs are genotyped and included together in the training analysis. This also has the effect of removing the base or the breed effects in multibreed analyses. The DEBV have the same genetic variance regardless of the accuracy of the EPD, but the residual variance differs according to the accuracy. The deregression is undertaken in a preprocessing step, one genotyped animal at a time. It requires the EPDs and accuracies in trios representing the genotyped animal, its sire, and its dam. Deregression also requires knowledge of the heritability of the trait as used in the evaluation and this can sometimes be challenging to obtain.

There are two challenges in developing pipelines for these analyses. One is that the format of the data varies between breed associations and often also between extracts from the same breed association. Some formats involve one row for each genotyped animal including the trios of animal, sire and dam EPDs and accuracies, for all traits of interest. It is also helpful to have a row of column headings that are consistent descriptors of the fields between subsequent analyses. Other formats of these files involve a separate row for each genotyped animal, its sire, and its dam. The second challenge is that the nature of the data provided can vary from EPD to EBV, and the reported accuracy can vary from BIF accuracy, to accuracy as a correlation or reliability as the squared correlation.

## Cross-fold Validation

It is typically of interest to have some easily represented measure of the accuracy of genomic predictions. In practice, the accuracy of every animal's DGV could vary. The accuracy of close relatives of training animals will have more accurate prediction than distant relatives, and animals that are more inbred will have less accurate predictions than outbred animals that are more heterozygous. We compute an individual prediction error variance (PEV) corresponding to each DGV, but it is not easy to represent these as reliabilities or BIF accuracies unless the genetic variance is known for every animal. The use of predicted genetic variances can result in animals with apparent negative reliabilities. The computed PEV would be a good reflection of accuracy if the features used in genomic prediction represented haplotypes known to contain particular QTL alleles, or the causal alleles at the QTL rather than just markers in linkage disequilibrium (LD) with the QTL. One of the problems with using high-density SNP markers as features for genomic prediction is that the DGV will exhibit variation as a result of segregation in the SNP markers regardless of the accuracy of the corresponding prediction. Ideally, the variance of the DGV should be reduced as the accuracy is reduced, but this does not happen to the required extent with most currently used methods.

We prefer to use a conservative measure of accuracy of DGV. We would prefer to be told by industry that predictions are more accurate than we claimed, rather than less accurate. Accordingly, we validate our predictions in distant rather than close relatives. We assign animals to groups or folds based on K-means clustering using a difference matrix based on the matrix of additive genetic correlations obtained by standardizing the numerator relationship matrix using the inbreeding of each individual (Saatchi et al., 2011). The correlations in validation populations have large standard errors, so we use all the folds for validation. We fit a bivariate model that uses the DEBV as one trait and the DGV as the other trait. Initially we used the pedigree-based numerator relationship matrix to describe the variance-covariance matrix of DGV and DEBV. However, this leads to unusual results for the heritability of the DEBV (which should be the trait heritability) because it does not account for residual covariance between the animals with DGV in one fold and the animals in the other folds that were used in their training. Accordingly, we now zero out the off-diagonal blocks of the relationship matrix corresponding to animals in different folds (Saatchi et al., 2012). This leads to predictions of overall accuracy that are more like pooled predictions from each fold, but are obtained from setting to zero the first derivative of the pooled likelihood from each fold. The resultant variance components from this analysis should present a DGV heritability near 1, a DEBV heritability near the trait heritability, and a genetic correlation between DGV and DEBV whose square represents the average proportion of genetic variance accounted for by the DGV. This accuracy will under represent the accuracy of DGV for animals closely related to the training population and over represent the accuracy of DGV for animals more distantly related than between the folds in training. The routine predictions used by breed association come from a final analysis that combines all the information from all folds into a single Bayesian regression analysis.

**Analytical Methods**

There are many possible alternative models that can be used for genomic prediction. The ranking of alternatives across traits depends upon the genomic architecture of each trait. We prefer to use methods that can take advantage of knowledge of genes with large effect, and allows simultaneous genome-wide association studies (GWAS) to find such large effect regions. We also prefer to use so-called variable selection models that can provide information as to how informative any particular SNP is in the analysis. We prefer to use methods that are scalable to large datasets. Our preferences are for BayesC and BayesB. More details of these methods including R-scripts that implement these methods are in Fernando and Garrick, 2013; and Garrick and Fernando, 2013. The R-scripts are not recommended for large datasets, but are useful to gain insight into the nature of the calculations involved in the Markov-chain Monte-Carlo (MCMC) methods we use for Bayesian inference. Analysis of field data is done using a cpp version of the software with many modifications to improve efficiency. That version of the software is publicly accessible via the web (bigs.ansci.iastate.edu). The web interface sends jobs to a high-performance computer cluster (HPC) that runs all our analyses for breed associations. We store the individual MCMC samples of effects and use these later to obtain PEV and DGV

on individual or groups of animals that were not part of training. The DGV are the expected values of the breeding values given the data (i.e. E[u/y]) and the PEV are the variance of breeding values given the data (i.e. var[u/y]). These are genomic analogues to the solutions to mixed model equations (MME) in NCE which given certain assumptions are E[u/y], and the PEV obtained from approximations of the inverse MME coefficient matrix to compute BIF accuracy. The software is also installed on virtual machines in a pipeline at GeneSeek, and with some breed associations, so that our research activities to develop improved predictions are independent of routine breed association servicing of genomic predictions.

**Blending**

The DGV do not include information from any phenotypes collected since the development of the prediction equation. Accordingly, the NCE EPD can include information that will enhance the accuracy of the DGV. Further, the information on the DGV should provide information to improve the accuracy of close relatives, such as immediate offspring of the genotyped individual. For this reason, the DGV are often blended with NCE EPD to obtain GE-EPD. One approach to blending is to use the DGV as correlated traits, such as is done by American Angus Association. Another approach is to use the DGV as external EPD each with some corresponding accuracy. Finally, the DGV could be blended using selection index principles, given knowledge of the variance-covariance structure among the selection criteria (i.e. **P** matrix) and the covariance between the selection criteria and objective (i.e. **g** vector). Once **P** and **g** are defined, it is a straightforward matrix calculation to determine the weighting factors **b** by solving **Pb**=**g**. These weighting factors will be different for each trait, according to the squared correlation of the DGV accuracy ($r_g^2$), and for each animal according to its EPD/EBV reliability ($R_{EBV}^2$) which can be obtained from its BIF accuracy for each trait.

Recognize $var\begin{bmatrix} BV \\ EBV \\ DGV \end{bmatrix} = \begin{bmatrix} 1 & R_{EBV}^2 & r_g^2 \\ R_{EBV}^2 & R_{EBV}^2 & R_{EBV}^2 r_g^2 \\ r_g^2 & R_{EBV}^2 r_g^2 & r_g^2 \end{bmatrix} var[BV]$, then $\boldsymbol{Pb} = \boldsymbol{g}$ is

$$\begin{bmatrix} R_{EBV}^2 & R_{EBV}^2 r_g^2 \\ R_{EBV}^2 r_g^2 & r_g^2 \end{bmatrix} \begin{bmatrix} b_{EBV} \\ b_{DGV} \end{bmatrix} = \begin{bmatrix} R_{EBV}^2 \\ r_g^2 \end{bmatrix},$$

with solution $b_{EBV} = \frac{1-r_g^2}{1-R_{EBV}^2 r_g^2}$ and $b_{DGV} = \frac{1-R_{EBV}^2}{1-R_{EBV}^2 r_g^2}$ as in Kachman (unpublished); or AGBU Technical Update May 2011. These parameters are presented here in terms of EBV and DGV, but apply equally to EPD (i.e. ½EBV) and ½DGV.

Selection index assumes the data have expected value equal to 0, so this requires the two sources of information (i.e. EPD and ½DGV) to be deviated from their means. In a multibreed context, the EPD need to be deviated from their breed effects so that only the within breed deviation component of the EPD is weighted in the index. The breed effects (including base)

must then be added back to obtain a GE-EPD on the usual scale and base. More details are in Kachman (unpublished).

Corresponding blended accuracies can be obtained from selection index theory and are a function of the accuracies of DGV and EPD as

$R^2_{GE-EBV} = \frac{1-(1-r_g^2)(1-R^2_{EBV})}{1-R^2_{EBV}r_g^2}$ which can then be converted back to BIF accuracy.

Ideally, when a young sire is genotyped and its GE-EPD produced, the additional information which results in its GE-EPD being better or worse than its EPD should flow through to influence the EPD of its progeny. One approach to achieve this is to use the relationship matrix to derive a pedigree-imputed DGV for relatives of genotyped animals. This has been implemented in Breedplan and is used in the genomic predictions published by American Hereford Association. The analytical details of the pedigree imputation have not been published, but calculations can be done very efficiently taking into account the sparse nature of the inverse relationship matrix. The equations to be solved are $\boldsymbol{A^{11}u_1 = -A^{12}u_2}$, where the inverse numerator relationship matrix and vector of real and imputed DGV are partitioned according to the non-genotyped (1) and genotyped (2) animals.

Corresponding reliabilities can be obtained for these imputed DGV by recognizing the variance of imputed DGV is given by $\boldsymbol{A_{12}A_{22}^{-1}A_{21}}$, and from partitioned matrix theory $\boldsymbol{A^{11}} = (\boldsymbol{A_{11} - A_{12}A_{22}^{-1}A_{21}})^{\boldsymbol{-1}}$, which leads to a convenient computing formula for reliability $R^2_{DGV}$ as the ratio $\{diag[\boldsymbol{A_{11}}] - diag[(\boldsymbol{A^{11}})^{\boldsymbol{-1}}]\}/diag[\boldsymbol{A_{11}}]$. For noninbred individuals this is simply 1-$diag[(\boldsymbol{A^{11}})^{\boldsymbol{-1}}]$. The imputed DGV reliabilities can then be backtransformed to BIF accuracies. The blending selection index weights can then be obtained by replacing $r_g^2$ in **P** and **g** from the selection index equations above and in $R^2_{GE-EBV}$ with the product $R^2_{DGV}r_g^2$ which will differ for every non genotyped animal according to the reliability of its imputed DGV.

Collectively this allows blending of imputed DGV and EPD to generate GE-EPD on non-genotyped relatives of genotyped animals. The blending process will put less emphasis on the imputed DGV than a DGV for a genotyped animal, that emphasis being dictated by the accuracy of the imputed DGV. Similarly, the accuracy of the GE-EPD will not increase for these relatives as much as it would for an animal with an actual high-density genotype.

**Diagnostics**

There are useful diagnostics to confirm that genetic predictions are behaving as expected. Consider a new more accurate EPD (NEPD) that is obtained from analysis that includes additional information collected since the calculation of a previous EPD (PEPD). It is expected that the regression of the more accurate on the less accurate (i.e. NEPD on PEPD) should be 1, as was the basis of so-called Method R (Reverter et al., 1994). Another diagnostic is that the correlation between the change in prediction (i.e. NEPD-PEPD) with the previous prediction

should be 0. This reflects that EPDs should be equally likely to go up or go down when additional information is collected to improve their accuracy. These diagnostics are useful in comparing results between different NCE, including comparing individual EPDs to parent average (PA) EPDs. We routinely use these diagnostics to check for problems with DGV, EPD and GE-EPD. One of the most common problems is that the variance of DGV exceeds the genetic variance times the square of the genetic correlation. This results in the regression of performance on the DGV varying from 1, indicating that there is some bias. If the regression exceeds 1, the DGV are shrunk too much, if it is less than 1, the DGV are not shrunk enough. Although both of these forms of bias occur in practice, it is more common for DGV variance to be too large and the regression less than 1, typically around 0.8. We scale the DGV slightly in these cases.

**Lower Density and Mixed Density Panels**

The original SNP panel that was most widely used in the cattle industry was the Illumina 50K panel. Due to inadequacies in predicting across breeds, a denser Illumina 770K panel was released, but that has not been used much outside research projects. A cheaper Illumina 3K panel was also introduced and then progressive versions added extra content. Illumina also offered the development of custom panels. GeneSeek created a number of custom panels that have included parentage markers, genetic defects tests, patented and other proprietary content. The most common versions of these panels currently being used are known as the GGP-LD (about 20K) and the GGP-HD (about 70K). There is also a custom HD panel for *Bos indicus* cattle. The GGP-LD content is almost entirely contained on the GGP-HD, and other than the custom content is a subset of the Illumina 770K panel. The GGP-LD and GGP-HD include about 9K and 28K content common to the Illumina 50K panel. The GeneSeek products are cheaper and more practical than the Illumina 50K panel, and we now use these almost exclusively. Collectively, this variety of chip densities requires us to impute virtually all genotypes to the 50K panel, as our current prediction equations are principally focused on the 50K content. We no longer make routine use of unmapped 50K content.

**GeneSeek Pipeline**

It is important that breeders get timely access to GE-EPD. This requires rapid genotyping of small numbers of samples, imputation of the genotypes to 50K features used in prediction, calculation of DGV, blending of DGV with EPD to generate GE-EPD, and communication to the breeders. We have developed a prediction pipeline that uses Beagle to impute GGP-LD or GGP-HD to 50K markers used in prediction, followed immediately by creation of DGV. The imputation is undertaken one chromosome at a time, in parallel, using previously constructed DAG files generated from the 50K populations used for initial training. The raw genotypes and DGV are then immediately accessible to the breed associations for blending and publishing.

**Pooling Breeds and Breed Associations**

Many breed associations record cattle of other breeds, particularly composites with Angus. For example, American Gelbvieh Association records Balancer cattle that are crosses with Angus or Red Angus, and North American Limousin Foundation records LimFlex cattle that are crosses with Angus. American Simmental Association and Red Angus Association of America record a wide range of breed crosses. For each of these associations, it is desirable that a single prediction equation be used across their pure and composite cattle. This is important because the breed is often not known at the time of sample submission – some samples are submitted from animals that do not have registration numbers. This precludes the ready use of breed-specific imputation techniques, or breed-specific prediction equations.

Fortunately, we have demonstrated that, in most cases, the training data from different breeds can be pooled, with the resultant training population producing cross-validation DGV accuracies that are similar to those obtained from within-breed training for any of the breeds being pooled (Saatchi and Garrick, 2013ab). Further, GeneSeek has made their 50K genotyped populations available to the National Beef Cattle Evaluation Consortium (NBCEC) to allow their use in training or validation to the benefit of their breed association clients. In particular, this allows Angus animals with 50K genotypes to be used in training of any other breed that has those animals represented in their NCE. That is, we do not pool the DEBV from different NCE, but use the NCE EPD for that animal in the target breed association. Training is therefore not breed specific, but breed association specific.

One example where pooling has reduced the accuracy of prediction is for some carcass traits when Limousin and Angus animal are pooled. We believe this results from the fact that Limousin segregates the F94L mutation that is fixed in British breeds. We are now genotyping and imputing that mutation to be used along with 50K markers in prediction.

**Future Developments**

We are actively researching many alternative approaches aimed at improving predictive accuracy and reducing computational burden associated with the generation of GE-EPD. These include the use of haplotype-based methods in prediction, and the fitting of QTL effects rather than SNP or haplotype effects in our models.

Those breed associations using selection index blending methods have the additional problem of interim calculations in order to provide breeders timely access to GE-EPD in between routine NCE. A sensible approach is to replace interim evaluations with more frequent, even continuous NCE. Increasing computing power is making it much easier to run NCE than was the case in the past.

Existing NCE all produce EPD by iterative solving of MME. The BIF accuracies are more problematic to obtain, and are approximated. The approximations provide reasonable

assessments of EPD accuracy in some cases, but much less so when genomic information is included. Alternative approaches to obtaining EPD and their accuracy involve sampling from the conditional distributions of the breeding values given the phenotypic, pedigree and genomic data, rather than simply solving MME. This provides much richer inference, and provides accuracies that are not approximations. Further, it allows accuracies to be readily obtained on comparisons between animals, or between groups of animals, that were not able to be estimated using the methods adopted to approximate the inverse elements of MME. We expect these methods to be extended to routine NCE.

A by-product of the genomic training analyses is that genomic regions accounting for the most variance in any particular trait are readily identified. Many of these regions appear common to more than one trait, and many are common to more than one breed. It appears that just 3 or 4 regions account for a significant amount of variation in some traits, particularly growth and calving ease. We are investigating methods to improve the prediction of effects in these regions, and in predicting these QTL as fixed rather than random effects. In future we expect to identify better markers or perhaps even causal mutations in these regions, and to genotype or impute these for use in prediction. Thus the features used in prediction are expected to migrate away from the 50K features used at present.

Whole-genome sequencing is becoming much cheaper and is now being undertaken on many individual sires. This data is expected to help identify improved features for genomic prediction. Further, it allows loss-of-function mutations to be identified, and these may be useful for prediction in a genome-driven rather than performance data-driven analysis. It may allow predictions of GE-EPD for embryo mortality or other factors based only on knowledge of loss-of-function mutations.

**Summary**

Genomic prediction is an immature technology, whose implementation has involved political, economic and technological considerations. Nevertheless, it has now been commercialized as a legitimate approach to improving the accuracy of EPD. That is now resulting in rapid growth in the number of genotyped individuals. The next growth phase will represent the genotyping of entire cohorts of calves but this will likely require slightly cheaper genotyping panels and improved accuracies of genomic prediction. These activities will require ongoing research effort.

**Acknowledgements**

We acknowledge the assistance provided by Neogen/GeneSeek in sharing Igenity genotypes and providing an umbrella agreement to NBCEC for breed association genotyping. We also acknowledge NBCEC affiliated researchers for sharing genotypes, and the participating breed associations for sharing their genotypes, and providing pedigree and EPD information. This project was supported by the USDA Cooperative State Research, Education and Extension

Service and National Research Initiative grants number 2009-35205-05100, and 2012-67015-19420 from the USDA National Institute of Food and Agriculture.

**Literature Cited**

Fernando RL, Garrick DJ. 2013. Bayesian methods applied to GWAS**.** In *Genome-Wide Association Studies and Genomic Predictions*. *Edited by Gondro C, van der Welf J, Hayes B. Springer: Humana Press,* 275-298.

Garrick DJ, Fernando RL. 2013. Implementing a QTL detection study (GWAS) using genomic prediction methodology**.** In *Genome-Wide Association Studies and Genomic Predictions*. *Edited by Gondro C, van der Welf J, Hayes B. Springer: Humana Press,* 275-298.

Garrick DJ, Taylor JF, Fernando RL. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetic Selection Evolution,* 41:55.

Goddard ME, Hayes BJ. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Review Genetics*, 10:381-391.

Habier D, Fernando RL, Dekkers JCM. 2007. The impact of genetic relationship information of genome-assisted breeding values. *Genetics*, 177:2389-2397.

Reverter A, Golden BL, Bourdon RM, Brinks JS. 1994. Method R variance components procedure: application on the simple breeding value model. *Journal of Animal Science* 72:2247-53.

Saatchi M, and Garrick DJ. 2013a. Genomic Prediction in Red Angus beef cattle is improved by using a multi-breed reference population. *Journal of Animal Science, 91(E-Suppl. 2) /Journal of Dairy Science, 96(E-Suppl. 1)*:22.

Saatchi M. and Garrick DJ. 2013b. Improving genomic prediction in Simmental beef cattle using a multi-breed reference population. In *Proceeding of the Western Section American Society of Animal Science*, 64:174-176.

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, Decker JE, Taxis TM, Chapple RH, Ramey HR, Northcutt SL, Bauck S, Woodward B, Dekkers JCM, Fernando RL, Schnabel RD, Garrick DJ, Taylor JF. 2011 Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetic Selection Evolution*, 43:40.

Saatchi M, Schnabel RD, Rolf MM, Taylor JF, Garrick DJ. 2012. Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. *Genetic Selection Evolution,* 44:38.

# Individual Reliabilities of Molecular Breeding Values

*Stephen D. Kachman*[1]

[1]*Department of Statistics, University of Nebraska–Lincoln*

## Introduction

Incorporation of genomic information through the use of molecular breeding values (MBV) into national beef cattle evaluations has now become widespread (Spangler 2013). A key component to giving the appropriate weight to genomic information is the reliability of the molecular breeding value. The reliability is the squared correlation between the individual's MBV and their true breeding value (TBV). Due to factors such as recombination, the reliability of a MBV decreases as the genetic distance between the animals used in training and the animals being evaluated increases (e.g. Habier et al. 2010).

## Role of reliability

The reliability of a MBV in genomic evaluations impacts both an animal's EBV and the reliability of the EBV after incorporating MBV information into the EBV. Ignoring base adjustments and scaling factors, and assuming conditional on the TBV that the phenotypic EBV and the MBV are uncorrelated, the EBV incorporating both phenotypic information and the MBV is

$$EBV + \frac{1-R_{EBV}^2}{1-R_{MBV}^2 R_{EBV}^2}(MBV - R_{MBV}^2 EBV),$$

where $EBV$ is the EBV based on phenotypic information, $MBV$ is the scaled MBV, $R_{EBV}^2$ is the reliability of the phenotypic EBV and $R_{MBV}^2$ is the reliability of the MBV . The reliability of the EBV incorporating both phenotypic information and the MBV is

$$R_{EBV}^2 + R_{MBV}^2(1-R_{EBV}^2)\frac{1-R_{EBV}^2}{1-R_{MBV}^2 R_{EBV}^2}.$$

From the above equations it can be seen that the greater the reliability of the MBV is the greater weight is given to the MBV and greater the resulting increase in the reliability.

## Statistical model

A simplified model will be used to obtain an approximate parameterization of individual animal reliabilities. Let $M$ be a $n \times p$ matrix of $p$ marker genotypes on $n$ animals and $b \sim (0, I\sigma_b^2)$ be a $p \times 1$ vector of marker effects. The $n \times 1$ vector of breeding values is defined to be $u = Mb$. We will further assume that the first two moments of $M$ are

$$Vec(M) \sim (0, L \otimes A)$$

where $L$ is $m \times m$ matrix capturing the covariance structure between marker covariates and $A$ is the numerator relationship matrix. Under this model, a genomic relationship matrix

$$H = MM'/tr(L)$$

is an unbiased estimator of $A$. The covariance matrix of breeding values conditional on the marker covariates is therefore $Htr(L)\sigma_b^2 = H\sigma_u^2$ where $\sigma_u^2$ is defined to be $tr(L)\sigma_b^2$ and the unconditional covariance matrix of the breeding values is $A\sigma_u^2$.

The model for the phenotypes is

$$y = X\beta + Zu + e$$

where $y$ is the vector of observed phenotypes, $X$ and $Z$ are the design matrices for the fixed and random effects, and $e \sim (0, R)$ is the vector of residuals.

Let $M_1$ be the marker genotypes of the animals used in training and $\hat{u}_1$ be the EBV of the animals used in training. Under this model an approximate parametric form of individual animal reliabilities was obtained.

**Results**

An individual animal reliability is the squared correlation between the animal's MBV and its true breeding value. The reliability is thus a function of the variance of the TBV ($\sigma_u^2 a_{ii}$), the variance of the MBV ($Var(MBV_i)$), and the covariance between the MBV and the TBV ($Cov(MBV_i, u_i)$).

An approximate variance of the MBV for animal *i* is

$$Var(MBV_i) \approx \frac{tr(L)^2 + tr(L^2)}{tr(L)^2} a_{1i}' A_{11}^{-1} Var(\hat{u}_1) A_{11}^{-1} a_{1i} + \frac{tr(L^2) tr(A_{11}^{-1} Var(\hat{u}_1))}{tr(L)^2} a_{ii}$$

where $A_{11}$ is the numerator relationship matrix of genotyped animals used in training and $a_{1i}$ is the vector of relationships between animals in training and the animal being evaluated. The approximate variance indicates that $a_{1i}' A_{11}^{-1} Var(\hat{u}_1) A_{11}^{-1} a_{1i}$ provides one measure of similarity between the animal being evaluated and the information used in training with the variance of the MBV decreasing to a background level as the similarity goes to zero.

An approximate covariance between the MBV and the BV of animal $i$ is

$$Cov(MBV_i, u_i) \approx \frac{tr(L)^2 + tr(L^2)}{tr(L)^2} a_{1i}' A_{11}^{-1} Cov(\hat{u}_1, \hat{u}_i) + \frac{tr(L^2) tr(A_{11}^{-1} Var(\hat{u}_1))}{tr(L)^2} a_{ii}.$$

Both the variance and the covariance approximations depend on two unknown parameters

$$\alpha_s = \frac{tr(L)^2 + tr(L^2)}{tr(L)^2}$$

and

$$\alpha_B = \frac{tr(L^2) tr(A_{11}^{-1} Var(\hat{u}_1))}{tr(L)^2}.$$

The variance approximation can now be written as

$$Var(MBV_i) \approx \alpha_s a_{1i}' A_{11}^{-1} Var(\hat{u}_1) A_{11}^{-1} a_{1i} + \alpha_B a_{ii}.$$

The covariance approximation can now be written as

$$Cov(MBV_i, u_i) \approx \alpha_s a_{1i}' A_{11}^{-1} Cov(\hat{u}_1, \hat{u}_i) + \alpha_B a_{ii}.$$

The form of the individual animal reliability is

$$R_{MBV}^2 = \frac{Cov(MBV_i, u_i)}{\sqrt{\sigma_u^2 a_{ii} Var(MBV_i)}}.$$

**Conclusions**

An approximate parametric form for individual animal reliabilities was obtained. It is equivalent to the usual form if $\alpha_s = 0$ and $\alpha_B = 1$. The approximate parametric form also provides a similarity measure for each animal being evaluated from the information used in training based on each animal's relationship to the animals in training and the phenotypic reliabilities of the individual animals. In obtaining the approximation, a number of simplifying assumptions were made which could have a direct impact on how well the approximation works in practice.

**Literature Cited**

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics, Selection, Evolution*: 42:5.

Spangler, M. 2013. Strengths and Weaknesses of Methods of Incorporating Genomics into Genetic Evaluation. *Proceedings BIF Genetic Prediction Workshop, Kansas City*.

# Bayesian Regression as an Alternative Implementation of Genomic-Enhanced Genetic Evaluation

*Dr. Rohan Fernando[1] and Dorian Garrick[1,2]*

*[1]Department of Animal Science, Iowa State University, USA; [2]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand*

## Introduction

In livestock, phenotypic and pedigree data are available on a large subset of the population, and genomic data are available only on a smaller subset. Genetic evaluation by combining all the available phenotype, pedigree and genomic data from the entire population results in genomic-enhanced genetic evaluations (GEBV).

Two approaches have been used to combine information from animals with and with- out genomic data for GEBV. The first is a multi-step approach. In the first step, high- density SNP genotypes are used to estimate the substitution effects at many thousands of loci in a reference population where genotypes and phenotypes are available on a few thousand animals. Typically, the number of substitution effects to be estimated has been much larger than the number of animals with data. Thus to estimate the substitution effects, Bayesian regression methods are used that combine prior information on the substitution effects with the information that is contained in the data [8, 3]. Once the substitution effects are estimated, the breeding value of a candidate with SNP genotypes can be predicted by using the estimated substitution effects in the regression model (step two). These predictions are called direct genomic values (DGV). In step three, these DGV are combined, using selection index theory, with EBV of the candidates obtained using the pedigree and phenotypic data from the entire population [11]. This approach is complicated by the fact that in many situations the animals with genotypes in the reference population do not have phenotypes. Deregressed EBV of these animals are therefore used to estimate the substitution effects [2]. This deregression requires the reliability of the EBV, which are often approximated. Further, as part of the deregression process parent information is often removed. This process will also result in approximations when data used to compute the EBV are not available for deregression. The selection index that is used to combine the DGV and the EBV needs to take into account that deregressed EBV were used in computing the DGV, involving further approximations.

In the second approach, BLUP GEBV are obtained in a single step, combining all phenotypic, pedigree and SNP data using Henderson's mixed model equations (MME) with a modified version of the additive relationship matrix **H** that reflects the additional information from the SNP genotypes [7, 1]. In this method, usually referred to as single- step BLUP (SS-BLUP), the SNP genotypes are used to construct a genomic relationship matrix (**G**) for the

animals that are genotyped, and the remaining relationships that are based on pedigree are modified to be consistent with **G**. The matrix **H** of relationships that are based on both pedigree and SNP information can be inverted efficiently, provided the number of genotyped animals is not too big to get a direct inverse of G and of the pedigree relationship sub matrix for the genotyped animals. Thus, SS-BLUP is an attractive method for GEBV that does not involve the many approximations that are encountered in the first approach. One of the disadvantages of SS-BLUP, however, is that it can quickly become computationally infeasible as the number of genotyped animals grows. The reason for this is that the computational burden for direct inversion of a matrix is proportional to $n^3$, where *n* is the order of the matrix. So, as the number of genotyped animals grows, the computational burden for SS-BLUP grows exponentially. Another disadvantage of SS-BLUP is that it is limited to assuming the marker effects follow a normal distribution. In some situations, a t distribution as in BayesA [8, 3], or a mixture of normally distributed effects as in BayesC [5, 4] or *t* distributed effects as in BayesB [8] give better results. Further, results from SS-BLUP will depend on the assumed relationship between genomic relationship coefficients and pedigree based relationship coefficients.

The limitations of SS-BLUP can be overcome by extending Bayesian regression models to accommodate animals with and without genotypes. Unlike SS-BLUP, the computational burden for Bayesian regression methods is linear in the number of genotyped individuals. Further, Bayesian regression models can accommodate a wider class of distributional assumptions for the marker effects.

# Methods

## Single-Step BLUP

We will first review the strategy underlying SS-BLUP and then show how this can be extended for Bayesian regression. Following Legarra et al. [7] the vector $\mathbf{g}$ of breeding values is partitioned as

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{M}_2\boldsymbol{\alpha} \end{bmatrix},$$

where $\mathbf{g_1}$ are BVs of the animals with missing genotypes $\mathbf{M}_1$, $\mathbf{g}_2$ are BVs of those with observed genotypes $\mathbf{M}_2$ and $\boldsymbol{\alpha}$ are the random effects of the genotypes. Assuming $\boldsymbol{\alpha}$ has null mean and covariance matrix $\mathbf{I}\sigma^2_\alpha$, the vector $\mathbf{g}_2$ has null mean and covariance matrix $\mathbf{M}_2\mathbf{M}_2'\sigma^2_\alpha$ conditional on $\mathbf{M}_2$. In order to derive the covariance matrix of $\mathbf{g}_1$ conditional on the change in distribution of $\mathbf{g}_2$, following Legarra et al. [7] the vector $\mathbf{g}_1$ is written as

$$\begin{aligned} \mathbf{g}_1 &= \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2 + (\mathbf{g}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2) \\ &= \hat{\mathbf{g}}_1 + \boldsymbol{\epsilon}, \end{aligned} \tag{1}$$

where $\mathbf{A}_{ij}$ are partitions of the relationship matrix, $\mathbf{A}$, that correspond to $\mathbf{g}_1$ and $\mathbf{g}_2$. The first term in (1) is the best linear predictor (BLP) of $\mathbf{g}_1$ given $\mathbf{g}_2$, and the second is a residual genetic effect to accommodate deviation between the true breeding value, $\mathbf{g}_1$, and its prediction from $\mathbf{g}_2$, $\hat{\mathbf{g}}_1$, which we refer to as $\boldsymbol{\epsilon}$, the "imputation residual". It is easy to see that $\boldsymbol{\epsilon}$ in (1) is uncorrelated to $\mathbf{g}_2$, and therefore if $\mathbf{g}_1$ and $\mathbf{g}_2$ are multivariate normal, $\boldsymbol{\epsilon}$ and $\mathbf{g}_2$ are independent. The conditional covariance matrix of $\hat{\mathbf{g}}_1$ given $\mathbf{M}_2$ is $\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\mathbf{M}_2'\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\sigma^2_\alpha$. The distribution of $\boldsymbol{\epsilon}$ is not affected by any change in the distribution of $\mathbf{g}_2$ because $\boldsymbol{\epsilon}$ and $\mathbf{g}_2$ are independent. Thus, the conditional covariance matrix of $\boldsymbol{\epsilon}$ is its unconditional covariance matrix: $(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\sigma^2_g$, where $\sigma^2_g$ is the genetic variance. Similarly, it can be shown that the covariance between $\mathbf{g}_1$ and $\mathbf{g}_2$ conditional on $\mathbf{M}_2$ is

$$Cov(\mathbf{g}_1, \mathbf{g}_2) = \mathbf{A_{12}}\mathbf{A}_{22}^{-1}\mathbf{M}_2\mathbf{M}_2'\sigma^2_\alpha.$$

Combining these results, the conditional covariance matrix of $\mathbf{g}$ given $\mathbf{M}_2$ is:

$$Var(\mathbf{g}|\mathbf{M}_2) = \mathbf{H} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}) & \mathbf{A_{12}}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathrm{A}_{21} & \mathbf{G} \end{bmatrix} \sigma^2_g, \tag{2}$$

where $\mathbf{G} = \mathbf{M}_2\mathbf{M}_2'/[\sum 2p_i(1-p_i)]$. This assumes that

$$\sigma^2_\alpha = \frac{\sigma^2_g}{\sum_j 2p_j(1-p_j)}, \tag{3}$$

Alternatively, $\mathbf{G}$ can be written as $\mathbf{G} = \mathbf{M}_2\mathbf{M}_2'\frac{\sigma_\alpha^2}{\sigma_g^2}$, which does not require assuming any relationship between $\sigma_\alpha^2$ and $\sigma_g^2$, but requires adding another unknown variable into the model. When the number of genotyped individuals is not too large, this matrix can be inverted efficiently as

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

but as the number of genotyped individuals grows, computing $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ will become infeasible.

## Single-Step Bayesian Regression

Consider the breeding-value model employed in SS-BLUP

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} + \mathbf{e}.$$

Setting up the mixed model equations for this model requires computing $\mathbf{H}^{-1}$. Thus, we replace $\mathbf{g}_1$ with (1) and $\mathbf{g}_2$ with $\mathbf{M}_2\boldsymbol{\alpha}$. Then, the model becomes

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \mathbf{M}_2\boldsymbol{\alpha} \end{bmatrix} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{M}}_1\boldsymbol{\alpha} + \boldsymbol{\epsilon} \\ \mathbf{M}_2\boldsymbol{\alpha} \end{bmatrix} + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\epsilon} + \mathbf{e}, \end{aligned}$$

where

$$\mathbf{U} = \begin{bmatrix} \mathbf{Z}_1 \\ 0 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1\hat{\mathbf{M}}_1 \\ \mathbf{Z}_2\mathbf{M}_2 \end{bmatrix}.$$

The matrix $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$ can be obtained without computing $\mathbf{A}_{22}^{-1}$ by solving the sparse system of linear equations:

$$\mathbf{A}^{11}\hat{\mathbf{M}}_1 = -\mathbf{A}^{12}\mathbf{M}_2, \tag{4}$$

where $\mathbf{A}^{ij}$ are partitions of $\mathbf{A}^{-1}$ that correspond to partitioning $\mathbf{g}$ into $\mathbf{g}_1$ and $\mathbf{g}_2$, and these are computed using Henderson's efficient algorithm [6].

If we assume normally distributed SNP effects, BLUP of $\boldsymbol{\alpha}$ can be obtained by solving the following mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} & \mathbf{X}_1'\mathbf{Z}_1 \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{W}_1'\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{X}_1 & \mathbf{Z}_1'\mathbf{W}_1 & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y}_1 \end{bmatrix}. \tag{5}$$

Note that these equations do not require $\mathbf{G}$ or its inverse. Nor does it require computing $\mathbf{A}_{22}$ or its inverse. The BLUP of breeding values are then given by

$$\tilde{\mathbf{g}} = \begin{bmatrix} \hat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix} \hat{\boldsymbol{\alpha}} + \mathbf{U}\hat{\boldsymbol{\epsilon}} = \begin{bmatrix} \hat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix} \hat{\boldsymbol{\alpha}} + \begin{bmatrix} \mathbf{Z}_1 \\ 0 \end{bmatrix} \hat{\boldsymbol{\epsilon}}. \tag{6}$$

More generally, inferences about marker effects or breeding values can be obtained by drawing samples from the posterior distribution of these variables from their posterior distributions. In practice, samples are not directly drawn from the posterior distributions. Rather, a Markov chain is constructed that has a stationary distribution identical to the posterior. Inferences from such a chain converge to those from the actual posterior [9, 10].

# Discussion

Single-step Bayesian regression (SSBR) has the same advantages as SS-BLUP. In both, phenotypes, pedigree and genotypes are combined in a single step to obtain GEBV. SSBR, however, is not limited to normally distributed marker effects. Further, as the number of genotyped animals grows, SSBR will have a computational advantage as computing time to draw samples from the Markov chains is linear in the number of animals.

# References


[1] I Aguilar, I Misztal, D L Johnson, A Legarra, S Tsuruta, and T J Lawlor. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *J Dairy Sci*, 93(2):743–752, Feb 2010.

[2] D J Garrick, J F Taylor, and R L Fernando. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*, 41(1):55–55, 2009.

[3] D Gianola, G de los Campos, W G Hill, E Manfredi, and R Fernando. Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363, Sep 2009.

[4] D. Habier, R. L. Fernando, K. Kizilkaya, and D.J. Garrick. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12:186, 2011.

[5] D. Habier, R.L. Fernando, K. Kizilkaya, and Garrick. D. J. Extension of the Bayesian alphabet for genomic selection. In *Proc. 9th World Congress on Genet. Appl. Livest. Prod.*, volume 9, page 468, 2010.

[6] C. R. Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32:69–83, 1976.

[7] A Legarra, I Aguilar, and I Misztal. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*, 92(9):4656–4663, Sep 2009.

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

[9] J. R. Norris. *Markov Chains*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, New York, 1997.

[10] D. A. Sorensen and D. Gianola. *Likelihood,Bayesian,and MCMC Methods in Quantitative Genetics*. Springer, 2002.

[11] P M VanRaden, C P Van Tassell, G R Wiggans, T S Sonstegard, R D Schnabel, J F Taylor, and F S Schenkel. Invited review: reliability of genomic predictions for north american holstein bulls. *J Dairy Sci*, 92(1):16–24, Jan 2009.