Consequences of Genomic Prediction in Cattle

D.J. Garrick[1,2]*******

[1]Department of Animal Science, Iowa State University, Ames, IA 50011, USA
[2]Institute of Vet., Animal & Biomedical Sciences, Massey University, Palmerston North, New Zealand
*(*corresponding author, e-mail: dorian@iastate.edu, Phone:001 515 509 1582)*

Abstract

Genomic prediction represents a revolutionary advancement in animal improvement by providing a means of improving the accuracy of estimated genetic merit for selection candidates with no individual or offspring records. Despite the fact that it has already been widely used in dairy cattle improvement and is now available in some beef cattle circumstances, its predictive ability in terms of accounting for mendelian sampling rather than parent average merit is still poorly characterized. Improved characterization will be required in order for alternative breeding strategies to be fairly compared. In the meantime, one consequence is that predictions will work better in offspring than in less related animals. Among many other consequences of genomic prediction, it promises opportunities for basic research to be undertaken using industry populations along with information from routine evaluation, it promises more balanced selection, and it necessitates major developments in national and international software and evaluation procedures.

Keywords: genomic prediction, across-breed prediction, selection response, relationship matrix

1. Introduction

Genomic selection is based on estimated breeding values (EBV) that have been obtained using high-density genotypes spanning the whole genome. The process of genomic prediction conceptually involves three steps, although these activities may be combined in a single analysis. The steps are: characterization of the EBV of chromosome fragments, called the discovery or training analysis; prediction or summing up of the values of all the fragments carried by each selection candidate; and blending to pool information from genomic and conventional pedigree and performance analysis. The approach promises faster genetic gain, and in certain circumstances, possibly at less cost or with lower rates of inbreeding than conventional breeding schemes based on individual measurement and progeny testing. In order to characterize the relative rates of gain, cost and inbreeding, it is necessary to know the predictive ability, and the costs of genotyping, among other factors. Bovine whole-genome genotyping has principally been undertaken using the Illumina 50k beadchip for the last two years, with current costs at US$175. However, higher density (≈800k @ US$270) and lower density (3k@unspecified price) Illumina beadchips have been promised for release later in 2010 as has a competitive high-density Affymetrix product, making objective comparison of alternative breeding schemes difficult with the cost uncertainty. This paper focuses on what is known about current predictive ability using 50k panels in cattle, and introduces some consequences of the current status of genomic prediction in cattle.

2. Current Genomic Prediction

2.1 Current Predictive Ability

Breeding Values (BVs) can be considered as the sum of Parent Average (PA) effects plus Mendelian Sampling (MS) effects. In an unselected population, each of these two components contributes 50% genetic variance. The goal of genomic prediction is to accurately estimate MS, because if this can be achieved PA can also be accurately estimated as PA is no more than accumulated MS effects from grandparental and earlier generations. The converse is not true; it is possible to accurately estimate PA without being able to estimate MS. In fact, that is the circumstance in conventional BV prediction using a pedigree-based relationship matrix to predict the merit of individuals without their own or offspring records.

Simulated data including quantitative trait loci (QTL) and markers such as single nucleotide polymorphisms (SNP) have demonstrated predictive abilities of 0.7-0.9 (Meuwissen *et al.*, 2001), accounting for 50-80% total genetic variance. It is tempting to assume such predictions do not discriminate between PA and MS and estimate both components equally well.

Field data representing observed performance and actual SNP genotypes often achieve high correlations in the training data, but it is the correlation in new subpopulations that are of more interest. Early whole genome analyses of the North American Holstein population (VanRaden *et al.*, 2009) reported the PA reliability (R^2) of animals without records or offspring to average 0.19 across traits and the genomic prediction to improve on that value by a further 0.18 increase. In the international collaborative analysis of Brown Swiss performance, Jorjani and Zumbach (2010) similarly compared either the PA predictions from conventional evaluations or the genomic evaluations from 50k SNP panels, to the subsequent performance of progeny tested daughters, four years later. Those analyses also demonstrated an increase in reliability of 0.18. Some of that increase in predictive ability is likely due to improved prediction of PA, but assuming all of it was due to MS, and MS accounted for 50% genetic variance, then it would seem that these genomic predictions are predicting up to 36% genetic variance.

There are fewer published reports of genomic predictions in beef cattle. Analyses of US Angus bulls based on their published expected progeny differences (EPD) for a range of traits resulted in correlations between two-thirds data used in training and one-third used for validation as in Table 1 (from Garrick, 2009). In that study, the AI bulls were randomly allocated to three subsets according to the sire of the bull, such that paternal half-sibs were not represented in more than one of the subsets. The pooled correlations between genomic and realized performance ranged from 0.5-0.7, accounting for 25-50% genetic variance.

Table 1. Correlations between 50k genomic prediction and realized performance for validation of Angus sires in independent Angus datasets for backfat (BFat), calving ease direct (CED) and maternal (CEM), carcass marbling (Marb), carcass ribeye area (REA), scrotal circumference (SC), weaning weight direct (WWD) and yearling weight (YWT).

Trait	Train 2 & 3 Predict 1	Train 1 & 3 Predict 2	Train 2 & 3 Predict 3	Overall[1]
BFat	0.71	0.64	0.73	0.69
CED	0.65	0.47	0.65	0.59
CEM	0.58	0.56	0.62	0.53
Marb	0.72	0.73	0.64	0.70
REA	0.63	0.63	0.60	0.62
SC	0.60	0.57	0.50	0.55
WWD	0.65	0.44	0.66	0.52
YWT	0.69	0.51	0.72	0.56

[1] Overall correlation estimated by pooling the estimated variances and covariances from each separate validation.

2.2 Spurious Markers that are Predictive in Training

Previous research aimed at discovering QTL for marker-assisted selection (MAS) was often characterized by media releases reporting discoveries that would revolutionize selection. In the US, enlightened producers demanded such discoveries be validated before their widespread adoption. The validation studies (www.nbcec.org) were typically undertaken in sub populations that were reasonably independent of the discovery population, and spurious markers were readily identified. However, the extent of genomic training populations has now commonly expanded to include entire populations of AI bulls, making it impossible to validate discoveries in unrelated animals of the same breed.

Consider a discovery population that is segregating a QTL, such that training animals can be categorized as having 0, 1 or 2 copies of the favorable allele. The goal of genomic selection is to identify a physically linked polymorphism that can be used as a surrogate to indicate the number of copies of the QTL allele. The strength of the relationship between the QTL and the marker is measured by their linkage disequilibrium (LD). Given sufficient markers, there will commonly be a marker or a linear combination of the markers that are predictive of the number of QTL alleles in the training population, but that marker (or markers) may not necessarily be physically linked to the QTL. The informative marker in the training population may even be on another chromosome from the QTL. Nevertheless, the marker would be predictive in training. Such markers would be unlikely to demonstrate any predictive ability in independent validation. However, applying these markers to offspring of the training animals can demonstrate spurious predictive ability, due to linkage rather than LD. If the marker alleles were perfectly predictive of the number of QTL alleles in training, then the markers would correctly identify the gametes of bulls with 0 copies of the favorable allele, and correctly identify the gametes of bulls with 2 copies of the favorable allele. Only in heterozygous bulls whereby half the gametes would carry the favorable QTL allele, and an independent half of the gametes carry the favorable marker allele would the relationship breakdown. Even in those gametes, the marker would by chance correctly predict the favorable QTL allele half of the time.

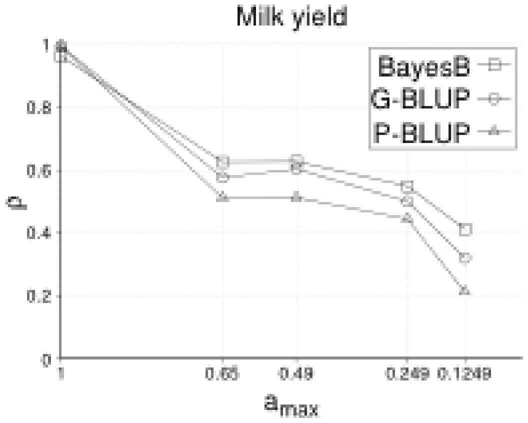
Genomic predictions are based on summing the effects of many genomic regions, some which might be correctly identified in training while others are spurious. Validation applied collectively to all the genomic effects would therefore exhibit some loss of predictive ability, with continued erosion in successive generations.

2.3 Validation in Relatives and in Other Breeds

Two options to further characterize the predictive ability of genomic approaches might be considered. First, one could partition available populations into training and validation subsets, in a manner that controls the degree of relatedness between the populations. Second, validation could be undertaken in another breed.

A recent publication by Habier *et al.* (2010a) partitioned the German Holstein population into training sets, in such a way to control the maximum pedigree-based additive genetic relationship between any bull in validation and all bulls in training. This partitioning was repeated in four scenarios to vary the level of relationship. Random partitioning resulted in additive relationships as high as 0.6 between training and validation bulls. Restricting the maximum relationship to 0.49 produced partitions that prevented parent-offspring relationships or splitting of full-sibs across training and validation subsets. Restricting the maximum relationship to 0.249 prevented grand-parental and half-sib relationships across training and validation subsets. A further scenario prevented maximum additive relationships from exceeding 0.1249. Creation of these scenarios required that some bulls be excluded from both the training and validation subsets. Interestingly, these scenarios had little impact on the average maximum relationship between training and validation subsets that remained at about 9% for the first three scenarios. The results in terms of correlations are in Figure 1, for predictions based on 1,048 training bulls using conventional pedigree analysis (P-BLUP), and for methods using genomic relationship matrices with equal (G-BLUP) or heterogeneous SNP weighting (BayesB). Clearly the genomic predictions outperform pedigree-based methods, justifying their continued implementation, but the reduction in predictive power for the 0.1249 scenario is alarming for the use of genomic predictions for traits that are not routinely phenotyped every generation, as is the goal for beef cattle implementations for traits associated with reproduction, feed intake, disease and eating quality.

Figure 1. Correlations (ρ) between genomic predictions based on samples of 1,048 German Holstein training bulls and observed performance in validation subsets with training and validation animals partitioned to control the maximum additive relationship (a_{max}) between any validation bull and all training bulls.



Further validation analyses have been undertaken using North American Holsteins for a small (1,000 bull) or large (4,000 bull) training set comprising animals born after 1994, validated in animals born before 1975 (Habier *et al.*, 2010b). Those results (Table 2) fail to account for more than 28% genetic variance. However, the validation bulls would have been assessed from progeny performance in management circumstances quite different from today, so both heterogeneous variance and genotype-environment interaction could have contributed to the reduction in predictive ability.

Table 2. Correlations between North American Holstein 50k genomic predictions from 1,000 or 4,000 training bulls born after 1994 and realized performance for bulls born before 1975.

	Training bulls	
Trait	1,000	4,000
Milk	0.42	0.44
Fat	0.48	0.52
Protein	0.15	0.18
Somatic Cell Count	0.14	0.28

Validation of genomic predictions in other breeds provides a worst-case scenario in terms of predictive ability. Across-breed predictions could perform poorly because of dominance, epistasis, genotype-environment interactions, variation in LD among breeds, among other reasons. Training analyses based on North American milk yields from 8,512 Holstein bulls resulted in correlations of 0.194 in 742 Brown Swiss bulls and 0.198 in 1,915 Jersey bulls from Bayes A, and 0.141 in Brown Swiss and 0.244 in Jersey from Bayes B. Training in two of the three breeds and validating in the third, resulted in correlations of 0.077 in Brown Swiss, 0.197 in Jerseys and 0.253 in Holsteins. Linkage cannot be contributing to these predictions, only LD and that accounts for less than 10% genetic variance.

Any improvement on the accuracy of PA predictions provides opportunities for improved breeding schemes. These results clearly indicate that genomic techniques can increase predictive ability and therefore have an immediate role in breeding schemes. However, there remains enormous potential for increasing the predictive ability to the levels that can be obtained in simulated data. Research is urgently required to further investigate methods in which some of this potential might be exploited in the near term.

3. Some Consequences of Current Circumstances

3.1 Convergence of Basic and Applied Research

Animal evaluation is applied research, and its implementation identifies, among selection candidates, those animals that are the best prospects to be parents of the next generation. Basic research includes investigating the biological mechanisms by which the offspring of some individuals outperform the offspring of other individuals. Such studies have long been of interest to animal scientists, although to date physiologists have had more success in identifying the mechanisms involved in average levels of performance than in identifying mechanisms that are responsible for variation in performance. Over the last two decades, numerous studies were undertaken to discover QTL, but most of these studies had little involvement with applied animal evaluation, other than perhaps using national EBV as data.

The BV can be defined from a basic viewpoint as the sum of the average effects of alleles, summed over all the loci influencing a trait and the pair of alleles at each locus, from an applied viewpoint it has been a black box approach involving little more than a regressed estimate of twice the deviation of offspring performance. The black box has the potential to be opened up with the introduction of genomic predictions. Genomic merit is computed as the sum of the allelic effects, and routine evaluation characterizes the value of every genomic location as part of the prediction process. Accordingly, it makes sense for basic and applied researchers to work together and exploit all the information obtained from routine evaluation. Knowledge of genomic locations that influence variation will allow the incorporation of biological information in the statistical analyses that drive genomic prediction. Inspection of the variation accounted for by genomic regions demonstrates that the reduced ability to predict milk yield in Brown Swiss from across breed predictions shown in the previous section is at least in part due to the fact that based on 50k SNP, DGAT1 does not appear to be segregating in Brown Swiss. Recognition of the portfolio of genomic regions or QTL that are common or unique to particular breeds would increase the accuracy of across-breed prediction.

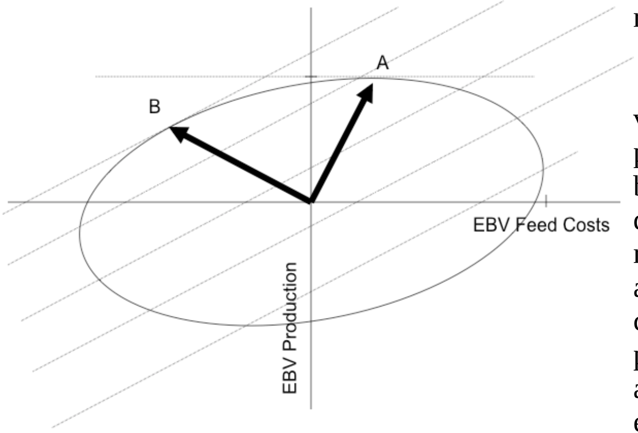
3.2 The Nature of Genetic Gain

Selection should be focused on an objective derived from a breeding goal. For profit-based goals, it should include a list of economically-relevant traits along with their relative emphasis. Such a list for beef or dairy cattle would typically include production, reproduction, disease, longevity and feed requirements. In practice, many of these traits in the selection objective do not have corresponding selection criteria that can be cheaply and readily observed for use in animal evaluation. In those circumstances, animal breeders have searched for indicator traits that can exploit correlations to predict merit, but for some traits such indicators are not apparent. Selection may therefore be practiced on only a subset of the traits in the selection objective.

Genomic prediction offers the potential to predict merit for traits that are difficult, expensive or impossible to measure in conventional circumstances. Such traits could then be included in the selection objective, without phenotypic information. That would lead to a change in the nature of genetic gain. Current implementation of genomic selection in dairy cattle has focused on improving on the PA prediction for phenotyped traits such as production, whereas in beef cattle the major focus has been on difficult traits such as feed intake, reproduction, animal health and eating quality.

Consider the nature of genetic gain for a circumstance relevant to beef and dairy cattle where profit is influenced by productivity and feed costs. The selection objective would ideally reward more productive animals according to the value of their production at the same time as penalizing them in relation to their corresponding feed costs. The best animals would be those that have the largest margin of productive income over feed costs.

Figure 2. A schematic ellipse representing the selection frontier for productivity (y-axis) and feed costs (x-axis). Relative to the average of the population at the origin of the ellipse, point A represents the outstanding selection candidates for production that would be selected if EBV were not available for feed costs. Point B represents outstanding candidates for profit, defined as the value of production less feed costs, for the angled iso-profit lines determined by the ratio of economic values for production and feed costs. Animals at B would be selected if candidates could be ranked for both traits. The arrow to A indicates the selection response increases production and feed costs if EBV were only available for production, whereas the arrow to B indicates that goal-based selection increases production, reduces feed costs, and collectively achieves a greater response in profit than selection for production alone.



3.3 *Ongoing Development of Genetic Evaluation Software*

A major consequence of genomic selection is that existing national evaluation software requires ongoing enhancements to exploit genomic information. Initial implementations involved several steps, namely training analyses, genomic prediction of selection candidates followed by blending of genomic and conventional pedigree information. Combining these steps into a single analysis has appeal, and has been promoted by Mizstal *et al.* (2009) Legarra *et al.* (2009) and Aguilar *et al.* (2010). The latter paper exploits a variance-covariance matrix among genotyped animals according to a genomic matrix (**G**), whereas variance-covariances among pedigree animals and between pedigree and genotyped animals involve modifications to the pedigree relationship matrix according to departures between the pedigree and genomic relationship matrices. Those departures are $(\mathbf{G} - \mathbf{A}_{22})$ and $(\mathbf{G}\mathbf{A}_{22}^{-1})$, respectively, as shown below.

$$\text{var}\begin{bmatrix} u_{pedigree} \\ u_{genotyped} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}\sigma_g^2$$

This representation raises an interesting issue. Suppose some pedigree animals have genotyped relatives without phenotypic records. In that case, rather than using the usual inverse of the pedigree-based relationship matrix $(\mathbf{A}_{11}^{-1})$, a better choice might be the inverse of the matrix adjusted for the known genomic relationships among the relatives.

The relationship matrix defines the variance-covariances on the basis of probability theory. Probabilities must be between 0 and 1, and cannot be negative. The diagonal elements of the relationship matrix represent twice the probability that one allele at a locus is identical by descent to a randomly chosen allele at that locus. Since there is a probability of 50% that the randomly chosen allele is the original allele, the diagonal element cannot be less than unity. Similar arguments demonstrate that the covariance cannot be negative. These features are consequences of an infinitesimal model.

In practice, gametes do not have a random sample of alleles from an infinite number of loci, as the number of chromosomes is finite, and a typical pair of parental chromosomes experiences one crossover event at meiosis. Accordingly, the realized probability that alleles are identical by descent between relatives such as full-sibs may be a little larger or a little smaller than the average probabilities used in the relationship matrix. Genotypic information can track genomic inheritance, and allow such departures from expectation to be quantified in the genomic relationship matrix. Consider the impact of such departure on modifying the assumed average relationships of parents. Suppose two parents are unrelated and non-inbred. Their relationship matrix is therefore an identity matrix of order 2. Suppose they produce two full-sibs, and the genomic relationship matrix shows the fullsibs to be a little more closely related than expected, having a relationship with each other of 0.53 rather than 0.5. Note that the genomic relationship matrix is estimated as a covariance matrix, rather than based on probability. Applying Legarra *et al.* (2009), the parents are estimated to be slightly inbred, and slightly related. This makes sense, in order to have produced fullsibs that are more closely related than 0.5, from an infinitesimal model. However, suppose as is equally likely, the fullsibs were less rather than more related than expected, with a relationship with each other of 0.47 rather than 0.53. In that case, Legarra *et al.* (2009) results in the parents having relationships with themselves of less than unity, and a negative covariance. Although such an outcome is possible from a variance-covariance framework, it is inconsistent with our usual probabilistic approach to genetic relationships. It is also different from the answer that is obtained if the exact approach is used to predict the parental relationship matrix conditional on the genotypes observed on the offspring. It would be interesting to apply these calculations to national evaluation data to determine if this modification improves the predictive ability of PA calculations. Such modifications not only have implications in the prediction of genetic merit, they also impact the calculation of reliabilities. Ongoing research is clearly warranted to determine the most appropriate methods and the computational and interpretational considerations of these methods.

4. Conclusions

Genomic prediction has been confirmed in several studies to usefully increase the accuracy of prediction for young animals without individual or offspring records. Nevertheless, its predictive ability remains below values predicted from simulation studies. Further, the extent to which it exploits linkage disequilibrium to predict mendelian sampling effects free from distortion due to linkage signals remains poorly characterized. This knowledge gap is more important in circumstances whereby training analyses are being undertaken for prediction in unrelated animals than when offspring are being predicted.

Genomic prediction offers new opportunities for interaction between biologists trying to identify the causal nature of variation in inherited performance and those applied scientists involved in routine evaluation of selection candidates.

The use of genomic prediction to rank animals for traits that have not been routinely phenotyped provides opportunities for more balanced selection than is the case when animals can only be evaluated for a subset of economically-relevant traits. However, practitioners need to be a cautious in modeling the effects of such selection when the long-term predictive ability of genomic-based methods is still uncertain.

Software and other components of the information systems used to collect, analyze and report estimates of national and international genetic merit and corresponding reliabilities will require ongoing revision over the next few years as the philosophical basis, statistical approach and interpretation becomes progressively resolved by ongoing research endeavors.

References

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

Garrick, D.J. 2009. The nature and scope of some whole genome analyses in US beef cattle. *Proceedings of the Beef Improvement Federation,* 41st Annual Research Symposium April 30 – May 3, 2009, Sacramento, California, USA. *Volume 41,* 92-102.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P. & Thaller, G. 2010a. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol. 42,* 5.

Habier, D., Fernando, R.L., Kizilkaya, K. & Garrick, D.J. 2010b. Extension of the Bayesian alphabet for genomic selection. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production.* In press.

Jorjani, H. & Zumbach, B. 2010. Preliminary results from validation tests. *Ibid*.

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

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information. *J. Dairy Sci. 92,* 4648-4655.

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

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92,* 16-24.