Milestones in genetic prediction for applied livestock improvement

D.J. GARRICK*

Lush Chair in Animal Breeding and Genetics, Iowa State University, Ames, Iowa, USA
*Corresponding author: dorian@iastate.edu

ABSTRACT

Realised rates of genetic improvement depend upon structural, technological, economic and cultural factors. The potential nature and rate of genetic improvement of livestock has largely been dictated by accuracy of the predicted genetic merit of selection candidates. Major milestones in predicting merit include the use of objective measurements rather than visual appraisal; adjustment of measured performance for nongenetic factors; use of aggregate measures of economic merit for multiple trait selection; combining of information from all available sources of relatives and traits; and properly accounting for previous selection. Despite these advances, use of pedigree and performance information has major limitations. Genomic technologies can trace inheritance of chromosome fragments and with knowledge of the fragment values can improve accuracy of prediction in young selection candidates. Current genomic methods perform more poorly with real than with simulated data, and predictive abilities erode when applied to target animals not closely related to the population used to quantify fragment values. Several strategies are being investigated to improve genomic predictions, but in the meantime the principal benefit of the technology has been to increase accuracy of predicted merit of young selection candidates for routinely measured traits. In some cases this has reduced the generation interval.

Keywords: genetic improvement; BLUP; mixed models; genomic prediction.

INTRODUCTION

Genetic improvement is the straightforward result of using above-average candidates as parents of the next generation. Ongoing improvement requires co-ordinated science and technology, in concert with market circumstances and many other non-genetic factors that influence business success. The rate and nature of genetic gain of any seedstock operation must be competitive with alternative seedstock suppliers, and the breeding programs and associated infrastructure must cost less than the margins that can be routinely harvested on the sale of improved seedstock. Many of the critical factors vary in different livestock species, leading to quite different business models and industry structures. Nevertheless, there have been several research milestones that are common to improvement in practically all industries. These are: the concept of aggregate economic merit for multiple trait selection; the selection index to combine information from various sources; best linear unbiased prediction to account for systematic non-genetic effects and the effects of prior selection; the inverse relationship matrix to include information from all known relatives of the selection candidates; and most recently, genomic prediction to incorporate into prediction DNA information on inheritance of chromosome fragments. This paper will describe the context of and then focus on, the current nature and status of genomic prediction.

The prediction problem and historical developments

Prior to the advent of performance recording, selection was undertaken by choosing as parents those individuals that most closely fitted the visual image of the ideal animal. The use of recorded performance advocated by Dr. J.L. Lush (Lush, 1937) and others, heralded a more objective basis for selection. That approach begins with defining the goal for selection. The goal, such as profit per unit of land, describes what is to be achieved by selection and leads naturally into the development of the selection objective. The selection objective describes the list of traits that influence the goal, and their relative emphasis. In the context of a profit-based goal, the list of traits should include those that influence income and/or costs of production. Their emphasis should be based upon their economic or relative economic values. Economic values might be defined as the partial derivative of the profit function, equivalent to the change in profit from a unit change in one trait in the objective, with all other traits in the objective held constant.

The idea of formally defining a selection objective as a linear function of trait information is a concept introduced by Hazel (1943). It is now widely known as a selection index and is used in many livestock industries. The determination of economic values remains a practical problem, particularly in the context of determining the future value of various traits given uncertainties in economic, political and technological arenas.

Hazel (1943) argued that net genetic improvement is the sum of the genetic gains for the traits of economic importance, each weighted by the relative economic value of that trait. That paper went on to demonstrate an approach to directly derive the aggregate economic merit, or "$Index" value of each candidate, as a linear function of the adjusted phenotypic measurements. An equivalent approach is to predict the merit of each trait in the breeding objective, and weight these by their relative economic value. Thus

$$\$Index_j = \hat{\mathbf{g}}_j'\mathbf{v} = \sum_{i}^{traits} \hat{g}_{ij} v_i \qquad \text{Equation 1}$$

where **v** is a vector whose ith element is the relative economic value for the ith trait and $\hat{\mathbf{g}}_j$ is the vector of predicted breeding values for the jth animal.

The development of selection index theory (Hazel, 1943), included more than an index of aggregate economic merit that could reward superior performance in some traits, and concurrently penalize those same individuals for inferior performance in other traits. It included statistical theory, developing a concept now known as best linear prediction (BLP), to combine information from various sources. In this context "best" means minimum prediction error variance, which is the variance of the differences between predicted and true unobservable merit. The method of BLP guarantees the best prediction possible, among linear methods of predicting performance, from observations jointly influenced by random genetic and residual causes. It requires that observations have been correctly adjusted for systematic or fixed non-genetic effects. This assumes parameters for factors such as flock or herd, year, age at measurement, sex, and cohort group are known without error.

Operationally, a vector **y** of phenotypic observations is adjusted for the parametric values of the fixed effects. That is, $\mathbf{X}\beta$, and the resultant vector of deviations $(\mathbf{y} - \mathbf{X}\beta)$ are multiplied by a vector of weights $\mathbf{b_{ij}}$ for the ith trait being predicted on the jth animal, as in

$$\hat{g}_{ij} = \mathbf{b_{ij}}'(\mathbf{y} - \mathbf{X}\beta) \qquad \text{Equation 2}$$

with the vector of weights varying according to the nature and scope of available phenotypic information.

There were two circumstances where combining information from different sources had been problematic. Both were addressed within the theory of BLP. First, that of combining information from various relatives, such as the parents, offspring, half-sibs and full-sibs. Closer relatives provide more information than distant relatives. Ancestors are informative as to average merit of the parents but cannot contribute to prediction of the effects of Mendelian sampling that allows the merit of offspring to deviate from their parent average. Descendants provide information on the realized merit of the ancestor, including both its parent average merit and its deviation from the parent average. Second, information on the merit of an individual can be obtained from the trait of interest, and from correlated traits. Correlated traits may demonstrate phenotypic similarity due to genetic or residual covariation. Phenotypic similarity from these two sources needs to be separated, and the contributions of each trait weighted according to their information content that will depend upon trait heritabilities and genetic correlations. Weights for various information sources derived using BLP take account of all these factors and generate the best predictions in the context of the smallest variance of the prediction errors, or equivalently, the largest correlation between true and predicted merit.

The vector of weights is obtained by setting up and solving a set of simultaneous equations. In those equations, the left-hand side matrix **P** contains the phenotypic variance-covariance matrix for the selection criteria, and the right-hand side contains a vector comprised of the ith column of the genetic-variance covariance matrix, **G**, as in

$$\mathbf{Pb_{ij}} = \mathbf{G}_{column\ i} \qquad \text{Equation 3}$$

a special case of $\mathbf{Pb_{ij}} = \mathbf{Gv}$ with **v**, an elementary vector with unity for its ith element and zeros elsewhere. This approach of deriving the weight for each information source using Equation 3 was introduced in 1950 to sheep improvement in New Zealand by Professor A.L. Rae, and was eventually implemented in the National Flock Recording Scheme, and then Sheeplan, to obtain estimated breeding values (EBV) for each trait using Equation 2 and then combining them into an index using Equation 1.

Rather than explicitly deriving the vector of linear weights for each trait on each animal, it is possible to directly compute breeding values in one operation. The approach is relevant to a more general model equation (Henderson, 1963)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zg} + \mathbf{e}, \qquad \text{Equation 4}$$

which includes an incidence matrix **Z** relating the vector of breeding values to particular observations in **y**. This allows for sire, sire and dam, sire and maternal grandsire, or animal models with maternal effects. In an animal model with only one record per animal, **Z** is an identity matrix.

The general form of equations to be solved to obtain EBV, such as $\hat{\mathbf{g}}$, are

$$\left[\mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}}\right]\left[\hat{\mathbf{g}}\right] = \left[\mathbf{Z'R^{-1}}\left(\mathbf{y} - \mathbf{X}\beta\right)\right].$$

Equation 5

This formulation is appealing, but has several limitations. First, it requires the fixed effects to be known without error. Second, it requires knowledge of the variance-covariance matrices, **R** and **G**. Third, it requires these matrices to be inverted, the inverses being used in matrix multiplication and addition in order to obtain the equations to solve to estimate breeding values. Solving the equations themselves was less of a problem, even in the early days of computing, as iterative methods such as Gauss-Seidel were known to be tractable, provided the number of elements in the left-hand side or coefficient matrix was limited in relation to available computer storage.

These limitations with BLP were largely overcome as a result of the research efforts of Dr C.R. Henderson, who like L.N. Hazel and A.L. Rae, was a graduate student of Dr Lush. Henderson was interested in estimating unknown fixed effects from the data, recognizing the difficulties of assuming these effects be known without error. He was motivated to estimate the non-genetic effect of having a second versus first lactation in dairy cattle, recognising poor performing cows were culled after their first lactation. Least squares estimates of the parity effect would be biased upwards by selection. During the course of his PhD, Dr Henderson developed equations whose solutions he believed would have favourable properties. He referred to these as the mixed model equations, not to be confused with the equation for the mixed model given in Equation 4. The mixed model equations are

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

. Equation 6

The second row in Equation 6,

$$\begin{bmatrix} \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

can be rearranged to

$\left[\mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}}\right]\left[\hat{\mathbf{g}}\right] = \left[\mathbf{Z'R^{-1}}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)\right]$, which is identical to Equation 5 except that observations are adjusted for a particular estimate of the fixed effects rather than their parametric values.

The coefficient matrix in Equation 6 is like the generalized least squares coefficient matrix from treating all effects as if fixed, and using the residual variance-covariance matrix as the only source of variation, except that the inverse of the variance-covariance matrix of the random effects, such as breeding values, is added to its corresponding elements of the coefficient matrix. However, it was some time before the favourable properties of these equations could be proven; a prerequisite for widespread acceptance of the method. That proof (Henderson *et al.,* 1959) was achieved by Shayle Searle while Dr Henderson was on sabbatical with the New Zealand Dairy Board, and involved demonstrating estimable linear functions of the fixed effects had the properties of best linear unbiased estimates (BLUEs). They coined the term best linear unbiased predictions (BLUP) to describe the solutions to the random effects.

In single trait settings, residual effects are often independent, so matrix **R** is diagonal, and trivially inverted. In multiple trait settings, residual covariances typically exist between traits, but not between animals, resulting in a block diagonal **R** matrix that can be readily inverted by inverting the small matrix of residual covariances among the traits. A major limitation was the construction of $\mathbf{G}^{1}$, the inverse of the variance-covariance matrix of breeding values.

In general, **G** consists of two components. One is the genetic variance-covariance matrix among traits measured on the same animal. That matrix can be constructed from knowledge of the phenotypic variances of the traits, the heritabilities and genetic correlations. The other component is the variance-covariance matrix among the animals represented in the pedigree. That matrix is the numerator relationship matrix and can be constructed directly from knowledge of the pedigree. In a typical livestock pedigree, the relationship matrix becomes dense, as non-zero coefficients accumulate between ancestors and descendants and among family members such as half- and full-sibs.

The inverse of **G** can be directly constructed from inverses of the two components – the variance-covariance matrix between traits on the same animal which has a low order and is easy to invert, and $\mathbf{A^{-1}}$, the inverse of **A**, the numerator relationship matrix. Even if **A** could have been constructed, computing its inverse represented a formidable task until Dr Henderson recognized its simple structure and that it could be directly constructed by accumulating elements based on knowledge of the pedigree, without ever forming **A** (Henderson, 1976; Quaas, 1976). That development, more than 30 years ago, opened the door to routine genetic evaluation for multiple traits taking account of entire national populations of recorded animals. Since that time, there have been numerous small advances in computing algorithms, such as iteration on data and conjugate gradient for solving linear equations, and

other factors for specific models, such as maternal effects, genetic groups, multi breed, heterogeneous variance, international evaluations and categorical traits. Nevertheless the foundations of animal ranking rely on the selection index, the mixed model equations and the inverse of the numerator relationship matrix. The procedures were adopted in New Zealand for sire referencing analyses of sheep from 1989, and became routine for all sheep traits from 2000 after the formation of Sheep Improvement Limited. Prototype systems using this approach for NZ dairy cattle were developed in 1992 (Garrick *et al.*, 1993) and became the basis for the current across-breed animal model evaluation system used by Livestock Improvement Corporation (Harris *et al.*, 1996). Similar procedures were used for beef cattle in New Zealand evaluated using Breedplan in Australia (Meyer & Garrick, 1995).

One practical problem is that the model must be known in order for the estimates to have the desired properties. This requires knowledge of the model equation that describes the systematic non-genetic effects, and all the random causes of variation, as well as the distributional properties, particularly the variance-covariance parameters for the random effects. Statistical analysis can be used to estimate these parameters and compare model equations, but these are more technically demanding than predicting breeding values. Enormous advances have been made in that area, including the use of the average information matrix (Gilmour *et al.*, 1995) and sparse matrix techniques that have led to software such as ASREML (Gilmour *et al.*, 2009), the current state of the art for a diverse range of models. Dr Gilmour has strong links with New Zealand, having studied for his Ph.D. at Massey University with Professors A.L. Rae and R.D. Anderson.

Despite these advances, there is a fundamental problem with pedigree-based prediction that has limited genetic improvement in practically all livestock species. That limitation arises because of several factors. These include selection decisions need to be made no later than puberty to minimize generation intervals and costs of maintaining selection candidates prior to their widespread use, selection is typically for a multiple trait objective, many of the traits cannot be cheaply and readily recorded in both sexes by puberty, and the accuracies of predicted breeding values are limited in the absence of measured performance records for every trait on the individual or its offspring. That limitation has led to different breeding scheme designs in different species to optimize generation intervals, selection intensities, accuracies of prediction and costs of measurement within that industry.

The limitation of conventional approaches

Estimates of genetic merit have associated accuracies and reliabilities. Accuracy of prediction refers to the expected correlation between estimated and true unobservable breeding values. It is a function of the particular sources of information used in prediction and the respective weights in Equation 3. The square of the accuracy is known as reliability (r^2) and can be interpreted as the proportion of variation in true merit that can be accounted for using the information at hand. Another interpretation is that 1 - r^2 is the proportion of genetic variance that cannot be explained given the available information.

The weights from solving Equation 3 used to combine information from alternative sources are a complex function of additive relationships. These additive relationships obtained from pedigree information reflect the expected or average relationship. For example, the relationship between non-inbred unrelated parents and their offspring is one-half, as is the relationship between full-sibs of non-inbred unrelated parents. However, given a finite number of genes influencing a trait, the actual relationship between a pair of full sibs might average one-half, but some full sibs could be more related and others less related to each other. Any departure between average and actual relationships cannot be inferred from the pedigree alone.

The additive merit of a selection candidate without its own records or any offspring can be estimated from knowledge of the merit of its parents using

$$r^2_{offspring} = \left(r^2_{sire} + r^2_{dam}\right)/4 . \qquad \text{Equation 7}$$

Even with near perfect information, the upper limit is $r^2 \leq 0.25$ if one parent is known and $r^2 \leq 0.5$ if both parents are known. These values correspond to $r \leq 0.5$ or $r \leq 0.7$, respectively, limiting gain per generation to between 50% and 70% of what could be achieved if candidates could be perfectly assessed by selection age.

Many of the traits that are economically important cannot be measured by the time of puberty, in both sexes, such as milk yield and egg size; without sacrificing the animal as a breeding prospect such as meat tenderness and disease resistance; or waiting until the end of its life to assess its longevity. Others are simply too expensive to measure on every candidate, such as feed intake. Even when the selection candidate can be measured early in life, the value of that information is limited by its heritability. Breeding values for low heritability traits cannot be accurately assessed without measuring large numbers of progeny. The net effect is that reliability of estimated overall merit is typically sub-optimal at puberty.

The biological explanation for the 0.50 ceiling on r^2 is apparent when the genetic merit of an individual is expressed as the average of its parents, plus a term representing its deviation from mid-parent value (ϕ), referred to as Mendelian sampling. That is

$$g_{individual} = 0.5g_{sire} + 0.5g_{dam} + \phi \qquad \text{Equation 8}$$

where the coefficients of 0.5 on the merit of the sire and dam can be justified by paired chromosomes only one member of each pair being passed on to the offspring. Consider a random mating unselected population exhibiting the same genetic variance from one generation to another, and in both sexes. That is,

$$\text{var}(g_{individual}) = \text{var}(g_{sire}) = \text{var}(g_{dam}).$$

It is a statistical fact that for a constant k, var(kX) = k^2 var(X). Applied to Equation 8, assuming random mating giving zero covariance between the sire and dam, and zero covariance between the sire or dam, and the Mendelian sampling term, the choice of sire contributes $0.5^2 = 0.25$ genetic variance, as does the choice of dam, implying that var(ϕ) = 0.5 in order for genetic variance to be maintained. If there were no variation due to Mendelian sampling, genetic variance would halve each generation and soon disappear.

The biological mechanisms for deviation from parent average are chance meiotic sampling of one member from each chromosome pair and creation of new chromosome combinations from crossing-over between existing homologous pairs.

The goal in predicting genetic merit of selection candidates would be to achieve high accuracy by puberty, or earlier. This cannot be achieved by pedigree-based performance recording, except for a few traits of high heritability measured cheaply before puberty.

The nature of breeding values and the role of molecular information

There are two other ways apart from Equation 8 to define a breeding value. It can be defined as twice the deviation in performance of offspring of the individual compared to random offspring. Finally, a breeding value can be described as the sum over all relevant loci of the average effects of the pair of alleles carried by an individual (Fisher, 1941). This requires knowledge of the loci responsible for variation in a trait, a means of identifying the specific alleles present at that locus, and known values for the average effects of all the alleles. In the presence of non-additive gene action, such as with dominance or epistasis, the average effects of alleles depend upon allele frequency. Selection, migration, mutation and drift can alter allele frequency and therefore average effects, so breeding values will vary between populations and over time.

Without records from relatives other than offspring, the elements of $\mathbf{A}^{-1}$ in the mixed model equations allow rearrangement to show estimated breeding values are a weighted function of the mean of the offspring phenotypes adjusted for their fixed effects and the merit of the mates (VanRaden & Wiggans, 1991). Regardless of heritability, the weight on the offspring deviation approaches two if there are many hundreds of offspring in large multi-sire contemporary groups. For lesser amounts of information, the weight on offspring is closer to zero, reflecting BLUP being a shrinkage estimator. The shrinkage for the same number of offspring is greater for low heritability traits, as deviations are less "believable" than when heritability is high.

Inspection of the mixed model equations in circumstances where the individual and its parents have records, but offspring have no observed phenotypes, demonstrates the sole contribution of relatives to the prediction of a young candidate is the mean of its parental estimated breeding values. This reiterates the total inability to predict from ancestors, the Mendelian sampling in Equation 8. Further, when the offspring does have its own record, its estimated breeding value is a linear function of its parent average and its own performance adjusted for fixed effects. The relative emphasis on the parent average compared to the individuals own performance depends upon the heritability. Low heritability traits provide little information to predict Mendelian sampling.

The mixed model equations and relationship matrix have their foundations in an infinitesimal model. As such they are a black box from a molecular point of view, neither requiring nor utilizing any knowledge of the number or location of individual genes. Supposing that one or more causal genes could be identified and their inheritance traced using molecular techniques, this information could be used to improve the prediction of estimated breeding values, particularly for young animals without individual phenotypic records. That thinking, along with an expectation that most traits might be influenced by one or a few major genes was the motivation for almost two decades of studies to detect quantitative trait loci (QTL) in livestock.

Those studies, investigating numerous traits in many species, identified a number of regions that influenced variation, but few major genes were found for traits influencing production. Some livestock programs were modified to account for such discoveries, but in most cases there is little evidence of improved annual rates of genetic gain. The QTL era would not be characterized as having delivered a milestone in livestock improvement,

although it did add to biological knowledge. The mechanism used to infer the presence of QTL from markers was based on the concept of linkage, whereby genomic regions with close physical location tend to be inherited as a unit.

There were two major problems with most QTL detection studies. First, genotyping was costly and time-consuming. The downstream effect was that most studies were too small to detect other than a few of the largest genes. Second, the marker density was insufficient to provide high levels of linkage disequilibrium (LD) across the entire genome. These experiments relied on genetic markers, and therefore LD to predict likely QTL genotype for a putative QTL at a particular genomic location. One method for creating LD was through linkage, by producing second generation intercross or back-cross individuals from disparate breeds or lines. Such studies would be well suited to detect QTL responsible for breed or line differences, but had less, if any, power to detect genes segregating in foundation breeds or lines (Spelman *et al*., 1998). An advantage of these dedicated studies was that they could be deeply phenotyped, but a disadvantage was that the populations had to be generated, incurring time delays, management and phenotyping costs in addition to the genotyping costs.

An alternative approach to avoid the time delay and minimize costs was to use characterized industry populations, such as widely-used parents with progeny information (Weller *et al*., 1990). This limited studies to traits with existing phenotypes, which had already been subjected to considerable selection pressure so that major genes were highly likely to have been selected towards fixation. Nevertheless, the approach generated some useful QTL, most notably in dairy cattle studies (Grisart *et al.*, 2002).

High-density genotyping and genomic prediction

Using markers to infer realized rather than average relationships in the context of mixed model methods was studied by Nejati-Javaremi *et al*. (1997). That approach relied on using every marker to infer relationships, or knowing chromosome regions that were informative for predicting merit. The idea that a large number of dense markers such as single nucleotide polymorphisms (SNP) could be simultaneously used to determine genomic regions responsible for variation in performance was popularized by Meuwissen *et al*. (2001). At that time, there was not enough information on genomic sequence in livestock species to create a set of SNP with genome-wide coverage, nor were there systems for simultaneously genotyping many SNP available for livestock applications. These circumstances have changed over the last few years as livestock species have been progressively genotyped, and reliable high-density whole-genome SNP platforms have been created. The first publicly-available Illumina SNP panel was for bovine, and has since been followed with swine, ovine and canine arrays. Panels have typically comprised 40k to 60k SNPs, and cost US$180 to US$300 per animal. Denser Illumina and Affymetrix bovine panels of >500k are expected to be released in 2010 or 2011.

Current status of genomic approaches

Regressing phenotype or breeding value on QTL genotype would estimate the additive effect of a QTL. However, since the location of QTL are unknown, nor are the QTL genotypes, this is not an option, but regression of phenotype or breeding value on SNP marker genotypes can be calculated. A SNP in perfect LD with one QTL would show a steeper regression than other SNPs with less LD with that QTL. Such regressions could be undertaken on every SNP. In the human area, regressions are typically carried out one SNP at a time, from a hypothesis-testing framework. In livestock, our interest is not so much in finding significant SNP, but in accurately predicting merit from available SNP, regardless of statistical significance.

It is not possible to simultaneously fit more fixed SNP effects than there are observations. Stepwise least squares procedures could be used to derive an informative SNP subset. However, least squares approaches are known to overestimate effects, particularly when the overall power of the experiment is low. Methods that shrink estimates tend to be more reliable. Two methods of shrinking estimates are to fit random rather than fixed regressions, or to fit mixture models that assume regression effects can come from different distributions, one of which might have zero effects. Both approaches are common in genomic predictions.

Regression of performance on all SNPs simultaneously can be achieved by fitting effects as random. That requires knowledge of the variance ratio appropriate for each SNP effect. One option is to fit the same variance ratio for every effect, known as ridge regression. That variance ratio could be estimated by various means. Fernando and Garrick (2009) refer to this method in a Bayesian context as Bayes C0. Another option is to fit a different variance ratio for each SNP, allowing estimates of some loci to be shrunk more than others. A Bayesian method for such an analysis was developed by Meuwissen *et al*. (2001) who called the method Bayes A. However, it is unlikely that every SNP would be in LD with a QTL, so a more appealing method might be one which allows some fraction of loci to have zero effect. Meuwissen *et al*. (2001) developed such a method in the context of a

mixture model, which they referred to as Bayes B. One problem with their method is that it required the mixture fraction (π) to be known. Kizilkaya *et al*. (2010) used a mixture model within the framework of ridge regression, describing that method as Bayes C. That method with π=0 is Bayes C0 or ridge regression. Fernando and Garrick (2009) extended it to simultaneously estimate the mixture fraction from the data, a method they refer to as Bayes Cπ.

Regressions tend to overfit the so-called training data used in the analysis; therefore quantifying the accuracy of the predictions cannot be done from that same analysis. Cross-validation where the training data is partitioned into subsets, one used for training and another used for validation is one method for quantifying accuracy. Another alternative is to validate using an independent dataset.

TABLE 1: Correlations between 50k genomic prediction and realized performance for validation of Angus sires in independent Angus datasets for backfat (FAT), calving ease direct (CED) and maternal (CEM), carcass marbling (MRB), carcass ribeye area (REA), scrotal circumference (SC), weaning weight direct (WWD) and yearling weight (YWT). Overall correlations are estimated by pooling the estimated variances and covariances from each separate validation.

Trait	Train 2 & 3 Predict 1	Train 1 & 3 Predict 2	Train 2 & 3 Predict 3	Overall
FAT	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
MRB	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

Simulated data with a finite number of QTL has demonstrated that models like Bayes B, which exploit the correct mixture fraction, result in more reliable predictions than Bayes A or Bayes C0 that fit very SNP. Estimating the mixture fraction from the data using Bayes Cπ gives better genomic predictions than using an improper mixture fraction. Simulation studies have shown correlations between genomic predictions and true underlying merit of 0.7 to 0.9 (Meuwissen *et al*., 2001), accounting for between 50% and 80% of total genetic variance.

Early whole genome analyses of the North American Holstein population (VanRaden *et al.*, 2009) showed that the parent average reliability (r^2) of animals without records or offspring averaged 0.19 across traits and that genomic prediction increased that value by 0.18 to 0.37. Livestock Improvement Corporation, an early adopter of genomic prediction in dairy cattle, found similar increases. In the international collaborative analysis of Brown Swiss performance, Jorjani *et al*. (2010) compared parent average predictions from conventional evaluations versus genomic evaluations from 50k SNP panels with the subsequent performance of progeny tested daughters four years later. Those analyses also demonstrated an increase in reliability by 0.18.

There are fewer published reports of genomic predictions in beef cattle. Analyses of USA Angus bulls based on published expected progeny differences (EPD) 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, so paternal half-sibs were not represented in more than one of the subsets. Pooled correlations between genomic and realized performance ranged from 0.5 to 0.7, accounting for between 25% and 50% of the genetic variance.

Habier *et al.* (2010) partitioned the German Holstein population into "training sets" to control the maximum pedigree-based additive genetic relationship between any bull in validation and all bulls in training. 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 exceeding 0.1249. Creation of these scenarios required that some bulls be excluded from both training and validation subsets. These scenarios had little impact on the average maximum relationship between animals in the training and validation subsets with the value remaining at about 9% for the first three scenarios. The correlations are shown in Table 2, for predictions based on 1,048 training bulls using conventional pedigree analysis (P-BLUP), and for methods using genomic relationship matrices with equal (Bayes C0) or heterogeneous SNP weighting (Bayes B). Genomic predictions outperformed pedigree-based methods, justifying their 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, and for any inferences to poorly related populations.

TABLE 2: Correlations 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.

Genomic prediction	a_{max}			
	0.65	0.49	0.249	0.1249
P-BLUP	0.51	0.51	0.45	0.21
Bayes C0	0.58	0.60	0.50	0.32
Bayes B	0.62	0.62	0.55	0.12

TABLE 3: Correlations between North American Holstein 50k genomic predictions from 1,000 or 4,000 training bulls born after 1994, and realized performance for ancestral bulls born before 1975.

Trait	Number of training bulls	
	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 analyses 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 are shown in Table 3 and fail to account for more than $0.52^2 = 28\%$ of the 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 poor predictive ability.

Predictive ability is seldom comparable in field data to levels that are readily achieved in simulated data. Poorer performance could be due to non-additive gene action as dominance or epistasis will alter the average effect of alleles as gene frequencies vary. Training populations may not have been large enough so predictions include spurious effects. Panels may not be large enough to ensure high LD for every QTL. Unlike simulated data, real data exhibits variation in LD in different genomic locations. This may result in good predictive ability in some genomic regions and poor predictive ability in others, limiting the overall predictive ability. Variation may result from structural variation such as deletions, insertions, inversions or copy number variants not well characterized by SNP genotypes. Epigenetic factors, such as caused by DNA methylation may sometimes be inherited, creating covariances between relatives that cannot be predicted from SNP analyses.

Validation of genomic predictions in other breeds provides a worst-case scenario in terms of predictive ability. 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 across-breed predictions, only LD, and that accounts for no more than 10% of genetic variance.

Results from simulated data whereby actual SNP genotypes were used to simulate phenotype by randomly selecting some SNP as QTL, demonstrated that for the same QTL in various beef cattle breeds, the across-breed predictive power was much poorer than within-breed, due to breed differences in the extent of LD between QTL and marker (Kizilkaya *et al.,* 2010). Marker panels that unrealistically included the causal polymorphisms, performed almost as well across breed as within breed.

The early adoption of genomic prediction in livestock has not been to improve accuracy on previously difficult to predict traits, but to improve the accuracy at a young age of the easily predicted traits. Further, the predictive ability is most reliable in offspring of the training animals, implying that for the near future, each seedstock population must have its own training analyses for every trait and that phenotypic data will need to be continually collected in order to provide for ongoing training in successive generations.

Likely future directions

Enormous potential remains for increasing the predictive ability from genomic data to levels closer to that obtained in simulated data. The solutions might involve the use of higher-density SNP panels such as 500k or 1m SNPs, the use of haplotypes, DNA sequence on widely-used parents and joint analysis that models both linkage and LD relationships. The cost-effective use of genomic technology may warrant cheaper lower density panels for screening non-parents, to impute genotypes of selection candidates for high-density markers based on high-density haplotype information on their parents and low-density information on the individuals (Habier *et al.*, 2009).

In the short-term it is difficult to design breeding schemes to optimally exploit genomic prediction because reliability of these methods beyond immediate offspring is not well characterized. In the medium- to long-term, market and economic issues that go beyond the predictive ability of the genomic evaluations will dictate the role of genomic prediction. The cost of genotyping panels may remain sensitive to demand due to shelf-

life and economies of manufacturing scale, creating challenges for sheep in comparison to bovine applications. The value of the technology in terms of livestock improvement not only has to exceed the costs, the benefits have also to be equitably shared along the production and technology chain, including ram and bull breeders and their servicing organisations. This may create challenges for companies that have invested in the genotyping and phenotyping of training populations simply to market genetic tests to industry. Return on investment could be more readily achieved in vertically-integrated organisations, such as Landcorp Farming Ltd., provided appropriate bioinformatics expertise is available to handle data storage, analysis and interpretation aspects that can be overwhelming compared to that required for mature pedigree and performance-based evaluation systems.

REFERENCES

Fernando, R.L.; Garrick, D.J. 2009: GenSel—User manual for a portfolio of genomic selection related analyses. Animal Breeding and Genetics, Iowa State University, Ames. Iowa, USA. http://taurus.ansci.iastate.edu/gensel Accessed May 2, 2010.

Fisher, R.A. 1941: Average excess and average effect of a gene substitution. *Annals of Eugenics* **11**: 53-63.

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. **41**: 92-102.

Garrick, D.J.; Harris, B.L.; Shannon, P.; Sosa-Ferreyra, C. 1993: A prototype sire evaluation for New Zealand dairy cattle. *Proceedings of the New Zealand Society of Animal Production* **53**: 91-94.

Garrick, D.J. 2010: Consequences of genomic prediction in cattle. *Proceedings of the Interbull Workshop. Genomic Information in Genetic Evaluations. 4-5 March, 2010. Paris, France*. Interbull Bulletin 41. http://www.interbull.org/index.php?option=com_content&view=article&id=78&Itemid=112 Accesssed 29 May 2010.

Gilmour, A.R.; Thompson, R.; Cullis, B.R. 1995: Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**: 1440-1450.

Gilmour, A.R.; Gogel, B.J.; Cullis, B.R.; Thompson, R. 2009: ASReml User Guide Release 3.0. VSN International Ltd., Hemel Hempstead, Hertfordshire, UK. www.vsni.co.uk

Grisart, B.; Coppieters, W.; Farnir, F.; Karim, L.; Ford, C.; Berzi, P.; Cambisano, N.; Mni, M.; Reid, S.; Simon, P.; Spelman, R.; Georges, M.; Snell, R. 2002: Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research* **12**: 222–231.

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

Habier, D.; Fernando R.L.; Dekkers, J.C.M. 2009: Genomic selection using low-density marker panels. *Genetics* **182**: 343-353.

Harris, B.L.; Clark, J.M.; Jackson, R.G. 1996: Across breed evaluation of dairy cattle. *Proceedings of the New Zealand Society of Animal Production* **56:** 12-15.

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

Henderson, C.R. 1963: Selection index and expected genetic advance, *In*: Statistical genetics and plant breeding. Hanson, W.D., Robinson, H.F. eds. National Academy of Sciences-National Research, Washington, D.C., USA. p. 141-163.

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

Henderson, C.R.; Kempthorne, O.; Searle, S.R.; von Krosigk, C.M. 1959: The estimation of environmental effects and genetic trends from records subject to culling. *Biometrics* **15**: 192-218.

Jorjani, H.; Zumbach, B.; Dürr, J.; Santus, E. 2010: Preliminary results from validation tests. *Proceedings of the Interbull Workshop. Genomic Information in Genetic Evaluations. 4-5 March, 2010. Paris, France*. Interbull Bulletin 41. http://www.interbull.org/index.php?option=com_content&view=article&id=78&Itemid=112 Accesssed 29 May 2010.

Kizilkaya, K; Fernando, R.L.; Garrick, D.J. 2010: Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *Journal of Animal Science* **88**: 544-551.

Lush, J.L. 1937: *Animal Breeding Plans*. Iowa State University Press, Ames, Iowa, USA. 443 pp.

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.

Meyer, K.; Garrick, D.J. 1995: Scope for a joint genetic evaluation of New Zealand and Australian Angus cattle. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* **11**: 250-253.

Nejati-Javaremi, A.; Smith, C; Gibson, J.P. 1997: Effect of total allelic relationship on accuracy of evaluation and response to selection. *Journal of Animal Science* **75**: 1738-1745.

Spelman, R.J.; Lopez-Villalobos, N.: Garrick, D.J. 1998: Experimental designs for quantitative trait loci detection in the New Zealand dairy industry. *Proceedings of the New Zealand Society of Animal Production* **58**: 6-9.

Quaas, R.L. 1976: Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* **32**: 949-953.

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. *Journal of Dairy Science* **92**: 16-24.

VanRaden, P.M.; Wiggans, G.R. 1991: Derivation, calculation, and use of national animal model information. *Journal of Dairy Science* **74**: 2737-2746.

Weller, J.L.; Kashi, Y.; Soller, M. 1990: Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. *Journal of Dairy Science* **73**: 2525-2537.