Genomic Selection in Farm Animal Species - Lessons Learnt and Future Perspectives

M.E. Goddard[*,2]*, B.J. Hayes*[2] *and T.H.E Meuwissen*[3]

Introduction

Selection based on phenotype and pedigree has been very successful in livestock despite lack of knowledge of the genes responsible for the genetic variation – we have practiced genetics without genes! However, a long term goal of research in livestock genetics has been to identify specific polymorphisms that contribute to variation in economically important traits. If successful this research should increase biological understanding of the traits and hopefully lead to faster genetic improvement of our livestock.

Each time technology revealed a new class of simply inherited polymorphisms or genetic markers, such as blood groups, they have been examined for any association with economic traits (Rendel 1961). There are three reasons why such an association might exist. Firstly, the polymorphism might cause an effect on the trait. Secondly, the marker might be in linkage disequilibrium (LD) with a polymorphism causing variation in the trait (ie with a QTL or quantitative trait locus). Thirdly the marker might be linked to the QTL but in linkage equilibrium with it so that the association exists only within families but not across the whole population.

The use of genetic markers in selection for economic traits is called marker assisted selection (MAS). The three reasons for association between markers and traits lead to three different types of MAS. Where a causal polymorphism is known, it has been relatively easy to incorporate it into selection decisions (Smith 1967) but this process has been limited by lack of knowledge of genes affecting most traits. Fernando and Grossman (1989) showed how markers linked to a QTL but in linkage equilibrium with it could be used. However, because the linkage phase between the marker and the QTL must be established for each family, this methodology has only really been used in dairy cattle in France and Germany (Boichard et al 2002).

Meuwissen et al (2001) proposed using dense markers covering the whole genome for MAS. The aim was to have sufficiently dense markers so that they were in LD with all QTL controlling the trait. This meant that all the genetic variance was tracked by the markers and that the effect of a marker on a trait could be estimated across the population, not just within a family. This form of MAS has become known as genomic selection. It is already being

[*] Melbourne School of Land and Environment, University of Melbourne, Victoria 3010, Australia
[2] BioSciences Research Division, Department of Primary Industries, Victoria, 3010, Australia
[3] IHA, Norwegian University of Life Sciences, Box 5003 As, Norway

widely used in dairy cattle and other industries are expected to follow shortly (VanRaden et al 2009, Harris et al 2008).

In this paper, we will describe the method of using markers to predict breeding value in genomic selection, the factors affecting the accuracy of this prediction, the effect of this technology on breeding programs, the effect on long term genetic progress and the prospects for the future.

Estimation of breeding value from marker data

The method of calculating estimated breeding values (EBVs) that maximises their accuracy and hence the response the selection is

$$EBV = E(BV \mid data)$$

where the 'data' is whatever information we have available. Traditionally the data has been phenotypes and pedigrees and the optimum method is selection index or BLUP. Now the data includes genotypes at marker loci but the principle is unchanged.

To implement this principle in selection indices or BLUP we require knowledge of the distribution of breeding values, usually assumed to be normal and so described by mean and (co)variances. Similarly, in genomic selection we need to know the distribution of the effects associated with markers. If this is a normal distribution of known variance we can use BLUP. However, it seems likely that the distribution is leptokurtotic with occasional markers of large effect (Hayes and Goddard 2001). Meuwissen et al (2001) modelled this with a t-distribution and called the method BayesA. It is also possible that many markers have no marginal effect once the other markers are included in the model. The method assuming a prior distribution that included a proportion of markers with zero effect and otherwise followed a t-distribution was called BayesB.

In simulated data the method of analysis that best reflects the model used for simulation is expected to give the highest accuracy of EBVs and this is the case (Meuwissen et al 2001). For many simulated datasets BayesB performs well. In real data there is often little difference in accuracy between different methods and a BLUP method, that assumes all marker effects are drawn from the same normal distribution, performs well (Hayes et al 2009a and b). This implies that there are many genes affecting each trait, all with small effects. However, in some traits such as fat % in milk, where the existence of genes of larger effect is known, BayesB outperforms BLUP.

Several other methods have been described in the literature. Generally the reason for these methods is to reduce the computing time needed for BayesB which is slow because it requires a reversible jump sampler step (equivalent to a Metropolis-Hastings step in this case) to sample between the zero or non-zero parts of the distribution. For instance, replacing the zero part of distribution with a normal or t-distribution of very small effects, allows Gibbs sampling instead of Metropolis-Hastings (Verbyla et al 2009). BayesB can also be implemented using a fast EM algorithm (Meuwissen et al 2009). Alternatively, an empirical non-linear regression can be used (VanRaden et al 2009) or a LASSO estimator (Usai et al 2009).

Methods that eliminate some markers from the model have the advantage that selection candidates can be genotyped for a smaller panel of markers and this may save costs.

An equivalent model

Instead of modelling the trait in terms of the effects of markers, it is possible to use a normal animal model where the breeding values are assumed to be normally distributed but the covariance among them is defined by a relationship matrix estimated from the markers. If properly specified, the marker model and the relationship model are equivalent and so give the same EBVs (Goddard 2009).

Accuracy of EBVs calculated from marker data

Two parameters control the accuracy of the EBVs calculated from marker data (GEBVs): the proportion of the genetic variance at QTL explained by the markers due to LD, and the accuracy with the marker effects are estimated (Goddard 2009).

The LD between markers and QTL, assuming the markers are randomly placed with respect to the QTL, depends on the density of markers. The expectation of the r^2 measure of LD between two loci c Morgans apart is $1/(1+4N_ec)$ where N_e is the effective population size assuming no selection, random mating, no mutation and constant N_e (Sved 1971). Thus if the number of markers is M and the length of the genome is L Morgans, the LD between adjacent markers depends on LN_e/M. In fact LN_e can be shown to be a critical parameter in determining accuracy of EBVs. It can be thought of as controlling the number of chromosome segments that are segregating in the population and whose effect on the trait must be estimated (Goddard 2009). If there are many independent segments (ie large N_eL) this has two consequences: more markers are needed so that each segment is tagged and the effects of typical segments is smaller and so more records are needed to estimate their effects accurately.

The importance of N_e can also be understood by the equivalent model based on relationships. The advantage of marker data over pedigree data can be thought of as the ability to estimate the realised relationship between two animals (ie the amount of the genome that they share) as opposed to the expected relationship derived from their pedigrees. The more realised relationships depart from pedigree relationships, the greater the accuracy that marker data gives to the EBVs. The variance of realised relationships about the pedigree relationship is equal to the LD r^2 averaged over all pairs of loci. In a random mating population, this average r^2 depends on N_eL. However, the markers do not exactly reflect the relationship between two animals at the QTL. If we do not know where QTL are located we estimate the relationship over the whole genome using all the markers. If there are M SNPs used as markers, the variance of the error in the estimated relationship is 1/M. Thus the ratio of the real variance in relationship to the error variance is dependent on M/N_eL. This ratio can be considered as a measure of the LD between all the QTL and all the markers. In general,

without assuming random mating, if there is a lot of variation in the relationship between pairs of animals, this increases the accuracy of genomic EBVs.

If the number of markers M is sufficiently high M/NeL is large and the nearly all the genetic variance is tracked by LD with the markers. However, it is still necessary to estimate the effect of each marker accurately. If the markers are close together, multiple markers may be in LD with the same QTL. In this case it is not the accuracy with which individual marker effects are estimated that is important, but the accuracy with which the total effect of a chromosome segment is estimated, regardless of the number of QTL or markers on the segment. If NeL is large, there are many effectively independent chromosome segments and so the effect of the average segment must be small. The variance of the effects of chromosome segments is proportional to h^2/N_eL. The effects will be estimated from a 'phenotype' y which we assume can be modelled as $y = \Sigma a + e$ where Σa is the sum of all the chromosome segment effects and e is an error. y may be a normal phenotype but in dairy bulls it is usually a daughter yield deviation or DYD. Let h^2 be the variance of Σa as a fraction of var(y). If y is a phenotype then h^2 is the heritability, if y is a DYD then h^2 is its reliability as an EBV. When the effect of any one chromosome segment is estimated, the other segments contribute to the error so the model is approximately $y = a + \varepsilon$ and $v(\varepsilon) \approx 1$. The accuracy of estimating the effects a depends on $Tv(a)/v(\varepsilon)$ where T is the number of records with phenotypes and genotypes. Since v(a) is proportional to h^2/N_eL and $v(\varepsilon)$ is nearly 1 (ie the complete phenotypic variance), the accuracy of estimating the effects a depends on Th^2/N_eL.

Thus there are two important parameters in determining the accuracy of EBVs – M/N_eL and $T\ h^2/N_eL$ (Meuwissen and Goddard in press). The other important parameter is the true distribution of marker effects. If the distribution is leptokurtotic then there are a few markers with large effects which are easier to estimate accurately than if the effects are normally distributed and consequently all quite small.

Simulation studies have confirmed that these three inputs determine the accuracy with one small exception: as h^2 increases towards 1.0 the accuracy increases even if $T\ h^2$ is kept constant. This occurs because in this case the residual variance $v(\varepsilon)$ is reduced by the estimates of all other a. Daetwyler et al (2009) describe a correction to the predicted accuracy to account for this.

The effect of T/N_e is important. It means that with populations of small N_e (eg 100), T = 2000 and h^2=0.8 can give good accuracy of EBVs even if the distribution of QTL effects is normal. However, if N_e>1000, then T needs to be very large. Similarly, if N_e=100 and L=30, then 12000 SNP markers is sufficient because the effective number of chromosome segments is <6000. However, when N_e>1000, ten times as many markers are needed.

Within breeds of livestock recent N_e is usually small (<200) although there are exceptions. However, when many breeds are economically important it may be difficult to obtain sufficient numbers of experimental animals within each breed especially for traits that are expensive to measure and of low heritability. It would be desirable to estimate a common prediction equation across breeds. We do not know exactly how to predict accuracy of

GEBVs in a multi-breed population but it seems likely that the appropriate Ne is much higher than that within breeds and may be larger than the Ne before breeds diverged eg >1000. This has implications for the number of markers needed and the number of animals used to estimate a prediction equation, both which must be increased. The need for increased density of markers can be understood by examining the density of markers needed to obtain a consistent LD phase between breeds. In B.taurus breeds this is about 1 SNP per 10kb or 300,000 SNPs per genome (DeRoos et al 2008). The importance of Th^2 explains why genomic selection is working better in Holstein dairy cattle than in beef cattle.

Distribution of QTL effects

It has proven difficult to estimate this distribution accurately. On the one hand GWAS demonstrate that there are QTL for almost any given trait distributed widely across the genome. The fact that a BLUP method of calculating GEBVs performs as well as methods such as BayesB for many traits argues for an almost infinitesimal model in which individual QTL effects are very small. On the other hand, there are some traits with polymorphism of moderate effect (eg DGAT and milk fat concentration) and for these traits methods like BayesB perform better then BLUP. Therefore it seems sensible to use prior distributions that at least admit the possibility of individual polymorphisms of moderate effect.

The effect of genomic selection on the design of breeding programs

As with other forms of MAS, the benefit from genomic EBVs depends on the proportion of genetic variance explained by the markers (which was discussed above) and the accuracy of EBVs available without markers (Meuwissen and Goddard 1996). If existing EBVs are of high accuracy at the earliest age at which reproduction is possible, then there is little to be gained. Consequently, traits which are displayed only in females, or late in life or post mortem or are expensive to measure, show the greatest benefits from genomic selection. For instance, dairy traits are displayed only in females and Schaeffer (2006) and Pyrce et al (these proceedings) conclude that genomic selection could double the rate of genetic gain. However, if traits such as growth rate are important in the breeding objective, then the gains from genomic selection are less. In practice, most breeding programs contain traits that are difficult to select for directly including female fertility, feed intake, and performance in commercial animals that may be reared in a different environment or are crossbred and/or castrated.

To make most advantage of GEBVs it may be necessary to reduce the generation interval eg by not waiting until males have been progeny tested. There is a natural synergy between GEBVs that reduce the age at which selection decisions can be made, and reproductive technology which reduces the age at which reproduction is possible. The combination of the two could dramatically reduce generation lengths and increase rate of genetic gain (Georges and Massey 1991, Hayley and Visscher 1997).

Implementation

Full implementation of genomic selection requires that industry level systems for calculating EBVs implement methods for using marker data. The first step is usually to calculate GEBVs on a small subset of the data with genotypes and to blend these with traditional EBVs from the main dataset. However, a combined analysis would be preferable. The easiest to implement is one based on relationships calculated from markers or, in ungenotyped animals, pedigrees. This naturally leads to a BLUP model for the SNP effects. We do not believe this is optimal. It would be better to use a prior that allowed for some moderate effects and for many SNPs to have zero effect. As more animals get genotyped for SNPs it will be logical to change to completely 'gene based' models and do away with our traditional model based on relationships entirely (Goddard 1998).

Long term genetic gain

The accuracy of GEBVs predicts the genetic gain in the next generation but will this rate of gain be maintained? Simulation studies show that the rate of genetic gain rapidly declines unless the prediction equation is continuously re-estimated each generation (Muir 2007; Sonesson and Meuwissen 2009). There are at least two explanations for this. Firstly, selection increases the frequency of the marker allele more than the favourable QTL allele if the LD between them is not perfect. This decreases the LD and hence decreases the proportion of QTL variance explained by the marker. Secondly, genomic selection does not trace QTL with a rare favourable allele as well as it does when the favourable allele is at an intermediate frequency. In the short term this is not too important because the QTL at intermediate frequency provide the genetic gain, but as they approach fixation the rate of genetic gain slows unless rare favourable alleles are being selected (Goddard 2009).

If QTL were assayed directly instead of linked markers, then the first reason for declining genetic gain would be eliminated. However, we would still need a mechanism to increase the frequency of rare favourable alleles. This might be helped by use of markers with low MAF or use of haplotypes or identification of causal polymorphisms. However, it might still be necessary to employ some selection directly on phenotype to drive rare favourable alleles to high enough frequency that they are picked up by genomic selection.

If the QTL can be identified, then Goddard (2009) shows how selection can maximise long term gain by placing additional selection pressure on initially rare favourable alleles.

The Future

The need to base prediction equations on very large numbers of animals with phenotypes and genotypes is beginning to influence planning for the future. This realisation should foster collaboration between breeders and companies who are competitors. Few countries or companies will be able to use so large a population that they could not benefit from enlarging it through collaboration. Hopefully this will lead to a return to the international collaboration in EBV calculation that has benefitted everyone in the past.

High density SNP chips will soon be available for cattle and hopefully other species. This will improve the accuracy of GEBVs especially in mixed breed populations. In the near future we will move to the use of full genome sequence data on individual animals by using a combination of physical sequencing and imputing sequence from SNP genotyping (Goddard 2008; Goddard and Hayes 2009). The main advantage of sequence data will be the ability to use the causal polymorphisms as the markers on which GEBVs are based. As well as an increase in accuracy of GEBVs (Meuwissen and Goddard in press), this will lead to a surge of new knowledge about the biology of economic traits. To identify causal polymorphisms it will be necessary to combine analysis of phenotype and genotype data with outside information about which polymorphisms are likely to have an effect on the phenotype.

Conclusions

Genomic selection will radically change genetic improvement of livestock. Potentially rates of genetic change can be greatly increased. However, it is difficult to obtain highly accurate EBVs from marker data alone. To achieve high accuracy we need very large reference datasets, high density SNPs, and analytical methods that use a distribution of QTL effects that matches the real world and allows some QTL of moderate effect. We still do not have a method that will calculate the accuracy of the GEBVs produced in a proven and practical way. In the future, we will use genome sequence data to identify causal mutants and the resulting surge in biological knowledge will suggest novel designs of breeding program. Combining genomic selection with reproductive technology will lead to short generation intervals and rapid genetic gain. However, we do not know how to maintain this rapid gain in the longer term.

References

Boichard, D., S. Fritz, M. N. Rossignol, M. Y. Boscher, A. Malafosse& Colleau. J. J. (2002). *Proc. 7th World Cong. Genet. Appl. Livst. Prod.* 33: 19-22.

De Roos, A. P. W., Hayes, B. J., Spelman, R. and Goddard, M. E. (2008). *Genetics.* **179**, 1503-12.

Daetwyler, H. D., B. Villanueva and J. A. Woolliams, (2008). *PLoS ONE* 3(10): e3395.

Fernando, R. & Grossman, M. (1989) *Genet. Sel. Evol.* **21,** 467-477 (1989).

Georges, M. and Massey, J.M. (1991) *Theriogen.* 35: 151-159.

Goddard M.E. (1998) *Proc 6th Wrld Congr. Genet. Appl. Livest. Prod.* **26** : 33-36.

Goddard M.E. (2008) In Pinard, M-H., Gay, C., Pastoret, P-P. and Dodet, B. (eds): Animal Genomics for Animal Health. Dev Biol. (Basel) , Basel, Karger volume 132: pages 383-9.

Goddard, M. E. (2009). *Genetica* 136: 245-257.

Goddard, M.E. and Hayes, B.J. (2009) *Proc. Assoc. Advmt Anim. Breed. Genet.* 18: 26-29

Harris, B.L., Johnson, D.L. and Spelman, R.J. (2008) *Proceedings of the Interbull Meeting, Niagara Falls, Canada* 2008.

Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K. and Goddard, M.E. (2009a) *Genetics Selection Evolution* 2009, **41**:51

Hayes, B.J., Daetwyler, H.D., Bowman, P., Moser, G., Tier, B., Crump, R.,. Khatkar, M., Raadsma, H.W. and Goddard, M.E. (2009b) *Proc. Assoc. Advmt Anim. Breed. Genet.* 18: 34-27

Hayes, B. J. and Goddard, M.E. (2001) *Genet. Sel. Evol*. **33**, 209-229.

Hayes, B. J. Visscher, P. M., McPartlan, H. and Goddard, M. E. (2003) *Genome Res.* **13**, 635-643 (2003).

Hayley C.S. and Visscher, P.M. (1998) *J. Dairy Sci*. 81: 85-97.

Meuwissen, T.H.E. and Goddard, M.E. (1996) *Gen. Sel. Evol*. **28**: 161-176.

Meuwissen, T. H .E., Hayes, B. J. and Goddard, M. E. (2001) *Genetics* **157**,1819-1829.

Meuwissen, T.H.E., Solberg, T.R., Shephard, R. and Wooliams, J.A. (2009) *Gen. Sel. Evol.* 41:2.

Muir, W. M., (2007). *J. Anim. Breed. Genet*. 124: 342–355.

Rendel, J. (1961) Relationships between blood groups and the fat percentage of the milk in cattle. *Nat. (Lond)* **189**, 408–409.

Schaeffer, L R (2006) *J. Anim. Breed. Genet*. **123**, 218-223.

Smith, C. (1967) *Anim. Prod.* **9**, 349-358.

Sonesson, A.K. and Meuwissen, T.H.E (2009). *Gen. Sel. Evol*. 41: 37.

Sved, J.A. (1971) *Theoretical Population Biology* **2,** 125-141.

Usai, M.G., Goddard, M.E. and Hayes, B.J. (2009) *Genetics Research* 91: 427-436.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. and Schenkel, F. S. (2009) *J Dairy Sci* **92**, 16-24.

Verbyla, K.L., Hayes, B.J., Bowman, P,J, and Goddard, M.E. (2009) *Genet. Res.* 91:307-311.