Genet. Res., Camb. (2010), **92**, *pp.* 413–421. © Cambridge University Press 2011
doi:10.1017/S0016672310000613

Genomic selection in livestock populations

MICHAEL E. GODDARD[1,2]*, BEN J. HAYES[2] AND THEO H. E. MEUWISSEN[3]
[1] *Department of Agriculture and Food Systems, University of Melbourne, Parkville 3010, Australia*
[2] *Victorian Department of Primary Industries, 1 Park Drive, Bundoora, Australia*
[3] *University of Life Sciences, P.O. Box 5003, 1432 Ås, Norway*

(*Received 13 October 2010 and in revised form 17 November 2010*)

Summary

Most traits of economic importance in livestock are either quantitative or complex. Despite considerable efforts, there has been only limited success in identifying the polymorphisms that cause variation in these traits. Nevertheless, selection based on estimated breeding values (BVs), calculated from data on phenotypic performance and pedigree has been very successful. Genomic tools, such as single nucleotide polymorphism (SNP) chips, have led to a new method of selection called 'genomic selection' in which dense SNP genotypes covering the genome are used to predict the BV. In this review we consider the statistical methodology for estimating BVs from SNP data, factors affecting the accuracy, the long-term response to genomic selection and the design of breeding programmes including the management of inbreeding.

1. Introduction

The objectives of genetic improvement of livestock are usually quantitative or complex traits such as milk yield or meat quality. Traditional genetic improvement has relied on using the recorded phenotype of each animal together with the knowledge of its pedigree to estimate its breeding value (BV), most often using the statistical method, known as best linear unbiased selection (BLUP) (Henderson, 1984). This technology has been very successful, leading to genetic gains in most farmed species (e.g. see Van Vleck *et al.*, 1986; Havenstein *et al.*, 1994). Despite this success, there has long been an interest in using simply inherited genetic markers to increase the rate of genetic gain and to identify the genes and polymorphisms controlling traits in the breeding objectives (as summarized in Dekkers & Hospital, 2002).

Ideally one would identify causal polymorphisms affecting an objective trait and incorporate these in the selection criterion (Dekkers, 2004). This has occurred for some mutations that cause genetic abnormalities and a small number of polymorphisms with large effects on quantitative traits (Dekkers, 2004). However, these known causal polymorphisms explain only a small proportion of genetic variance in the breeding objective and have contributed only a small amount to the genetic gain achieved. This approach has been limited by our inability to identify most of the causal polymorphisms affecting our objective traits.

As new categories of genetic markers were discovered they have been tested for an association with quantitative traits, even though there was no *a priori* reason to expect an association. For instance, bovine blood groups were sometimes found to be associated with milk production traits (Neimann-Sorensen & Robertson, 1961; Rendel, 1961). It is possible that this association was causal but more likely that it was due to linkage between the blood group loci and quantitative trait loci (QTL) that cause variation in milk production. These associations proved too weak and too unreliable to be useful in the selection of livestock.

Microsatellites were the first class of genetic markers that covered the genome and therefore, had the possibility to detect QTL no matter where they were located. Typically 100–200 microsatellites were used to cover the genome and they detected QTL by

* Corresponding author: Department of Agriculture and Food Systems, University of Melbourne, Parkville 3010, Australia. e-mail: mike.goddard@dpi.viv.gov.au

linkage within full-sib or half-sib families (Georges *et al*., 1995). The limitations of these studies were that they mapped the QTL very imprecisely (often to confidence intervals of 50 cM) and the marker and QTL were in linkage equilibrium so that the linkage phase varied between families. Consequently the linkage phase had to be determined within each family before the marker could be used for selection. Fernando & Grossman (1989) presented a general method for estimating BVs using markers in linkage equilibrium with QTL but, in practice, the gains were small and this method of marker assisted selection has only been used rarely (but for an exception see Boichard *et al*., 2006). By saturating a QTL region with additional markers, the causal mutation has occasionally been discovered (Grisart *et al*., 2003), but only when it explained an unusually large proportion of genetic variance.

The QTL mapping studies showed that many QTL affect a typical quantitative trait (Hayes & Goddard, 2001; Chamberlain *et al*., 2007). Meuwissen & Goddard (1996) showed that the gain in selection response from marker assisted selection was nearly proportional to the proportion of genetic variance explained by the markers. Therefore, a new type of marker assisted selection was needed that utilized all QTL and that did not require linkage phase to be determined for each family. Meuwissen *et al*. (2001) showed with simulation that using a dense panel of markers covering the whole genome and in linkage disequilibrium (LD) with the QTL could lead to large increases in response to selection. This type of marker assisted selection has become known as genomic selection. It became feasible with the availability of panels of thousands of single nucleotide polymorphisms (SNPs) that could be genotyped at reasonable cost. It is already widely used in dairy cattle breeding (Dalton, 2009) and is expected to revolutionize all livestock genetic improvement programmes and can be extended to plants (Bernardo & Yu, 2007; Heffner *et al*., 2009; Zhong *et al*., 2009), aquaculture (Sonesson & Meuwissen, 2009) and prediction of genetic risk in humans (Wray *et al*., 2007). In this review, we will describe the methodology used, the factors determining the accuracy of selection, the implementation in breeding programmes, the effect on long-term genetic gain and the use of genomic selection for QTL mapping.

2. Methodology

The BV (bv) or additive genetic value of an individual j can be written as $bv_j = \sum_{i=1}^{N_q} x_{ij} a_i$ where a_i is the additive effect of the ith QTL and x_{ij} is the genotype of the individual at the ith QTL coded as 0, 1 or 2 for homozygote, heterozygote and other homozygote respectively, and N_q is the number of QTL. In practice the QTL position and effects are not known. Instead we detect the QTL by their LD with markers such as SNPs. If there is sufficient LD, the genotype at a QTL, x_i, can be predicted from a linear combination of marker genotypes and so BVs can be estimated by a linear combination of markers $\tilde{bv}_j = \sum_{i=1}^{N_m} m_{ij} b_i$, where b_i is the apparent effect of the ith marker due to its LD with one or more QTL, m_{ij} is the genotype of the jth individual at the ith marker and N_m is the number of markers. However, b_i has to be estimated from data and so the estimated breeding value (EBV) for individual j becomes $\widehat{bv}_j = \sum_{i=1}^{N_m} m_{ij} \hat{b}_i$.

Selection theory shows that an EBV is most accurate if $\widehat{bv} = E\,(bv|\text{data})$ where data includes whatever information is available from which to estimate the BV. Here this means that the vector of marker effects $\mathbf{b}$ should be estimated as $\hat{\mathbf{b}} = E(\mathbf{b}|\text{data})$. The data ($\mathbf{y}$) usually consists of a reference sample of the population that has been measured for the trait and genotyped for the markers. Assuming the data ($\mathbf{y}$) have been corrected for all other effects, then, as presented in Goddard (2009),

$$\hat{\mathbf{b}} = E(\mathbf{b}|\text{data}) = \int \mathbf{b} p(\mathbf{y}|\mathbf{b}) P(\mathbf{b}) \mathrm{d}b \Big/ \int p(\mathbf{y}|\mathbf{b}) p(\mathbf{b}) \mathrm{d}b, \tag{1}$$

where $p(\mathbf{b})$ is the prior distribution of $\mathbf{b}$, and $\mathrm{P}(\mathbf{y}|\mathbf{b})$ is the likelihood of the data given $\mathbf{b}$. This shows that the best estimate of $\mathbf{b}$ depends on the distribution of $\mathbf{b}$. If $\mathbf{b}$ follows a normal distribution with the same variance for all markers $\mathbf{b} \sim N(0, \mathbf{I}\sigma^2_{\mathrm{b}})$ then (1) reduces to a BLUP estimate of $\mathbf{b}$. Since 10 000–1 000 000 SNPs may be used, this assumption implies that all SNPs have very small effects and this is akin to the traditional infinitesimal model for quantitative traits.

Other assumptions for the distribution of $\mathbf{b}$ do not lead to closed form solutions for $\hat{\mathbf{b}}$ but $\hat{\mathbf{b}}$ can be calculated by Markov Chain Monte Carlo (MCMC) methods. For instance, Meuwissen *et al*. (2001) considered the case where the marker effects are assumed to follow a scaled t distribution. Marker effects of large size are more probable under a t distribution with a small number of degrees of freedom than under a normal distribution (i.e. the t distribution has 'thicker tails' or greater kurtosis than a normal distribution). This assumption might more correctly reflect the true situation than assuming that marker effects follow a normal distribution since some polymorphisms with large effects on quantitative traits are known. Meuwissen *et al*. (2001) called this model of marker effects 'Bayes A' and showed how Gibbs sampling could be used to estimate the marker effects and hence the BV of individuals. Although it allows for some markers with large effects, the Bayes A model still assumes that all markers have a non-zero effect. If the number of QTL is much smaller than the number of markers, one might expect that many markers have

no effect after those in higher LD with the QTL have already been included in the model. Therefore, Meuwissen *et al*. (2001) introduced a model (that they called 'Bayes B') in which a proportion of the marker effects follow a scaled t distribution but the remainder of markers have no effect. Bayes B was implemented using a combination of Gibbs sampling and Metropolis–Hasting steps to estimate marker effects and hence individual's BVs.

As usual, the BLUP estimate of a marker effect can be interpreted as a least squares estimate that has been shrunk or regressed towards zero. This is also the case for estimates of marker effects under Bayes A and B models, but they shrink the estimates in a non-linear manner so that a least squares estimate that is small relative to its standard error is shrunk almost to zero, while estimates that are large are shrunk less severely. Other prior distributions of **b** including the double exponential also were considered (Yi & Xu, 2008) and Meuwissen *et al*. (2009) gives a closed form solution for this.

For the case where **b** is normally distributed there is an equivalent model that is informative (Habier *et al*., 2007; VanRaden 2008; Hayes & Goddard, 2008). Using the matrix notation, if $\mathbf{y}=\mathbf{Mb}+\mathbf{e}$ and $\mathbf{bv}=\mathbf{Mb}$ then $\mathbf{y}=\mathbf{bv}+\mathbf{e}$ with $\mathbf{bv} \sim N(0, \mathbf{MM}'\sigma_g^2)$, where **M** is the matrix of marker genotypes with elements m_{ij} defined above. This is a conventional animal model where **y** is the sum of a **bv** and environmental error (**e**) but where the relationships among the individuals are estimated as **MM**′. Thus estimating the BV of an individual by adding the effects of all markers carried (sometimes called a SNPBLUP model) is equivalent to estimating the BV using the realized relationship among the individuals estimated from the markers (sometimes called a GBLUP). If a set of unphenotyped individuals are all equally related, they will all receive the same EBVs and so the correlation between the true BV and EBV for this set of individuals is zero. This shows the importance of variation in relationships between pairs of individuals – it is this variation that provides power to estimate BV from marker genotypes.

The best method for estimating the relationships is a slight modification of **MM**′ set out in Yang *et al*. (2010). Their method has the lowest standard error for estimated relationships when the true relationships are small.

3. The accuracy of genomic selection

For individuals with marker genotypes but without phenotypic records we can calculate their EBV simply as $\widehat{bv}_j=\sum_{i=1}^{N_m} m_{ij}\,\hat{b}_i$. The accuracy of this EBV depends on two factors – the proportion of variance in the QTL explained by the markers due to LD, and the accuracy with which the **b** are estimated (Goddard, 2009).

(i) *The proportion of variance in the QTL explained by the markers*

The first of these factors can be quantified by the accuracy with which the relationships between individuals are estimated by the markers. Consider an infinitesimal model where there are an infinite number of QTL spread evenly over the chromosomes. If the markers are a random subset of these QTL they will estimate the relationship at the QTL except for a sampling error caused by the finite number of markers. The variance of the difference between the estimated relationship and the true relationship of a pair of individuals is called the prediction error variance (PEV). The PEV of the relationship between individuals i and j (G_{ij}) caused by the finite number of markers (N_m) is PEV (G_{ij}) $=1/N_m$ (Yang *et al*., 2010). The degree by which this error degrades the estimate of the true relationship depends on the true variation in relationship. If pairs of individuals vary widely in relationship then a small error may be unimportant but if the true variation is similar to the PEV then this error will severely affect the accuracy of the estimated relationship and hence, the EBVs. If individuals vary in pedigree relationship (e.g. some are closely related and some are not) then the variation in true relationship will be great and the markers should be able to estimate these differences relatively easily. However, in that case EBVs could be calculated from the pedigree information without genetic markers. The real power of genomic selection is to estimate BV more accurately than could be done using pedigree data. Therefore, it is the variation in G_{ij} in excess of that due to variation in pedigree that is important.

Hayes *et al*. (2009*b*) showed how variation in realized relationship occurs among individuals with the same pedigree, such as a group of full sibs. Among pairs of full sibs, some pairs share more than 50% of their DNA and some share less than 50%. This variation around 50% only exists because genes on the same chromosome are linked and so not inherited independently – if there were an infinite number of unlinked genes, all pairs of full sibs would share 50% of their autosomes. The variation about 50% combined with phenotypes on a group of full sibs, allows us to estimate the BV of an additional full sib from the same family, even one with no recorded phenotype (Hayes *et al*., 2009*b*). The estimation of BV of an additional full sib is possible because the new individual is more closely related to some of its full sibs than to others. Full sibs inherit large segments of chromosome from their parents without recombination, so whole segments of chromosome are either shared or not shared between a pair of full sibs. Use of the relationships in this way to estimate individual's BV is equivalent to estimating the effect of

https://doi.org/10.1017/S0016672310000613 Published online by Cambridge University Press

chromosome segments on the trait and using these estimates to predict the BV of the additional full sib.

In a random mating population there is some variation in pedigree relationship and additional variation in realized relationship. The variance of the relationship around the pedigree relationship is approximately $\log(2N_eL)/(2N_eLc)$ where N_e is the effective population size, L is the average length of a chromosome in Morgans and c is the number of haploid chromosomes. If two individuals share a common ancestor, there is a probability, defined by their relationship that they both inherit an allele identical by descent (IBD) from this common ancestor. If they inherit a common allele at one locus they will also inherit common alleles at neighbouring loci due to linkage. On average, the length of this IBD segment will decrease the more distant the common ancestor is. The average time to a common ancestor increases as N_e increases, so the length of chromosome segments shared IBD decreases as N_e increases. Thus, in a population of large N_e, individuals share many small chromosome segments and the realized relationship averages out to close to the relationship expected from the pedigree. This explains the occurrence of N_e in the formula for the variance of relationship about the pedigree relationship. Therefore, for large genomes and populations with large N_e the variance of true relationship is very small and so the PEV must be small if the relationships are to be estimated with precision and this implies that a large number of markers are needed.

Using the model based on SNP effects, $\mathbf{y}=\mathbf{Mb}+\mathbf{e}$, it is possible to estimate the total genetic variance explained by the SNPs. The same answer can be achieved by using the equivalent model based on relationships estimated from the SNPs (Yang *et al*., 2010). In either case the variance estimated will be less than the total genetic variance if the QTL are not in perfect LD with the SNPs or, equivalently, if the estimated relationship is not an unbiased estimate of the relationship at the QTL. In cattle breeds such as Holsteins, the recent N_e has been small (~ 100) so the variation in relationship is large and 50 000 SNPs can estimate the relationships well and so the genetic variance explained by the SNPs is close to the full genetic variance (VanRaden *et al*., 2009). This is equivalent to saying that the QTL genotypes can be predicted by the SNP genotypes due to LD between them. However, in humans recent N_e has been very large and so the variance of true relationships is small and even with 600,000 SNPs the PEV is significant and results in the genetic variance explained by the SNPs being only about half the known genetic variance (Yang *et al*., 2010). This is due partly to the use of a finite number of SNPs to estimate the relationship but also to systematic differences between the SNPs and QTL. If QTL and SNPs have different evolutionary histories, there may be systematic differences in the relationships at QTL and at SNPs. For instance, if QTL mutant alleles are typically eliminated by selection, they will tend to be young, and so ancient relationships estimated from the SNPs may not be relevant. An equivalent description of this situation is that QTL will have low minor allele frequency (MAF) and so cannot be in high LD with SNPs that have higher MAF. Consequently, the SNPs will explain less of the genetic variance of a trait than expected simply by accounting for the PEV due to a finite number of SNPs. Yang *et al*. (2010) found that SNPs only explained about half the genetic variance for human height but they should have explained 80 % if the QTL had behaved like SNPs.

Since the true variance in relationships is dependant on N_eLc and the PEV with which it is estimated is $1/N_m$, it is not surprising that the accuracy of EBVs depends on $N_m/(N_eLc)$ (Meuwissen, 2009; Meuwissen & Goddard, 2010).

In the argument above we assumed an infinite number of QTL. However, BLUP estimates of EBVs are insensitive to the true genetic model for the trait and give a similar accuracy regardless of the actual number of QTL governing variation in the trait (Meuwissen & Goddard, 2010).

(ii) *The accuracy with which the marker effects are estimated*

If the BLUP model is used to estimate SNP effects, the standard theory provides estimates of the accuracy of the estimated SNP effects and the EBVs of individuals provided that the correct variance components are used. To the extent that the SNPs do not explain all of the genetic variance, an additional 'polygenic' term ($\mathbf{u}$) should be included in the statistical model with $V(\mathbf{u})=\mathbf{A}\sigma^2_u$ where $\mathbf{A}$ is the relationship matrix constructed from pedigree information and σ^2_u is the genetic variance not explained by the SNPs (Hayes *et al*., 2009*a*). In the equivalent model based on realized relationships the G matrix should be estimated by regressing $\mathbf{MM}'$ back towards $\mathbf{A}$ to account for the error in the relationships estimated by $\mathbf{MM}'$.

If all SNPs were independent (i.e. no LD) then the accuracy of estimating any one SNP effect is approximately $\sqrt{n/(n+\lambda)}$ where n is the number of animals with genotypes and phenotypes and $\lambda=\sigma^2/\sigma^2_b$, where σ^2 is the phenotypic variance, that is, the variance of y. However, the LD between SNPs and QTL located close together on a chromosome causes a segment of chromosome to act almost as a block and the accuracy of estimating the effect of the block is given by the above formula but with $\lambda=s\sigma^2/\sigma_g{}^2$, where s is the effective number of chromosome segments. The best value to use for s has not been fully resolved but is approximately $2N_eLc/\log(2N_eL)$ (Hayes *et al*.,

https://doi.org/10.1017/S0016672310000613 Published online by Cambridge University Press

2009*b*). The accuracy of estimating a single SNP effect $\sqrt{n/(n+\lambda)}$ also equals the accuracy of estimating that part of the BV that is predicted by the SNPs, which is the sum of many SNP effects.

If the SNP effects (**b**) do not follow a normal distribution, the accuracy achieved using the BLUP model is relatively unaffected. However, greater accuracy can be achieved by a statistical method whose assumption about the distribution of **b** more closely approximates the true distribution. For instance, if a trait is controlled by a small number of QTL, some of which have a moderately large effect, then Bayes B yields higher accuracy than the BLUP analysis (Verbyla *et al*., 2009). This is not surprising because Bayes B assumes that many of the SNP effects are zero and the remainder follow a scaled *t*-distribution which allows for some larger than normal effects. However, whether a BLUP or Bayes B analysis is used, λ is still a key parameter in determining the accuracy (Meuwissen & Goddard, 2010).

Unfortunately, we do not know the true distribution of apparent SNP effects but for some traits there are clearly a small number of QTL with effects that are larger than would be sampled from a normal distribution. Also it seems likely that as the number of SNPs used increases the assumption that many have zero effect is more likely to be true. Therefore, Bayes B seems to be widely useful – it seldom performs worse than BLUP and sometimes is significantly better (see experimental results discussed later in this paper).

Many empirical methods have been tried for predicting BV from SNP genotypes (e.g. Gianola *et al*., 2009; Moser *et al*., 2009). In most cases to date many methods give similar accuracy. However, it seems logical to attempt to use an explicit assumption about the distribution of **b** and to make this assumption as close to reality as possible. Most of these methods imply an additive model of QTL effects. This seems appropriate when the aim is to estimate BV because this is by definition a linear combination of QTL effects. However, if the aim was to estimate total genetic value a model assuming non-additive genetic effects might be better. Non-additive effects can be included in the model explicitly or a non-parametric or semi-parametric method such as kernel regression may be used (Gianola *et al*., 2006). Lee *et al*. (2008) showed that mouse colour could be predicted better by including dominance in the model but the difficulty with such non-additive models is likely to be the inability to estimate numerous small effects that typically explain a small amount of the variance (Hill *et al*., 2008).

The methods to estimate BV from marker genotypes presented in this paper have a natural Bayesian interpretation which includes a prior distribution of marker effects. However, very similar methods can be derived from non-Bayesian perspectives. For instance, they can also be derived as the expected value of BV in a frequentist setting where marker effects are regarded as random samples from a population of random effects. Penalized least squares and other machine learning methods can also yield similar results (Moser *et al*., 2009).

(iii) *Experimental results*

The accuracy of genomic prediction of BV has been assessed by estimating a prediction equation using one dataset and then testing the prediction in a second independent dataset. When this has been done the results are qualitatively in line with the theory above. For instance Wiggans *et al*. (2010) observed the accuracy to increase from 0·80 to 0·84 as the number of records (n) used increased from 3700 to 7173.

In many respects, the conditions examined by VanRaden *et al*. (2009) and Wiggans *et al*. (2010) are the most favourable. All the animals were within one breed of cattle (Holstein) and the recent effective population size of this breed is low ($\sim$100). This means that the variation in true relationship is large or equivalently that the LD between SNPs and QTL is high, so that the approximately 40 000 SNP explain most of the genetic variance. Also, the low N_e means that the effective number of chromosome segments (s) is small and so the accuracy of estimating their effects is high. This is further aided by the use of progeny tested sires as the experimental animals since they have relatively accurate estimates of BV and so the residual error in the data (σ_e^2) is low.

Experiments in other livestock have not yielded such high accuracy. For instance, in sheep breeds the accuracy achieved has been lower than reported in Holsteins (Daetwyler *et al*., 2010). This is expected because the number of animals with marker genotypes and phenotypes is smaller, these animals belong to multiple breeds and the phenotypic data consists of individual animal phenotypes instead of the progeny test used in the Holstein case. In humans, where recent N_e is very large, the accuracy of predicting phenotype has been low despite large datasets (Manolio *et al*., 2009). However, formal prediction methods such as Bayes B have not been attempted.

As the number of records increases in other breeds and species the accuracy of the EBVs is expected to increase as it did in Holsteins. However, for many breeds and species it may not be possible to assemble such high quality datasets as has been done for Holsteins. In these cases it would be desirable to combine data from several breeds within a species. This is only beneficial if the phase of LD between SNPs and QTL is the same in different breeds. This is not the case when 50 000 SNPs are used in cattle breeds (de Roos *et al*., 2008) but consistent LD phase should occur if denser SNPs are used (e.g. 500 000).

Unfortunately, when multiple breeds are used, the effective number of chromosome segments (s) increases, implying that even larger datasets are needed. Therefore, we expect that the increased accuracy achieved from high density SNP panels will be greater if methods such as Bayes B, which assume many SNPs have zero effect, are used.

Methods such as Bayes B do yield higher accuracy of EBV than BLUP in traits with segregating QTL of moderate effect (Hayes *et al*., 2010). For instance, EBVs for fat concentration in milk and proportion of white colour in the coat of Holstein cattle where more accurate when Bayes B was used than when BLUP was used and there are known genes segregating which effect these traits (Hayes *et al*., 2010).

4. Implementation of genomic selection

In most livestock breeds there are systems in place to calculate EBVs from traditional phenotypic records and pedigrees. In dairy cattle these operate at a national and international level. The use of DNA data to increase the accuracy of EBVs needs to be integrated into these existing systems. At present most systems have very large databases of traditional phenotypic records (from millions of animals) and comparatively small databases of SNP genotypes. Consequently strategies have been devised to minimize the additional computing load in the analysis of the large database. For instance, the SNP genotypes can be combined in a prediction equation to yield an estimate of BV coming only from SNPs. This has been called a direct genetic value (DGV) or marker breeding value (MBV). A selection index is then used to combine this estimate with that generated by the traditional analysis, which does not use SNPs at all, resulting in a final published EBV (Harris & Johnson, 2010). Alternatively, the DGV can be treated as an additional trait, genetically correlated with the phenotypic trait, in a multi-trait BLUP analysis of the large dataset. This has the advantage that it propagates the DGV to the relatives of an animal with SNP genotypes but at the cost of increased computing burden. A third method is to use the equivalent model based on relationships. For animals with SNP genotypes these are used to calculate the relationship and for other animals the pedigree relationship is adjusted for the knowledge contained in the relationships based on genotypes (Legarra & Misztal, 2008). This method requires raw genotypes rather than DGVs and it is most useful if the BLUP method of estimating SNP effects is to be used. However, other methods such as Bayes B could be used by calculating a relationship matrix from the SNPs but weighting the SNPs according to the variance associated with them. If, in the future, large numbers of animals have SNP genotypes, it may be that the genetic evaluation system will be completely changed to one that uses genetic markers rather than pedigree relationships, as suggested by Goddard (1998).

If the data contains many animals each with many genotypes, then MCMC methods to estimate SNP effects can take so much computer time as to become impractical. Approximations to Bayes B that use an EM algorithm instead of sampling (Shepherd *et al*., 2010) may overcome this problem.

5. The design of breeding programmes that utilize genomic selection

Marker assisted selection is most useful for traits which cannot be recorded on an individual prior to the (minimum) age of breeding (Meuwissen & Goddard, 1996). For instance, traits which are only displayed in females or only observable late in life or after slaughter benefit most. Traditionally traits such as milk yield, which is not displayed by bulls, have been improved by progeny testing bulls based on their daughters' milk yield. This leads to an accurate estimate of the bull's BV but at the expense of a long generation interval. The benefit of genomic selection is that bulls and heifers can be selected early in life and the generation interval reduced leading to approximately doubling genetic gain per year (Schaeffer, 2006; König *et al*., 2009; Pryce *et al*., 2010). This radically changes the design of dairy breeding programmes which have been based on progeny testing. By using genetic markers and genomic selection we can select the best bulls when they are born and breed from them at 1 year of age instead of waiting until they have completed a progeny test at 5 years of age. Despite the large change in breeding programmes needed to capture the benefit of genomic selection it is being widely adopted. In the USA, 52 786 Holstein dairy cattle alone had been genotyped with a SNP chip up to September 2010 (George Wiggans, personal communication).

In developing countries it has been hard to implement traditional genetic improvement programmes because they are logistically complex especially if they require recording the pedigree and production of thousands to millions of animals. Genomic selection might be more practical than traditional selection in these countries. The development of a prediction equation would still require recording the performance of many animals but pedigree would not be required and implementation would require only a DNA sample from each selection candidate and the laboratory facilities to genotype SNPs and compute EBVs from them.

6. Long-term response to genomic selection

If a prediction equation is estimated in the base generation and used for selection for several subsequent

https://doi.org/10.1017/S0016672310000613 Published online by Cambridge University Press

generations, simulation studies show that the selection response declines rapidly (Muir, 2007). Goddard (2009), shows that this is due to two processes. First, selection drives the selected SNP allele towards fixation more quickly than the favourable QTL allele so that the LD between them, which genomic selection relies on, diminishes. Second, traditional mass selection on phenotype does not result in a rapid decline in genetic variance because increasing the frequency of initially rare favourable alleles compensates for the movement towards fixation of common, favourable alleles. However, genomic selection is unlikely to select effectively for rare alleles because they are poorly correlated with the common SNPs. The decline in rate of response to genomic selection is likely to be slower if the trait is controlled by very many genes, each with very small effects, because then the change in allele frequency will be slower.

This reduction in response over time can be reduced in a number of ways. Re-estimating the prediction equation each generation would partially prevent the decline in response (Muir, 2007). Goddard (2009) presented a method to optimize long-term response which decreases selection pressure on QTL that are initially common and of large effect, compared with selection on EBV alone. When very high density SNP genotyping and the Bayes B method of estimation of SNP effects is used, only SNPs that are in close LD to the QTL obtain estimated effects $\neq 0$, with accuracies that persist over time, since the LD persists over time (Meuwissen & Goddard, 2010).

Long-term response is also reduced by inbreeding which of course also causes inbreeding depression. In traditional selection it is possible to balance maximizing the EBV of selected animals with minimizing long-term inbreeding by optimizing the contribution of individual animals to the next generation (Wray & Goddard, 1994; Meuwissen, 1997). By using the relationship matrix estimated from the SNPs this method can be extended to genomic selection (Sonesson & Meuwissen, 2010*b*).

7. Genomic selection in plants and aquaculture

In principle genomic selection could be applied to crops and species used for aquaculture as well as to livestock. However, some practical problems are likely to occur. Some species have very large N_e in the wild and hence, the LD extends over a very short distance. This means that very dense SNP genotyping would be necessary and possibly very large sample sizes as well. This may be uneconomical especially where individual plants or fish have a small value. To overcome these problems it may be necessary to reduce N_e in a breeding programme, for instance, by using only the best families or existing varieties to breed the new strain. A novel design has been suggested (Sonesson *et al*., 2010), where estimation of SNP effects is based on the genotyping of DNA pools and SNP density is reduced by estimating SNP effects within one or a few families.

Deliberating reducing N_e could lead to faster inbreeding and inbreeding depression but this can be avoided when the commercial animal or plant is a cross between two or more lines. Reciprocal recurrent selection is selection within pure strains based on the performance of their crossbred offspring. This is a selection method which increases crossbred performance and heterosis and thus can be described as minimizing inbreeding depression. Genomic selection would be particularly useful for reciprocal recurrent selection because it would eliminate the need for a progeny test and therefore, reduce generation interval.

8. Genomic prediction in humans

In humans the same techniques for predicting genetic value, from a genome wide panel of SNPs, could be used to predict the genetic risk of a particular disease that an individual faces (Wray *et al*., 2007). This could be a more accurate prediction than that already made from family history and used in disease prevention, diagnosis, treatment and counselling. However, the high recent N_e of humans implies that many SNPs and a reference population with very many people will be needed to achieve a highly accurate prediction. It is hoped that use of methods such as Bayes B that identify the SNPs in LD with causative variants will lead to higher accuracy methods such as BLUP.

9. Mapping and identifying QTL

The genome wide SNP genotypes that are used in genomic selection are also used in genome wide association studies (GWAS) to map genes for complex traits (Goddard & Hayes, 2009). Typically in a GWAS each SNP is tested for an association with the traits ignoring all other SNPs. Consequently the association at one SNP could reflect the action of more than one QTL and so the SNP with the largest association may not be the closest SNP to a QTL. In genomic selection all the SNPs are fitted at once which may result in more precise mapping of the QTL. However, when the BLUP model of SNP effects is used many SNPs are estimated to have small effects and the position of the QTL is again blurred. But when a method such as Bayes B is used the SNPs with large effects might be good indicators of the position of QTL, (Verbyla *et al*., 2009).

10. Future developments

The cost of genome sequencing is dropping rapidly so in the near future sequence data on individuals will

https://doi.org/10.1017/S0016672310000613 Published online by Cambridge University Press

supplement SNP genotype data. This will increase the accuracy of EBVs because it will provide very dense markers and will include the causal mutations (Meuwissen & Goddard, 2010). Only a sample of individuals from any species of livestock will be sequenced but other animals will have sequence imputed from SNP genotypes using the database of sequenced animals as a reference. This will provide a large number of animals with phenotypic records and imputed genome sequence and this should constitute a powerful resource for discovering the causal mutations. This should lead to development of prediction equations that persist across generations and across breeds, since the LD between the SNPs and the QTL is (nearly) complete.

References

Bernardo, R. & Yu, J. (2007). Prospects for genome-wide selection for quantitative traits in maize. *Crop Science* **47**, 1082–1090.

Boichard, D., Fritz, S., Rossignol, M. N., Guillaume, F., Colleau, J. J. & Druet, T. (2006). Implementation of marker assisted selection: practical lessons from dairy cattle. In Proceedings of the Eighth World Congress in Genetics Applied to Livestock Production, Electronic communication 22-03.

Chamberlain, A. J., McPartlan, H. & Goddard, M. E. (2007). The number of loci that affect milk production traits in dairy cattle. *Genetics* **177**, 1117–1123.

Daetwyler, H. D., Hickey, J. M., Henshal, J. M., Dominik, S., Gredler, B., van der Werf, J. H. J. & Hayes, B. J. (2010). Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Animal Science* In press.

Dalton, R. (2009). No bull: genes for better milk. *Nature* **457**, 369.

Dekkers, J. C. (2004) Commercial application of marker and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science* **82**, E313–E328.

Dekkers, J. C. M. & Hospital, F. (2002). Multifactorial genetics: the use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**, 22–32.

de Roos, A. P. W., Hayes B. J., Spelman R. & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. *Genetics* **179**, 1503–1512.

Fernando, R. L. & Grossman, M. (1989). Marker-assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution* **21**, 467–477.

Gianola, D., Fernando, R. L. & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**, 1761–1776.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–363.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., Sargeant, L. S., Sorensen, A., Steele, M. R., Zhao, X., Womack, J. E. & Hoeschele, I. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**, 907–920.

Goddard, M. E. (1998). Gene based models for genetic evaluation–an alternative to BLUP? *Proceedings of the Sixth World Congress in Genetics Applied to Livestock Production* **26**, 33–36.

Goddard, M. E. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245–257.

Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**, 381–391.

Grisart, B., Coppitiers, 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 *DG*AT1gene with major effect on milk yield and composition. *Genome Research* **12**, 222–231.

Habier, D., Fernando, R. L. & Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389–2397.

Harris, B. L. & Johnson, D. L. (2010). Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *Journal of Dairy Science* **93**, 1243–1252.

Havenstein, G. B., Ferket, P. R., Scheideler, S. E. & Larson, B. T. (1994). Growth, livability, and feed conversion of 1957 vs 1991 broilers when fed 'typical' 1957 and 1991 broiler diets. *Poultry Science* **73**, 1785–1794.

Hayes, B. J. & Goddard, M. E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics, Selection, Evolution* **33**, 209–229.

Hayes, B. J. & Goddard, M. E. (2008). Technical note: Prediction of breeding values using marker derived relationship matrices. *Journal of Animal Science* **86**, 2089–2092.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. & Goddard, M. E. (2009*a*). Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433–443.

Hayes, B. J., Visscher, P. M. & Goddard, M. E. (2009*b*). Increased accuracy of selection by using the realised relationship matrix. *Genetics Research* **91**, 47–60.

Hayes, B. J., Pryce, J. E., Chamberlain, A. J., Bowman, P. J. & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk fat percentage and type in Holstein cattle as contrasting model traits. *PLOS Genetics* **23**, 6:e1001139.

Heffner, E. L., Sorrels, M. R. & Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science* **49**, 1–12.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph: University of Guelph Press.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *Public Library of Science Genetics* **4**, e1000008.

König, S., Simianer, H. & Willam, A. (2009). Economic evaluation of genomic breeding programs. *Journal of Dairy Science* **92**, 382–391.

Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E. & Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *Public Library of Science Genetics* **4**, e1000231.

Legarra, A. & Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360–366.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. C., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. R., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E, Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A. & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Meuwissen, T. H. E. (1997). Maximizing the response to selection with a predefined rate of inbreeding. *Journal of Animal Science* **75**, 934–940.

Meuwissen, T. H. E. & Goddard, M. E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics, Selection, Evolution* **28**, 161–176.

Meuwissen, T. H., Solberg, T. R., Shepherd, R. & Woolliams, J. A. (2009). A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics, Selection, Evolution* **5**, 41–42.

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.

Meuwissen, T. H. E. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genetics, Selection, Evolution* **41**, 35–44.

Meuwissen, T. H. E. & Goddard, M. E. (2010). Accurate prediction of genetic values for complex traits by whole genome resequencing. *Genetics* **185**, 623–631.

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S. & Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics, Selection, Evolution* **31**, 41–56.

Muir, W. M. (2007). Comparison of genomic and traditional BLUP estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics* **124**, 342–355.

Neimann-Sorensen, A. & Robertson, A. (1961). The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agriculturae Scandinavia* **11**, 163–196.

Pryce, J. E., Goddard, M. E., Raadsma, H. W. & Hayes, B. J. (2010). Deterministic models of breeding scheme designs that incorporate genomic selection. *Journal of Dairy Science* **93**: 5455–5466.

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

Schaeffer, L. R. (2006) Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* **123**, 218–223.

Shepherd, R. K., Meuwissen, T. H. E. & Woolliams, J. A. (2010). Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC Bioinformatics* **11**, 529. [Epub ahead of print]

Sonesson, A. K. & Meuwissen, T. H. E. (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genetics, Selection, Evolution* **30**, 41 : 37.

Sonesson, A. K., Meuwissen, T. H. E. & Goddard, M. E. (2010*a*). The use of communal rearing of families and DNA pooling in multi-trait aquaculture genomic selection schemes. *Genetics Selection Evolution* 42 : 41.

Sonesson A. K., Woolliams J. A. & Meuwissen T. H. E. (2010*b*). Maximising Genetic Gain Whilst Controlling Rates Of Genomic Inbreeding Using Genomic Optimum Contribution Selection. Proceedings of the Ninth World Congress on Genetics Applied to Livestock Production. Paper 0892.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. & Schenkel, F. (2009). Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24.

Van Vleck, L. D., Westall, R. A. & Scneider, J. C. (1986). Genetic change in milk yield estimated from simultaneous genetic evaluation of bulls and cows. *Journal of Dairy Science* **69**, 2963–2965.

Verbyla, K. L., Bowman, P. J., Hayes, B. J. & Goddard, M. E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genetics Research* **91**, 307–311.

Wiggans, G., Cooper, T., VanRaden, P. & Silva, M. (2010). Increased reliability of genetic evaluations for dairy cattle in the United States from use of genomic information. In *Proceedings of the Ninth World Congress in Genetics Applied to Livestock Production*, Electronic communication 476.

Wray, N. R. & Goddard, M. E. (1994). Increasing long term response to selection. *Genetics, Selection, Evolution* **26**, 431–451.

Wray, N. R., Goddard, M. E. & Visscher, P. M. (2007). Prediction of individual risk to disease from genome-wide association studies. *Genome Research* **17**, 1520–1528.

Yang, J., Beben, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. & Visscher, P. M. (2010). Missing heritability of human height explained by genomic relationships. *Nature Genetics* **42**: 565–569.

Yi, N. & Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**, 1045–1055.

Zhong, S., Dekker, J. C. M., Fernando, R. L. & Jannink, J.-L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* **182**, 355–364.