Animal (2010), **4:2**, pp 157–164 © The Animal Consortium 2009
doi:10.1017/S1751731109991352

Genomic breeding value prediction: methods and procedures*

M. P. L. Calus[†]

Animal Sciences Group, Animal Breeding and Genomics Centre, Wageningen University and Research Centre, 8200 AB Lelystad, The Netherlands

(Received 17 November 2008; Accepted 10 July 2009; First published online 6 November 2009)

Animal breeding faces one of the most significant changes of the past decades – the implementation of genomic selection. Genomic selection uses dense marker maps to predict the breeding value of animals with reported accuracies that are up to 0.31 higher than those of pedigree indexes, without the need to phenotype the animals themselves, or close relatives thereof. The basic principle is that because of the high marker density, each quantitative trait loci (QTL) is in linkage disequilibrium (LD) with at least one nearby marker. The process involves putting a reference population together of animals with known phenotypes and genotypes to estimate the marker effects. Marker effects have been estimated with several different methods that generally aim at reducing the dimensions of the marker data. Nearly all reported models only included additive effects. Once the marker effects are estimated, breeding values of young selection candidates can be predicted with reported accuracies up to 0.85. Although results from simulation studies suggest that different models may yield more accurate genomic estimated breeding values (GEBVs) for different traits, depending on the underlying QTL distribution of the trait, there is so far only little evidence from studies based on real data to support this. The accuracy of genomic predictions strongly depends on characteristics of the reference populations, such as number of animals, number of markers, and the heritability of the recorded phenotype. Another important factor is the relationship between animals in the reference population and the evaluated animals. The breakup of LD between markers and QTL across generations advocates frequent re-estimation of marker effects to maintain the accuracy of GEBVs at an acceptable level. Therefore, at low frequencies of re-estimating marker effects, it becomes more important that the model that estimates the marker effects capitalizes on LD information that is persistent across generations.

Keywords: genomic selection, SNP, methods

Implications

The commercial application of genomic selection implies that relatively large numbers of important breeding animals are genotyped using high-density markers. This opens up the possibility to perform quantitative trait loci (QTL)-mapping studies with greater power and precision than was possible beforehand, while the estimated SNP effects are valid for the whole evaluated population rather than, for instance, only within a sire family. Some of the models applied for genomic breeding value prediction find their origin in QTL mapping. For instance, the internal model (or rather single-nucleotide polymorphisms (SNP)) selection step in BayesB, in fact assesses the probability that a SNP is linked to a QTL. This implies that applying such models, as a 'by-product' yield the probability that QTL exist across the genome. Although QTL mapping in itself is not the primary goal, identification of regions that heavily influence a number of traits, would greatly help to further disentangle the nature of different traits, as well as the genetic correlation between them.

Introduction

An important tool in genetic improvement of livestock species is the prediction of breeding values. Breeding value prediction depends on knowledge of relationships between individuals. Defining genetic relationships between animals allows estimation of the proportion of phenotypic variance that is heritable. A major breakthrough in animal breeding was the application of best linear unbiased prediction (BLUP) to predict breeding values, made possible by direct derivation of the inverse additive relationship matrix (Henderson, 1975). Three disadvantages from applying this method to predict breeding values are the following: (i) to estimate reliable breeding values for selection candidates, phenotypic information of the animal itself or close relatives is needed; (ii) BLUP favours close relatives leading to increased

* This paper has been presented at the session 'Genomics Selection and Bioinformatics' of the 59th Annual meeting of the European Association for Animal Production held in Vilnius (Lithuania), 24 to 27 August 2008. Dr A. Maki-Tanila acted as Guest Editor.

[†] E-mail: mario.calus@wur.nl

inbreeding; and (iii) the infinitesimal model is assumed, meaning that an infinite number of genes with small effect underlie a trait. Efforts to apply quantitative trait loci (QTL) mapping, to allow implementation of marker-assisted selection (MAS), tried to tackle both issues (Dekkers and Hospital, 2002). These approaches identify QTL that have a large effect on a trait and trace those to enhance reliability of predicted breeding values, before phenotypic information is available.

Early applications of QTL mapping have the disadvantage that still phenotypic records are needed from the animals themselves or from their close relatives, because they use linkage mapping, where the linkage phase between marker and QTL has to be estimated for instance for each (grand) sire family (e.g. Weller *et al*., 1990). More recent developed methods of QTL mapping rely on population-wide linkage disequilibrium (LD) information. This means that in the model it is assumed that across the population each separate allele of a QTL locus generally is transmitted together with a separate allele at a marker locus. This assumption holds when recombination between a QTL and a marker is limited, that is, the physical distance between a marker and QTL locus that are in LD should be small and thus a high marker density is needed. With increasing distance between two loci that are in LD, the chance of recombination increases, and therefore, the accuracy of forward prediction using estimated single-nucleotide polymorphisms (SNP) effects decreases. The recent availability of high-throughput SNP genotyping provides dense marker maps at reasonably low cost, and allows to map more and smaller QTL. However, the effect of small QTL is hard to estimate, implying that the inaccuracy and bias of estimated QTL effects may increase.

Both the efficiency and the accuracy of detecting QTL can be increased by applying multiple QTL models (Jansen, 1993; Zeng, 1994). Meuwissen *et al*. (2001) applied a multiple QTL model and skipped the QTL-mapping step in a method that is termed genomic selection. This method simultaneously estimates the effects for a large number of markers across the chromosome, for example, approximately 50 000 are currently applied in cattle (Van Tassell *et al*., 2008), and directly sums those effects to total so-called genomic estimated breeding values (GEBVs). The fact that all effects are estimated simultaneously implies that, even though individual marker effects may be overestimated, the expected variance of the total estimated breeding values, including all estimated marker effects, does not exceed the total genetic variance.

The objective of this paper is to review the methods and procedures applied for genomic prediction, as well as the prediction of GEBVs and the accuracy of GEBVs.

Genomic prediction – the process

The key issue in genomic prediction is estimation of effects of individual SNP alleles on a trait of interest. These SNP effects are estimated using a reference population, also termed training data (Meuwissen, 2007) (Figure 1). This reference population typically comprises at least 1000 individuals that have reliable phenotypic as well as genotypic information. This phenotypic information could be own phenotypic performance, but also breeding values obtained from (national) evaluations based on phenotypic information (De Roos *et al*., 2007; De Roos *et al*., 2009; Lund and Su, 2009), deregressed proofs (Berry *et al*., 2009; Schenkel *et al*., 2009; VanRaden *et al*., 2009), daughter-yield deviations or average offspring performance (González-Recio *et al*., 2008). By linking the genotypic and phenotypic information together, estimates for each of the SNPs are obtained. The last step in the process involves genotyping of young selection candidates, whose GEBVs are obtained by summing up all the relevant SNP effects.

An important question is which animals need to be included in the reference population. Several approaches can be taken. In dairy cattle, for instance, the most straightforward approach is to use proven bulls (De Roos *et al*., 2007; VanRaden *et al*., 2009), that have reliable national breeding values, which allows to derive reliable deregressed proofs. This approach is an obvious choice when obtaining reliable phenotypes is time consuming and expensive, compared with the cost of genotyping. Furthermore, the young selection candidates may be close relatives (such as offspring) of the phenotyped animals in the reference population, which may enhance the reliabilities of the breeding values (Habier *et al*., 2007). When animals for the reference population both need to be genotyped and

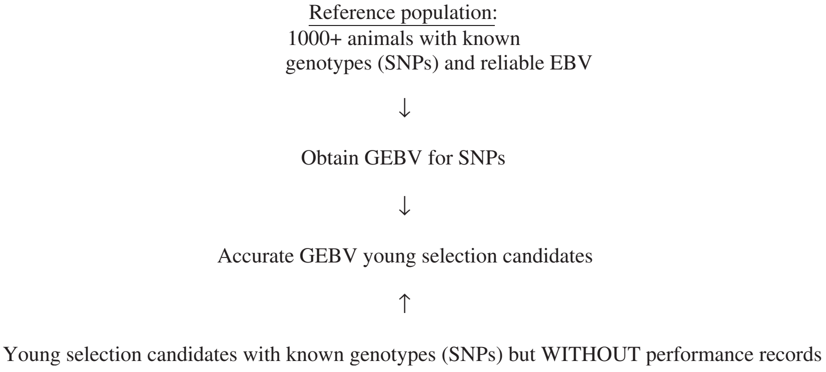
Reference population:
1000+ animals with known
genotypes (SNPs) and reliable EBV
↓
Obtain GEBV for SNPs
↓
Accurate GEBV young selection candidates
↑
Young selection candidates with known genotypes (SNPs) but WITHOUT performance records

Figure 1 Genomic selection – the process.

phenotyped, and costs of phenotyping are low compared with genotyping costs, the design of the reference population may be optimized cost effectively. Although so far not much research is done on the optimal composition of a reference population, theoretically the optimal reference population should comprise the whole range of phenotypes and genotypes, to allow reliable prediction across these ranges. However, this is not possible in real life, so the reference population should be designed to reflect the whole range of phenotypes and genotypes as close as possible. Since prediction of GEBVs is shown to be more reliable when juvenile animals share their recent pedigree with animals in the reference population (Habier *et al*., 2007), it seems straightforward that composing a reference population closely related to the juvenile selection candidates is an optimal strategy. De Roos *et al*. (2008a), however, showed, that it is particularly important that at least some of the animals in the reference population originate from the same pedigree or line as the juvenile animals. This indicates that a reference population that needs to serve multiple lines may optimally contain animals from all lines, and yield nearly as accurate predictions for each of the lines as a line-specific reference population may do. The likely reason for this is that combining animals from different families or lines ensures that only LD that persists across those families or lines is utilized in the prediction equations. This is, however, only possible if marker density is sufficiently high such that each QTL has at least one SNP that is in sufficiently high LD across multiple lines or breeds. De Roos *et al*. (2008b) showed that in order to accurately predict GEBVs for Jerseys, using prediction equations based on a Holstein–Friesian reference population, at least 300 000 SNPs are needed, while the current available SNPs (approximately 50 000) are sufficient for accurate predictions within the same breed. In addition, Harris *et al*. (2008) reported that based on 44 146 SNPs, GEBVs for Jerseys using SNP effects estimated in Holstein, and vice versa, were not accurately predicted. Since within the Holstein–Friesian breed, the average r^2 between adjacent SNPs at a marker density that resembles the currently used 50 000 SNPs is between 0.15 and 0.20 (De Roos *et al*., 2008b; Khatkar *et al*., 2008), it is expected that panels that may be used to predict breeding values across breeds or lines should capture at least the same level of LD across those breeds or lines.

Genome – wide breeding value estimation

Sources of data

In terms of sources of information, the simplest model to predict GEBVs only uses phenotypic and genotypic data (e.g. Meuwissen *et al*., 2001), where genotypic data in recent commercial applications nearly always consist of SNP genotypes. In addition, pedigree data may be used (Figure 2). This allows to derive an additive relationship matrix and incorporation of polygenic breeding values in the model (Calus and Veerkamp, 2007). Whenever pedigree information is not available, the additive relationship matrix can be constructed directly from the genotypic information (e.g. Fernando, 1998). In terms of pre-processing data, pedigree information can be compared with SNP information to discover possible genotype or pedigree errors, while pedigree and SNP information may be used jointly to derive marker haplotypes (e.g. Meuwissen and Goddard, 2001).

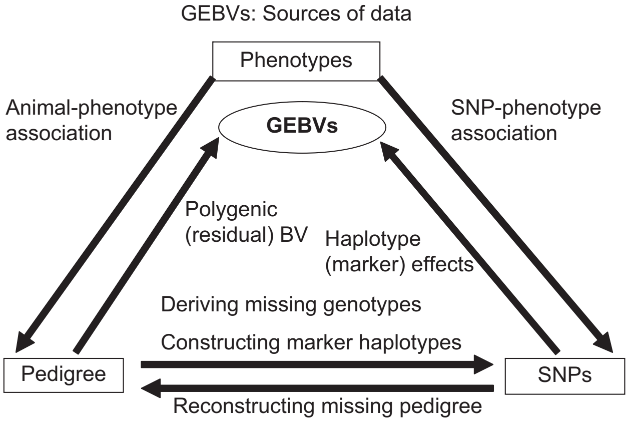
GEBVs: Sources of data
Phenotypes
Animal-phenotype association
GEBVs
SNP-phenotype association
Polygenic (residual) BV
Haplotype (marker) effects
Deriving missing genotypes
Constructing marker haplotypes
Pedigree
SNPs
Reconstructing missing pedigree

Figure 2 Sources of data that may be used in genome-wide prediction of breeding values.

The model

Generally, the model to predict GEBVs, considering only additive genetic effects, is described as:

$$y_i = \text{fixed effects} + \text{animal}_i + \Sigma(\text{SNP}_{ijk}) + e_i$$

where y_i may be a phenotype, national EBV, daughter-yield deviation, deregressed EBV or average offspring performance of animal *i*; fixed effects are a set of fixed effects, which may only be the overall mean, animal_i is a polygenic effect, $\Sigma(\text{SNP}_{ijk})$ is the sum of both SNP effects ($k = 1, 2$), summed across all loci (j) for animal *i*. Note that for both alleles, a separate effect may be estimated ($k = 1, 2$), or alternatively, one allele substitution effect *a* may be estimated per locus.

The GEBVs of animal *i* can be obtained as follows:

$$\text{GEBV}_i = \text{animal}_i + \Sigma(\text{SNP}_{ijk}).$$

Parameterization of the model

The simplest parameterization of the model consists of estimating effects for each SNP allele, or estimating allele substitution effects for each SNP locus (e.g. Xu, 2003b). This implies that both the order of the SNPs along the genome and the linkage phase of the SNP alleles do not need to be known. A disadvantage of fitting the SNP alleles is that the model ignores the effect of recombination, which may change the linkage between marker and QTL alleles. Estimating effects of constructed marker haplotypes solves this problem to some extent: copies of each unique haplotype are assumed to carry the same QTL alleles. In this case, the marker haplotypes are treated to be alike-in-state (AIS; e.g. Meuwissen *et al*., 2001; Villumsen and Janss, 2008).

The assumption that copies of each unique haplotype carry the same QTL allele can be relaxed by treating haplotypes to be identical-by-descent (IBD; Meuwissen and Goddard, 2001). Two haplotypes are IBD whenever they consist of the same marker alleles due to inheritance from a common ancestor. Since the probability that two haplotypes are IBD is nearly always larger than zero and smaller than one, predicted IBD probabilities between haplotypes are included in the model to account for the relationships between haplotypes (Meuwissen and Goddard, 2001). An additional difference between marker AIS and IBD haplotypes is that the IBD status of haplotypes is evaluated at one position in the genome (a so-called putative QTL position), whereas AIS haplotypes are considered to be AIS across the whole haplotype. When IBD probabilities between two haplotypes are close to one, those haplotypes are usually considered to be the same to reduce the number of effects that needs to be estimated and thereby increasing power to estimate effects (Yu *et al*., 2005) and speeding up convergence (Calus *et al*., 2009).

Underlying QTL distribution

Estimated SNP or haplotype effects capture at least partly the underlying QTL effects. Consequently, the distribution of estimated SNP or haplotype effects should resemble the distribution of the underlying QTL effects. Estimated distributions of QTL effects suggest that there are many loci of small (near zero) effect and that there are few loci with large effect (Hayes and Goddard, 2001). However, for many traits, the distribution of underlying QTL may largely be unknown *a priori*. The design of QTL-mapping studies may lead to overestimated effects for large QTL (Xu, 2003a), while small QTL may not be picked up at all. Furthermore, the effect of a single QTL may be explained by multiple SNPs, leading to more SNP supposedly linked to QTL, on average having relatively smaller effects than the QTL. This suggests that prior information on the distribution of QTL effects should be used with caution when designing models to estimate SNP effects, or, alternatively, that models should be able to make inferences from the data with regard to the distribution of QTL effects.

Methods to solve the model

The main challenge for models that estimate SNP effects is that generally phenotypes are available from a few thousand animals, while the number of SNP is much larger, for instance, approximately 50 000 SNPs for cattle. A practical problem is how this set of equations can be solved. Another issue is that the distribution of QTL effects implies that there are many loci of small (near zero) effect and that there are few loci with large effect (Hayes and Goddard, 2001). This implies that methods may have to be able to eliminate loci with (near) zero effect and/or have to be able to deal with the problem that the number of loci whose effect need to be estimated, is much larger than the number of records available. A straightforward approach is to select a reduced set of explaining loci. This can be done by just looking at the genomic variation, that is, by selecting a subset of SNPs (so-called tag-SNPs), that together explain a large part of the genomic variation in the full set of SNPs (e.g. Ke and Cardon, 2003). Another approach is to select a subset of SNPs that may be associated with a phenotype of interest. Applied approaches are, for instance, machine learning classification procedures (Long *et al*., 2007), or forward stepwise regression to determine for each SNP whether it significantly explains phenotypic variance, given the effects of the already selected SNPs (Habier *et al*., 2007).

Conceptually more appealing is to incorporate a step in the model that determines for each of the SNPs whether it is associated with a QTL or not, and therefore, whether it has a non-zero effect on the phenotype or not. Such an approach was implemented in a Bayesian model by Meuwissen *et al*. (2001), and termed BayesB. In BayesB, a Metropolis–Hastings step is used in every iteration to determine for each locus whether it has an effect on the phenotype or not. If it has not, the effect of that locus is set to zero. One concern about this model is that it relies quite heavily on the considered prior information to infer the SNP effects (Gianola *et al*., 2006).

An alternative approach is to apply a mixture model, where one distribution of variances is assumed for loci with an association with the phenotype, and a distribution with very small values is assumed for loci with no association with the phenotype (Meuwissen and Goddard, 2004; Calus *et al*., 2008; Janss *et al*., 2008). This method allows to apply Gibbs sampling, while avoiding the Metropolis–Hastings step that is used in BayesB. The benefit of this method is that all loci for which no clear evidence is found for a direct association, together still may explain a substantial part of the genetic variance (Calus *et al*., 2008). This makes that these mixture models are able to combine the properties of the BayesB and the 'genomic BLUP' (GBLUP) model; the distribution with large effects is expected to pick up the same SNPs as the BayesB model, while the other SNPs still all explain a roughly equal small amount of the variance as in the GBLUP model. The GBLUP model assumes that each of the SNPs equally contributes to the additive genetic variance (Meuwissen *et al*., 2001). It has been shown that constructing a genomic relationship matrix based on all markers and included in a regular mixed model, instead of a pedigree-based relationship matrix, is equivalent to the GBLUP method (Goddard, 2009). The contribution of each locus to the genomic relationship matrix can be weighted to account for different variances per locus or differences in local marker density (VanRaden, 2008).

Several other models have been considered for genome-wide prediction. Apart from a few methods that assume equal contribution to the genetic variance by all loci, a common feature of these methods is to reduce the dimension of the SNP data. Reported alternatives include non-parametric kernel methods (Gianola *et al*., 2006; Bennewitz and Meuwissen, 2008; Gianola and van Kaam, 2008), partial least squares (PLS) regression (Raadsma *et al*., 2008; Solberg, 2008), principal component analysis (PCA; Solberg, 2008), genetic algorithms, and Bayesian LASSO (de los Campos *et al*., 2009). Kernel methods were shown to yield similar results as the Bayesian methods for purely additive models (Bennewitz and Meuwissen, 2008), but may outperform the Bayesian models when considering non-additive effects (González-Recio *et al*., 2008). PCA reduces dimensionality of the SNP data matrix by finding a few variables explaining as much variance as possible, while PLS does the same, yet conditional on the dependent variable (i.e. the phenotypic information).

Nearly all of the mentioned studies only considered additive genetic effects in the model. Some models have been presented that explicitly include dominance (Xu, 2003b) or epistatic effects (Xu and Jia, 2007). The main problem with the epistatic effects is that for a high number of markers the number of potential interactions becomes unreasonably high, and evaluating all of them is impossible. A suggested solution involves modelling one genetic term that includes additive and epistatic effects, without discriminating explicitly between them (Gianola *et al*., 2006; Gianola and van Kaam, 2008).

Accuracies of GEBVs

Results from simulation studies

In the following, the term accuracy is used for the correlation between true and estimated breeding values, usually denoted as r_{TI}. The accuracy of GEBV depends partly on the parameterization of the model and the statistical model that is used to solve it. Comparisons, based on simulated data, indicated that at low marker density the IBD-haplotype parameterization yielded considerable higher accuracy followed by the IBS-haplotype parameterization and the single SNP model (Calus *et al*., 2008). At high marker density, with LD comparable to the marker density of commercially available SNP chips, the differences in accuracy between the models were negligible (Calus *et al*., 2008). Including a polygenic effect in the model also increased the accuracy of GEBVs at low marker density (Calus and Veerkamp, 2007), while at high marker density, or in a situation where the markers already explained most of the genetic variance, including a polygenic effect, hardly increased the accuracy of the GEBVs (Calus and Veerkamp, 2007; Solberg, 2008).

In simulation studies, Bayesian models, including BayesB and mixture models, have been reported to yield accuracies for animals without phenotypic records ranging from 0.7 to 0.8, when the marker density was equivalent to one SNP per cM and approximately 1000 animals were included in the reference population (Meuwissen *et al*., 2001; Habier *et al*., 2007; Calus *et al*., 2008; Solberg *et al*., 2008). In these studies, the simulated QTL followed a gamma (Meuwissen *et al*., 2001; Calus *et al*., 2008; Solberg *et al*., 2008) or normal distribution (Habier *et al*., 2007). Both distributions imply that a limited number of loci explains a large amount of the genetic variance, which fits the assumption of those Bayesian models. When the number of QTL (8700) was large relative to the number of markers (8729), a model that assumed equal contribution to the additive variance by each SNP (i.e. including a genomic relationship matrix instead of a pedigree-based relationship matrix) was shown to need at least 5000 (out of 8729) markers across the genome to give an accurate prediction of the simulated heritability (Hayes and Goddard, 2008). These results suggest that fine-tuning the model depending on the trait and the underlying QTL distribution greatly adds to the fit of the model to the data.

Results based on real data

In applications to real dairy cattle data, allowing unequal contributions of SNPs to the genetic variance yielded GEBVs that had only a slightly higher reliability than GEBVs estimated with a model that assumed equal contributions of SNPs to the genetic variance (Hayes *et al*., 2009; VanRaden *et al*., 2009). Reliabilities were, however, substantially higher using BayesB for traits that are affected by DGAT (VanRaden *et al*., 2009). These results suggest that the distribution of QTL effects in real data is generally less extreme than assumed in the above-mentioned simulation studies, but more research is needed to verify this.

Table 1 gives a summary of ranges of GEBV accuracies for several dairy cattle breeding traits obtained from published

Table 1 *Ranges of accuracies for genomic estimated breeding values across traits as estimated in different countries for various sizes of the reference population*

Reference	Size reference population	Number of SNPs included	Range accuracy of GEBVs	Range accuracy of GEBVs – accuracy of PA[1]
Table 2 Berry *et al.* (2009)	596	42 598	0.56 to 0.71	0.01 to 0.17
Table 1 Lund and Su (2009)	1238[2]	38 055	0.55 to 0.85	
Harris *et al.* (2008 and 2009)	2490	44 146	0.63 to 0.82	0.15 to 0.24
Table 1 De Roos *et al.* (2009)	3600	48 000	0.52 to 0.82	0.04 to 0.23
Table 3 Schenkel *et al.* (2009)	3966–4127	38 416	0.36 to 0.77	−0.01 to 0.29
Table 3 VanRaden *et al.* (2009)	5335	38 416	0.44 to 0.79	0.18 to 0.31

GEBVs = genomic estimated breeding values; PA = parent average.

[1] That is, pedigree index.

[2] Calculated as 80% of the reported 1548 bulls in the reference population.

studies based on real data with approximately 40 000 SNPs. Dairy cattle was used here as an example, because the results published on other species are limited so far. The accuracies from the different studies in Table 1 were obtained using different methods, apply for different groups of animals and may be for GEBVs or GEBVs blended with traditional proofs. Here, accuracies were used that were obtained for blended proofs of young bulls, or attempted to resemble those, whenever available. All studies presented in Table 1 showed that the accuracy of GEBVs was up to 0.31 higher than for pedigree indexes for bulls without any phenotypic data on offspring available, although for some traits, the GEBVs did not outperform pedigree indexes. Although inclusion of many different traits and use of different methods to assess the accuracy in different studies makes comparison of the obtained accuracies difficult, the reported maximum accuracy of each study was for a milk production trait. Those maximum values per study show that having 600 reference bulls resulted in lower accuracies (0.71) than when using at least 1200 reference bulls (accuracy $\geqslant$0.77). Despite the limited number of studies in Table 1, and differences among the studies, the results suggest that there is a trend of increasing accuracy of GEBVs with increasing numbers of bulls in the reference population. Therefore, it seems advisable to include $>$1000 bulls in the reference population to obtain GEBVs for juvenile selection candidates with accuracies that are higher than those for pedigree indexes for all traits.

Factors affecting the accuracy of GEBVs

Independent of the applied model to estimate marker effects, the accuracy of GEBVs strongly depends on the LD between marker and QTL loci that is consistent between the reference population and the animals for which GEBVs are predicted, and the accuracy of the estimated marker effects (Goddard, 2009). The accuracy of the estimated marker effects depends on the characteristics of the reference population, such as the number of included phenotypes (Hayes *et al.*, 2009; VanRaden *et al.*, 2009), sampling of animals from the population (Habier *et al.*, 2007) and the heritability of the trait (Calus *et al.*, 2008; Hayes *et al.*, 2009). The 'phenotypes' of the reference population that are used to estimate the SNP effects may be deregressed EBVs, as applied in many cases for dairy bulls. In this case, the heritability is effectively increased per animal, by adjusting the residual variance by a weight that is calculated from the number of offspring contributing to the EBVs of the bull (e.g. Fikse and Banos, 2001). The LD between marker and QTL loci depends largely on the marker density. This indicates that the accuracy of GEBVs can be increased by increasing the number of animals in the reference population (Meuwissen *et al.*, 2001), by sampling animals within families relevant for selection (Legarra *et al.*, 2008), by increasing the marker density (Meuwissen *et al.*, 2001; Calus *et al.*, 2008; Solberg *et al.*, 2008) and by increasing the heritability (Calus *et al.*, 2008; Solberg, 2008).

The accuracy of prediction of GEBVs across and within different families or lines may be different, because the marker effects may absorb polygenic effects from one pedigree structure that are different from another structure (Habier *et al.*, 2007; Legarra *et al.*, 2008). This does have a consequence not only for the construction of the reference population set, but also for the evaluation of different methods to predict GEBVs. An important additional parameter to compare different methods may therefore be the amount explained within and across family variance (Legarra and Misztal, 2008).

Persistency of LD and accuracy of GEBVs across generations

Accurate genomic breeding value prediction strongly depends on the persistency of LD between markers and QTLs across generations. LD within a population can be created and maintained by selection, migration, mutation and drift (Lynch and Walsh, 1998). The same forces may, however, decrease the LD between two loci, especially if the allele frequency at either locus is substantially changed. As a consequence, when marker effects are not re-estimated, the accuracy of GEBVs across generations is expected to decrease. In simulation studies, it was shown that the accuracy of GEBV decrease slowly when mating is random (Meuwissen *et al.*, 2001; Solberg, 2008), and more rapid when selection is considered (Muir, 2007). Since recombination will break up LD in both situations, this indicates that selection is an important factor to break up LD between markers and QTL.

The breakup of LD between markers and QTL, and consequently the reduction of accuracy of GEBVs across generations of selection, is one factor that determines the optimal frequency to re-estimate SNP effects. Another factor, from the perspective of the breeding program, is the frequency in which reliable phenotypes become available to re-estimate the SNP effects. In dairy cattle, a practical approach would be to continuously add phenotypic daughter information that becomes available for bulls that were selected as juveniles based on their GEBVs. In the time required to obtain this daughter information, two to three generations of selection maybe performed purely based on GEBVs (Villumsen *et al.*, 2009). Some studies showed that depending on the frequency of re-estimating the effects, different methods may be more suitable to obtain SNP effects whose accuracies are more persistent across generations. For instance, methods that are better able to disentangle between LD and linkage information show better persistency of GEBV accuracy across generations (Habier *et al.*, 2007; Zhong *et al.*, 2009). Despite the finding that approximately 50 000 SNPs are sufficient for within-breed genomic breeding value prediction in cattle, larger SNP arrays that undoubtedly will become available in the future will increase the LD between SNPs and QTL, leading to slower reduction of the accuracy of GEBVs across generations.

In most applications of genomic breeding value estimation, only additive genetic effects are considered. Despite this, a part of the estimated genetic effects may actually be non-additive by nature (Hill *et al.*, 2008). This may decrease the predictive ability of GEBVs across generations as well. For instance, in the situation that an epistatic effect between two unlinked loci is partially absorbed in the estimated additive effects due to extreme allele frequencies, the inheritance of this 'additive' effect will not follow the Mendelian rules. As a consequence, these fitted 'additive' effects may improve the fit of the model to the data, but actually decrease the accuracy of forward prediction.

Conclusions

Genomic selection is the ultimate application of markers in animal breeding. Using a reference population of animals with known genotypes and phenotypes enables to predict breeding values of young selection candidates that have reported accuracies that are up to 0.31 higher than those of pedigree indexes. Published results indicate that for dairy cattle at least approximately 1000 bulls are required in the reference population to obtain GEBVs with accuracies that compete with the accuracies of EBVs based on progeny testing for all traits. Several methods, of which most aim at reducing the dimension of the marker data, can be applied to estimate marker effects. The accuracy of genomic selection strongly depends on the characteristics of the reference population, as well as the relationship between the evaluated animals and the reference population. Breakup of LD between markers and QTL across generations advocates frequent re-estimation of marker effects to maintain the reliability of GEBVs at an acceptable level. Therefore, at low frequencies of re-estimating marker effects, it becomes more important that the model that estimates the marker effects capitalizes on LD information that is persistent across generations.

Acknowledgements

The author acknowledges the financial support of the Netherlands Organization for Scientific Research (NWO – Casimir), CRV and Hendrix Genetics. Two anonymous reviewers are thanked for useful and constructive comment on the manuscript.

References

Bennewitz J and Meuwissen THE 2008. Genomic breeding value estimation using kernel regression and additive models. In 12th Quantitative Trait Locus and Marker Assisted Selection Workshop, Uppsala, Sweden, p. 34.

Berry DP, Kearney F and Harris BL 2009. Genomic Selection in Ireland. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, Bulletin no. 39.

Calus MPL and Veerkamp RF 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal Of Animal Breeding And Genetics 124, 362–368.

Calus MPL, Meuwissen T, De Roos APW and Veerkamp RF 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178, 553–561.

Calus MPL, Meuwissen T, Windig JJ, Knol EF, Schrooten C, Vereijken ALJ and Veerkamp RF 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. Genetics Selection Evolution 41, 11.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K and Cotes JM 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375–385.

De Roos APW, Hayes BJ and Goddard ME 2008a. Reliability of genomic breeding values across multiple populations. In 12th QTL-MAS Workshop, Uppsala, Sweden, p. 33.

De Roos APW, Hayes BJ, Spelman RJ and Goddard ME 2008b. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179, 1503–1512.

De Roos APW, Schrooten C, Mullaart E, Calus MPL and Veerkamp RF 2007. Breeding value estimation for fat percentage using dense marker maps on *Bos taurus* autosome 14. Journal of Dairy Science 90, 4821–4829.

De Roos APW, Schrooten C, Mullaart E, Van der Beek S, De Jong G and Voskamp W 2009. Genomic Selection at CRV. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, Bulletin no. 39.

Dekkers JCM and Hospital F 2002. The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics 3, 22–32.

Fernando RL 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia.

Fikse WF and Banos G 2001. Weighting factors of sire daughter information in international genetic evaluations. Journal of Dairy Science 84, 1759–1767.

Gianola D, Fernando RL and Stella A 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173, 1761–1776.

Gianola D and van Kaam JBCHM 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178, 2289–2303.

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

González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM and Avendaño S 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178, 2305–2313.

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

Harris BL, Johnson DL and Spelman RJ 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. In Proceedings of the 36th ICAR Biennial Session Held in Niagara Falls, USA, pp. 325–330.

Harris BL and Montgomerie WA 2009. Current status of the use of genomic information in the national genetic evaluation in New Zealand. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, Bulletin no. 39.

Hayes BJ, Bowman PJ, Chamberlain AJ and Goddard ME 2009. Invited review: genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science 92, 433–443.

Hayes B and Goddard ME 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33, 209–229.

Hayes BJ and Goddard ME 2008. Technical note: prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86, 2089–2092.

Henderson CR 1975. Rapid method for computing inverse of a relationship matrix. Journal of Dairy Science 58, 1727–1730.

Hill WG, Goddard ME and Visscher PM 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genetics 4, 2.

Jansen RC 1993. Interval mapping of multiple quantitative trait loci. Genetics 135, 205–211.

Janss L, Gregersen V, Bendixen C and Lund M 2008. Validation of genomic predictions in pigs using medium-dense marker coverage. In Book of Abstracts of the 59th Annual meeting of the EAAP, Vilnius, Lithuania.

Ke XY and Cardon LR 2003. Efficient selective screening of haplotype tag SNPs. Bioinformatics 19, 287–288.

Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Al Cavanagh J, Barris W, Schnabel RD, Taylor JF and Raadsma HW 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

Legarra A and Misztal I 2008. Computing strategies in genome-wide selection. Journal of Dairy Science 91, 360–366.

Legarra A, Robert-Granié C, Manfredi E and Elsen J-M 2008. Performance of genomic selection in mice. Genetics 180, 611–618.

Long N, Gianola D, Rosa GJM, Weigel KA and Avendaño S 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal Of Animal Breeding And Genetics 124, 377–389.

Lund M and Su G 2009. Genomic selection in the Nordic countries. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, Bulletin no. 39.

Lynch M and Walsh B 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland.

Meuwissen T 2007. Genomic selection : marker assisted selection on a genome wide scale. Journal of Animal Breeding and Genetics 124, 321–322.

Meuwissen T and Goddard ME 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution 33, 605–634.

Meuwissen T and Goddard ME 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genetics Selection Evolution 36, 261–279.

Meuwissen T, Hayes BJ and Goddard ME 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Muir WM 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.

Raadsma HW, Moser G, Crump R, Khatkar MS, Zenger KR, Cavanagh JAL, Hawken RJ, Sölkner J and Tier B 2008. Predicting genetic merit for mastitis and fertility in dairy cattle using genome wide selection and high density SNP screens. In Conference of the International Society for Animal Genetics, Amsterdam, The Netherlands.

Schenkel FS, Sargolzaei M, Kistemaker G, Jansen GB, Sullivan P, Van Doormaal BJ, VanRaden PM and Wiggans GR 2009. Reliability of genomic evaluation of Holstein cattle in Canada. Proceedings of the Interbull International Workshop – Genomic Information in Genetic Evaluations, Uppsala, Sweden, Bulletin no. 39.

Solberg TR 2008. Methods for prediction of genome-wide breeding values using dense marker genotyping. PhD, Norwegian University of Life Sciences.

Solberg TR, Sonesson AK, Woolliams JA and Meuwissen THE 2008. Genomic selection using different marker types and densities. Journal of Animal Science 86, 2447–2454.

VanRaden PM 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF and Schenkel FS 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC and Sonstegard TS 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5, 247–252.

Villumsen TM and Janss L 2008. Genomic selection focusing on haplotype length and prior settings. In 12th Quantitative Trait Locus and Marker Assisted Selection Workshop, Uppsala, Sweden, p. 41.

Villumsen TM, Janss L and Lund MS 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics 126, 3–13.

Weller JI, Kashi Y and 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.

Xu SZ 2003a. Theoretical basis of the Beavis effect. Genetics 165, 2259–2268.

Xu SZ 2003b. Estimating polygenic effects using markers of the entire genome. Genetics 163, 789–801.

Xu SZ and Jia ZY 2007. Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics 175, 1955–1963.

Yu K, Xu J, Rao DC and Province M 2005. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. Annals of Human Genetics 69, 577–589.

Zeng ZB 1994. Precision mapping of quantitative trait loci. Genetics 136, 1457–1468.

Zhong S, Dekkers JCM, Fernando RL and 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.