Genet. Res., Camb. (2009), **91**, *pp.* 47–60. © 2009 Cambridge University Press
doi:10.1017/S0016672308009981 Printed in the United Kingdom

Increased accuracy of artificial selection by using the realized relationship matrix

B. J. HAYES[1]*, P. M. VISSCHER[2] AND M. E. GODDARD[1,3]
[1] *Biosciences Research Division, Department of Primary Industries Victoria, 1 Park Drive, Bundoora 3083, Australia.*
[2] *Queensland Institute of Medical Research, Brisbane, Australia.*
[3] *Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia.*

(*Received 23 July 2008 and in revised form 4 December 2008*)

Summary

Dense marker genotypes allow the construction of the realized relationship matrix between individuals, with elements the realized proportion of the genome that is identical by descent (IBD) between pairs of individuals. In this paper, we demonstrate that by replacing the average relationship matrix derived from pedigree with the realized relationship matrix in best linear unbiased prediction (BLUP) of breeding values, the accuracy of the breeding values can be substantially increased, especially for individuals with no phenotype of their own. We further demonstrate that this method of predicting breeding values is exactly equivalent to the genomic selection methodology where the effects of quantitative trait loci (QTLs) contributing to variation in the trait are assumed to be normally distributed. The accuracy of breeding values predicted using the realized relationship matrix in the BLUP equations can be deterministically predicted for known family relationships, for example half sibs. The deterministic method uses the effective number of independently segregating loci controlling the phenotype that depends on the type of family relationship and the length of the genome. The accuracy of predicted breeding values depends on this number of effective loci, the family relationship and the number of phenotypic records. The deterministic prediction demonstrates that the accuracy of breeding values can approach unity if enough relatives are genotyped and phenotyped. For example, when 1000 full sibs per family were genotyped and phenotyped, and the heritability of the trait was 0·5, the reliability of predicted genomic breeding values (GEBVs) for individuals in the same full sib family without phenotypes was 0·82. These results were verified by simulation. A deterministic prediction was also derived for random mating populations, where the effective population size is the key parameter determining the effective number of independently segregating loci. If the effective population size is large, a very large number of individuals must be genotyped and phenotyped in order to accurately predict breeding values for unphenotyped individuals from the same population. If the heritability of the trait is 0·3, and $N_e = 1000$, approximately 5750 individuals with genotypes and phenotypes are required in order to predict GEBVs of un-phenotyped individuals in the same population with an accuracy of 0·7.

1. Introduction

In best linear unbiased prediction (BLUP) of breeding values, information from performance of relatives is incorporated through the use of a relationship matrix. Elements of this matrix are derived as the predicted proportion of the genome that is identical by descent (IBD) among two individuals given their pedigree relationship. However, Mendelian sampling during gamete formation results in variation in the realized proportion of the genome, which is IBD between pairs of individuals with the same predicted relationship coefficients (Franklin, 1977; Hill, 1993; Guo, 1996). For example, between full-sib individuals the

* Corresponding author. Tel: +61 (0)3 9479 5439. Fax: +61 (0)3 9479 3113. e-mail: ben.hayes@dpi.vic.gov.au

predicted proportion of the genome that is IBD is 0·5, while its standard deviation is 0·04 for a species with 30 chromosomes each of 1 M in length (Guo, 1996).

DNA marker information can be used to calculate the realized relationship matrix with elements the actual proportion of the genome that is IBD between two individuals, with a high degree of precision, provided that a sufficient number of markers are used. Nejati-Javaremi *et al.* (1997) demonstrated with simulation that if the loci contributing to trait variation were known, and the alleles at these loci were used to derive the realized relationship matrix, the accuracy of breeding values calculated using this matrix could be higher than that calculated using the predicted relationship matrix. In practice, all the loci contributing to trait variation are unlikely to have been identified. Villanueva *et al.* (2005) demonstrated by simulation that using the realized relationship matrix derived from markers rather than the predicted relationship matrix in the calculation of estimated breeding values (EBVs) could lead to higher accuracies of selection. They proposed that marker information used in this way could offer benefits in selection programmes when no quantum trait locus (QTL) has been mapped or when the underlying genetic model can be considered the infinitesimal model, where no individual QTL has a moderate to large effect on the trait. For some traits such as height in humans, this is indeed the case, with the largest reported QTLs explaining only a small fraction of the genetic variance (e.g. Sanna *et al.*, 2008; Visscher, 2008).

While Villanueva *et al.* (2005) considered estimating realized relationships conditional on a known pedigree (exploiting linkage information) realized relationship coefficients can also be estimated for 'unrelated' individuals within a population. This requires sufficient marker density to identify chromosome segments in two individuals that are descended from the same common, but unknown ancestor.

An alternative method by which DNA marker data can be used to estimate breeding values is genomic selection (Meuwissen *et al.*, 2001). In this method, the markers are used to track QTLs whose effects are estimated and summed to predict the breeding value of each individual. However, if there are many QTLs whose effects are normally distributed with constant variance, then genomic selection can be equivalent to the use of the realized relationship matrix (e.g. Fernando, 1998; Habier *et al.*, 2007; Van Raden, 2007 and Goddard, 2008).

Currently, there is no analytical method available to predict the accuracy of EBVs calculated using the genomic relationship matrix considering information from relatives. Analytical expressions would be desirable to guide the design of experiments aiming to achieve a given accuracy of genomic breeding values (GEBVs). Our objective was to derive such expressions for the accuracy of GEBV considering information from relatives. We also modify the expression of Goddard (2008) for the accuracy of GEBV in random mating populations to improve the predictions. Our starting point for all derivations was the equivalent genomic selection model. We then verified the analytical predictions using two simulation approaches. First, we derive a prediction of the accuracy based on the prediction error variance (PEV) where the realized relationship matrix is determined by a large number of informative markers. Secondly, we derive accuracy from simulations with both markers and QTLs segregating as the correlation between true and predicted breeding values. We then investigate the sensitivity of the results to the number of markers used, the number of QTLs and effective population size.

2. Methods

(i) *An equivalent model for genomic selection*

This material is also contained in Goddard (2008) but is included here for completeness. Consider a model of the true breeding value of the ith individual (g_i) based on a large number of QTLs of small effect. To simplify our analytical derivation, we will define a parameter q as the number of independent chromosome segments. This model can be pictured as dividing the chromosomes into segments that effectively segregate independently and defining the effect of the segment as the sum of the effects of the QTL carried on that segment. The assumption here is that there are at least as many QTLs as there are effective chromosome segments. Alternatively if QTLs are unlinked, then q is the number of unlinked QTLs. Then

$$g_i = \sum_{j=1}^{q} W_{ij} u_j,$$

where u_j is the allele substitution effect at the jth QTL and is normally distributed $u \sim N(0, \sigma_u^2)$, where σ_u^2 is the variance of the effect of QTL alleles sampled randomly from the population, and W_{ij} is 0, 1 or 2 if individual i carries 0, 1 or 2 copies of the second allele at the jth QTL. In practice, it is convenient to subtract the mean value of w from each element so that $W_{ij} = 0 - 2p_j$ or $1 - 2p_j$ or $2 - 2p_j$, where $p_j =$ the allele frequency of the second allele at locus j. This corresponds to the genomic selection model that Meuwissen *et al.* (2001) called the BLUP model. A simple version of genomic selection is to define the W_{ij} based on markers instead of the QTL. Then the best estimates of the u_j and hence g_i can be obtained by BLUP.

In matrix form $\mathbf{g} = \mathbf{Wu}$ and $V(\mathbf{g}) = \mathbf{WW}'\sigma_u^2$, where $\mathbf{W}$ is a design matrix allocating QTL allele effects to

individuals. **g** is also normally distributed since it is the sum of many normally distributed effects.

Now a vector of phenotypic records **y** can be modelled as either

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZWu} + \mathbf{e} \quad (1)$$

or

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}, \quad (2)$$

where **X** is a design matrix, **b** is a vector of fixed effects and **Z** is matrix allocating records to individuals. The two models are equivalent provided $V(\mathbf{g}) = \mathbf{G}\sigma^2_g = \mathbf{WW}'\sigma^2_u$, where $\mathbf{G} = \mathbf{WW}'\sigma^2_u/\sigma^2_g$ is the relationship matrix calculated from the markers, and σ^2_g is the genetic variance. Elements of **G** are G_{ik}, the proportion of the genome that is IBD between individuals i and k. Relationships, like inbreeding coefficients, are always relative to a base population. By subtracting the mean allele frequencies from the elements of W, the relationships G_{ik} are relative to the current population. Consequently, they average approximately zero and some are negative. This means that the genetic variance σ^2_g is also the genetic variance in the current population. The two models give the same estimates of $\mathbf{g} = \mathbf{Wu}$. That is, a genomic selection model (1) with normally distributed QTL effects is equivalent to a conventional individual model (2) with the relationship matrix among the individuals estimated from the markers. Note that it assumes that the genotypes are known without error.

(ii) *Derivation of accuracies for breeding values predicted with the equivalent model*

When the marker data have been collected on a sample of individuals, we can use either (1) or (2) to calculate GEBVs for those individuals and their reliabilities. However, it would be useful to predict in general, before collecting the marker data, the accuracy that this form of genomic selection would achieve. It is difficult to derive a formula for reliability based on (2) because **G** is a complicated matrix. It is perhaps easier to work with (1) but this is still difficult because the design matrix, **Z′W′WZ**, the inverse of which occurs in the system of equations required to predict the QTL effects, is complex and likely to have singularities. This complexity comes about because w_{ij} for closely linked markers are correlated. Therefore, we will approximate (1) by a model in which there are q independent chromosome segments as described above. In what follows, we first derive the number of independent chromosome segments in different family relationships or a random mating population, and then use these numbers in the derivation of accuracy of GEBV.

(a) *Effective number of independent chromosome segments within families*

We will determine the effective number of chromosome segments by considering the variation in relationship between pairs of individuals with the same pedigree. For instance, based on pedigree all full sibs have a relationship of 0·5 but in reality this relationship varies from about 0·4 to 0·6 (Hill, 1993; Visscher *et al*., 2006). This variation in relationship comes about because sibs inherit large segments of chromosomes from their parents. The more the independent chromosomes segments make up the genome the more closely all full sibs would come to sharing exactly 50 % of their genome.

Formulae for the variation in realized relationship between the different types of relatives have been published by Hill (1993) and Guo (1996) and we will use their formulae.

Consider a single locus and calculate the relationship between relatives i and j, i.e. G_{ij}. For full-sibs 25 % of the time $G_{ij} = 1$, 50 % of the time $G_{ij} = 0{\cdot}5$ and 25 % of the time $G_{ij} = 0$. So the variance is 1/8. If there are q independent chromosome segments, then the variance of $G_{ij} = 1/(8q)$. However, the variance in G_{ij} can also be calculated for a genome consisting of chromosomes of known length in Morgan. Hill (1993) and Guo (1996) present formulae for this. For instance, using their formula, if the genome consists of a single chromosome 35 M long, then the variance in relationship between pairs of full-sibs is 0·00177. Equating this to the variance of G_{ij} in our model with q independent chromosome segments, $1/(8q) = 0{\cdot}00177$. So the effective number of loci is $q = 1/(8 * 0{\cdot}00177) = 70{\cdot}6$, close to the reported empirical value of 82 for human full sibs (Visscher *et al*., 2006). So if two gametes produced by the same sire (corresponding to two sibs) are considered, then a 35 M chromosome will experience approximately 70 crossovers (35 for each gamete). Therefore, the two gametes can be considered as composed of 70 segments and for each segment the probability that the two gametes are identical is 0·5. Although we have assumed a single chromosome here, Hill (1993) showed that the variation in relationship is not particularly affected by assumptions on the number of chromosomes, provided the total length of the genome was kept constant.

With the same assumptions as above, for half sibs, the $V(G_{ij})$ for a single locus is 1/16 and the variance of relationship for a genome with one chromosome 35 M long is 0·00088 and so again $q \sim 70$. For double cousins, $V(G_{ij})$ for a single locus is 3/32 and the variance in relationship from the formula of Guo (1996) is 0·00107, so $q = 88$.

The number of effective loci is similar to the recombination index for humans, assumed by Rasmusson

(1993) to be the number of independently segregating units in the genome.

(b) *Effective number of chromosome segments in a random mating population*

To derive the effective number of loci in a random mating population, consider two gametes taken at random from the population. The position at one end of a chromosome in both gametes can be traced back until they coalesce. Positions close to this first point will coalesce in the same ancestor but, as one moves along the chromosome, a recombination will be reached such that the next position coalesces in a different ancestor. Thus the two gametes can be seen to be composed of a series of short chromosome segments that coalesce. The average length of these segments is $1/(4N_e)$, where N_e is the effective population size (Stam, 1980). Therefore, the two gametes of length L Morgan are divided into $4N_eL$ segments. However, some segments are larger by chance than others so that if the effective number of segments is calculated from the variation in relationship between individuals in the population it is approximately $2N_eL/\log(4N_eL)$ per chromosome (Goddard, 2008). However, this approximation does not consider the fact that the small segments may still contain as many QTL mutations as the larger segments since they have on average a longer time to trace back to the same common ancestor and hence a longer time for mutations to accumulate. Therefore, the most appropriate value for the number of effective segments might be in between $4N_eL$ and $2N_eL/\log(4N_eL)$ per chromosome. As an approximation, we will assume that the effective number of loci is $2N_eL$ and then test the validity of this assumption with simulated data.

The variation in relationship between two gametes arises for two reasons. First, some pairs of gametes are more closely related by pedigree than others. For instance, some pairs of gametes may share a common parent or grandparent, whereas other pairs do not. Secondly, even considering pairs of gametes that have the same pedigree relationship, they may share more or less alleles than the average expected for that relationship, due to Mendelian sampling. The first source of variation in relationship is used by a conventional individual model BLUP to estimate the breeding values of individuals including those with no phenotypic record. The second source of variation in realized relationship is the source of the increase in reliability of GEBVs. For a pair of gametes with constant pedigree relationship, the ancestor in which one chromosome segment coalesces is independent of the ancestor in which another chromosome segment coalesces. Consequently, the variation in relationship due to this source would be zero if there were an infinite number of unlinked loci. Even though the number of positions in the genome may be very large, linkage causes variation in relationship by generating chromosome segments that coalesce. Since each segment coalesces independently conditional on the pedigree, it again seems appropriate to estimate the effective number of loci as in between the number of segments ($4N_eL$) and the number of segments weighted by length ($2N_eL/\log(4N_eL)$ per chromosome), e.g. as $2N_eL$.

(c) *Accuracy of genomic EBVs with information from relatives*

In what follows we will assume that fixed effects can be adequately estimated and that the data have been corrected for them, so $\mathbf{y} = \mathbf{Zg} + \mathbf{e}$, where $\mathbf{y}$ is corrected for fixed effects.

Even without any genetic markers, the breeding value of an individual can be estimated from pedigree and phenotypic records. We will focus on predicting the increase in reliability due to markers of the GEBV of an individual that has marker data but no phenotypic record and no offspring because that is the most important use of genomic selection. That is, we will calculate the increase in accuracy of GEBV above that obtainable simply from the pedigree and records on ancestors and collateral relatives. In this scenario, the breeding value of the ith individual (g_i) can be expressed as the mean of individuals with the same pedigree as individual i (f) and a deviation from that mean caused by the actual genes the individual inherited:

$$g_i = f + \sum_{j=1}^{q} u_{sij} + \sum_{i=1}^{q} u_{mij},$$

where f = family mean breeding value, u_{sij} = paternal allele effect inherited by the ith individual at the jth independent chromosome segment as a deviation from family mean, u_{mij} = maternal allele effect inherited by the ith individual at the jth independent chromosome segment as a deviation from the family mean and summation is over all independent chromosome segments.

The variance of the breeding values is then

$$V(\mathbf{g}) = V(f) + \sum_{j=1}^{q} V(\mathbf{u_{sj}}) + \sum_{j=1}^{q} V(\mathbf{u_{mj}}).$$

If we analyse the data $\mathbf{y}$ with the model and estimate f, $\mathbf{u_{sj}}$ and $\mathbf{u_{mj}}$:

$$\hat{\mathbf{g}} = \hat{f} + \sum_{j=1}^{q} \hat{\mathbf{u}}_{\mathbf{sj}} + \sum_{j=1}^{q} \hat{\mathbf{u}}_{\mathbf{mj}}.$$

The components of this equation are independent, as the effects of the sire and dam alleles are expressed

as deviations from the family mean. Therefore

$$V(\hat{\mathbf{g}}) = V(\hat{f}) + \sum_{j=1}^{q} V(\hat{\mathbf{u}}_{\mathbf{sj}}) + \sum_{j=1}^{q} V(\hat{\mathbf{u}}_{\mathbf{mj}}). \quad (3)$$

The reliability of GEBVs is $V(\hat{\mathbf{g}})/V(\mathbf{g})$ and the accuracy is the square root of this reliability. The reliability can be calculated using eqn (3) and compared with that obtained using only the family mean to quantify the increase in reliability due to the marker data.

With N progeny per family the reliability from phenotypic and pedigree information alone is

$$\frac{V(\hat{f})}{V(f)} = \frac{N}{N+\lambda_f},$$

where

$$\lambda_f = (V(\mathbf{y}) - V(f))/V(f).$$

Now we calculate the increase in reliability with genomic information. Assuming that there are n paternal alleles per family that are equally represented in the data so that there are N/n individuals carrying each paternal allele. As explained in the appendix $\lambda_s = 2q/h^2$. Then

$V(\hat{\mathbf{u}}_s) = \sigma_u^2(1-1/n)N/(N+n\lambda_s)$ (derived in the appendix)

$\sum_{j=1}^{q} V(\hat{\mathbf{u}}_{sj}) = 0{\cdot}5\, V(\mathbf{g})(1-1/n)N/(N+n\lambda_s)$ because

$V(\mathbf{g}) = 2q\sigma_u^2$

$\sum_{j=1}^{q} V(\hat{\mathbf{u}}_{dj})$ is calculated in a similar manner.

As an example consider the case of selecting among a population of individuals consisting of full sib families. In this case $V(f) = 0{\cdot}5\, V(\mathbf{g})$, $\lambda_f = (V(y) - V(f))/V(f) = (1-0{\cdot}5h^2)/(0{\cdot}5h^2)$ and $V(\hat{f}) = \frac{0{\cdot}5V(\mathbf{g})N}{N+\lambda_f}$. As above, for full-sibs, if the genome consists of a single chromosome 35-Morgan long, the variance in relationship among pairs of full sibs is 0·00179 corresponding to $q = 70$ effective chromosome segments. Consequently, $\sigma_u^2 = V(\mathbf{g})/(2*70)$, $\lambda_s = 2q/h^2 = 140/h^2$. Within a family of full-sibs there are two paternal alleles and two maternal alleles, so $n = 2$,

$$\sum_{j=1}^{q} V(\hat{\mathbf{u}}_{sj}) = 0{\cdot}5\, V(g)(1/2)N/(N+2\lambda_s)$$

and $\sum_{j=1}^{q} V(\hat{\mathbf{u}}_{dj})$ is the same. If $V(\mathbf{g}) = 1$, $N = 99$ and $h^2 = 0{\cdot}5$, then

$$V(\hat{f}) = 0{\cdot}5*99/(99+3) = 0{\cdot}486.$$

So from eqn (3)

$$V(\hat{\mathbf{g}}) = 0{\cdot}486 + 0{\cdot}5*(1/2)*99/(99+2*280) + 0{\cdot}5*(1/2)*99/(99+2*280) = 0{\cdot}561.$$

As $V(\mathbf{g}) = 1$, the reliability is 0·561.

(d) *Accuracy of GEBVs in a random breeding population*

We also want to predict the increase in reliability of an EBV using the realized relationship matrix compared with the pedigree relationship matrix in a random breeding population. As before, assume the breeding value is the sum of many QTLs each of which is in perfect Linkage disequilibrium (LD) with a marker. That is, the breeding value of individual i is $g_i = \sum_{j=1}^{q} w_{ij}u_j$ as before, except here it is convenient to express w_{ij} as a deviation from the mean, e.g. $w_{ij} = x_{ij} - 2p_j$, where x_{ij} is 0, 1 or 2 representing homozygote, heterozygote and other homozygote and p_j is the allele frequency at independent chromosome segment j and as before u_j is the allele substitution effect at the jth independent chromosome segment assuming there are only two alleles per independent chromosome segment, and q is the number of independently segregating chromosome segments, which for a randomly mating population is derived above. The phenotypes are modelled as in (1).

This derivation of accuracy of breeding value is similar to those for full-sib families but differs in an important way. In the full-sib case, each parent is assumed to have two different alleles at each independent chromosome segment, so the number of alleles at one independent chromosome segment is four times the number of families. Consequently, the genetic variance at one independent chromosome segment in the parental generation is $2\sigma_u^2$. However, in the case of a random mating population, there are assumed to be only two alleles per independent chromosome segment and the variance contributed by the jth independent chromosome segment is $V(w_j)\ \sigma_u^2 = 2p_j(1-p_j)\ \sigma_u^2$ and the total genetic variance is $\sigma_g^2 = qV(\boldsymbol{w})\sigma_u^2$, where $V(\boldsymbol{w})$ is the average value of $V(w_j)$ over all independent chromosome segment. In the full-sib case, we were estimating the effect of markers within a family and so only the number of individuals within the family could be used to estimate $\mathbf{u}$. On the other hand, the effective number of loci within a full-sib family is small because large segments of chromosome segregate within a family. By contrast in a random mating population we are estimating the effect of $\mathbf{u}$ across the population, so all individuals with phenotypes (N) can be used but the effective number of loci is large because there must be a marker close enough to any QTL to be in high LD with it.

There is assumed to be no LD between the QTLs so the BLUP equations for estimating $\mathbf{u}$ are approximately block diagonal with the jth independent chromosome segment having a block of equations

$$[\mathbf{W}'_j\mathbf{W}_j + \lambda]\ \hat{u}_j = \mathbf{W}'_j\mathbf{y}$$

or alternatively

$$NV(w_j)/[NV(w_j)+\lambda]\hat{u}_j = \sum_i^N w_{ij}y_i$$

and the reliability of $\hat{u}_i$ is $NV(w_j)/[NV(w_j)+\lambda]$, where N = total number of individuals with phenotypic records, $V(w_j)=2p_j(1-p_j)$ is the marker variance and $\lambda=\sigma_e^2/\sigma_u^2$. The value of σ_u^2 is calculated from the total genetic variance as $V(\boldsymbol{g})/(qV(\boldsymbol{w}))$, where q is the number of independent chromosome segments as described above for a random mating population and $V(\boldsymbol{w})$ is the average heterozygosity of all independent chromosome segments. If we assume that independent chromosome segments are neutral, then the distribution of allele frequency p is $f(p)=k/[2p(1-p)]$, where $k=1/\log(2N_e)$ assuming that p is bounded by $1/(2N_e)$ and $1-1/(2N_e)$, and $V(\boldsymbol{w})=k$ (Hill *et al.*, 2008). Thus $\lambda=\sigma_e^2/\sigma_u^2=qk/h^2$ assuming, as in the Appendix, that σ_e^2 is close to the phenotypic variance.

Goddard (2008) showed that the reliability of the EBV from many marker effects is a weighted average of these individual marker reliabilities where the weights are $V(w_j)$, i.e. the heterozygosity of each marker. Thus, the heterozygosity enters in two ways – the greater the $V(\mathbf{w})$, the greater the reliability of the marker effect and the greater the weight. Assuming the distribution of allele frequencies is that predicted by the neutral model, Goddard (2008) gives the following formula for reliability of GEBVs calculated as the sum of the marker effects:

$$\begin{aligned}\text{Reliability of GEBV} &= [1-\lambda/(2N\sqrt{a}) \\ &\quad *\log((1+a+2\sqrt{a})/(1+a-2\sqrt{a}))],\end{aligned} \tag{4}$$

where $a=1+2\lambda/N$. We used this formula to calculate the reliability of GEBV with N_e of 25, 50, 100 or 250, with 1000 or 2500 phenotypic records.

For instance, consider the case where $N=1000$, $N_e=25$, $h^2=0{\cdot}5$ and $L=29$.

Then $k=1/\log(2N_e)=0{\cdot}256$, $q=2N_eL=1450$, $\lambda=qk/h^2=741$, $a=1+2\lambda/N=2{\cdot}48$. The reliability of GEBV in unphenotyped individuals is then

$$\begin{aligned}&[1-\lambda/(2N\sqrt{a})*\log((1+a+2\sqrt{a})/(1+a-2\sqrt{a}))] \\ &\quad=0{\cdot}29.\end{aligned}$$

(iii) *Simulation approaches to verify analytical predictions of accuracy of breeding values*

To verify the analytical approach to predicting accuracy, we have used two simulation approaches. The first approach calculates accuracy from the PEV of breeding values derived from the left-hand side of the mixed model equations, with realizations of the relationship matrix from simulation.

Given model (2) above and assuming that $\mathbf{b}$ is known, then $\text{BLP}(\mathbf{g})=\mathbf{AZ'V^{-1}(y-Xb)}$, where $\mathbf{A}$ is the predicted relationship matrix and $\mathbf{V}=\mathbf{ZAZ'}\sigma_g^2+\mathbf{I}\sigma_e^2$.

$$V(\hat{\mathbf{g}})=\mathbf{AZ'V^{-1}ZA},$$

$$\mathbf{PEV}=V(\mathbf{g})-V(\hat{\mathbf{g}})=\mathbf{A}-\mathbf{AZ'V^{-1}ZA}=\mathbf{A(I}-\mathbf{Z'V^{-1}ZA}).$$

If the variance components such as the additive genetic variance are known, then we can predict the accuracy of selection for every individual in the pedigree file.

With marker data we can estimate the actual relationship matrix $\mathbf{G}$, with $E(\mathbf{G})=\mathbf{A}$. Hence $\mathbf{G}$ is a random 'variable'. We are interested in the expected value of PEV over repeated samples of $\mathbf{G}$ given the pedigree, $\text{BLP}(\mathbf{g}|\mathbf{G})=\mathbf{GZ'V_G^{-1}(y-Xb)}$ where $\mathbf{V_G}$ is $\mathbf{V}$ above using $\mathbf{G}$ in place of $\mathbf{A}$, and $E_{\mathbf{G}}[V(\hat{\mathbf{g}}|\mathbf{G})]=E[\mathbf{GZ'V_G^{-1}ZG}]$.

The samples of $\mathbf{G}$ can be generated using simulation. The program Merlin (Abecasis *et al.*, 2002) was used to gene-drop whole genomes, and from these simulations $\mathbf{G}$ and $V(\hat{\mathbf{g}}|\mathbf{G})$ were estimated. We repeated these simulations 100 times for 10 relatives and 10 times for 100 or more relatives to obtain an expected PEV for each individual in the pedigree and therefore the expected increase in accuracy of selection by using the realized relationship matrix and compared the expected PEV with the PEV for $\mathbf{G}=\mathbf{A}$.

The pedigrees simulated were identical to those used in the analytical approach above, i.e. full sibs, half sibs or double first cousins. Marker genotypes at each locus were simulated with a number of equifrequent alleles. The limit of Merlin is 63 alleles/locus; so, the frequency at each locus was 1/63. Elements of $\mathbf{G}$ were genome-wide identical by state sharing statistics calculated by simply averaging identity by states (IBS) across the loci. For each pair of individuals and each locus, IBS was calculated as $\text{IBS}=\frac{1}{2}\sum\sum\delta_{jk}$, with δ_{jk} an indicator variable which is 1 if allele j ($j=1, 2$) in the first individual is identical to allele k ($k=1, 2$) in the second individual. Similarly for inbreeding/homozygosity, for each locus and for each individual, homozygosity identical by state (HIBS) = δ, with $\delta=1$ if the two alleles are identical. These statistics were averaged over all 3500 markers (spaced at 1 cM intervals). To obtain an unbiased prediction of the genome-wide IBD sharing, the following adjustments were made:

$$\pi_{g(i,j)}=[\text{mean(IBS)}_{ij}-2/m]/(1-1/m),$$

$$f_{g(i)}=[\text{mean(HIBS)}_i-1/m]/(1-1/m),$$

where m is the number of alleles per marker (=63), $\pi_{g(i,j)}$ is the estimate of genome-wide IBD sharing between individuals i and j, $f_{g(i)}$ the estimate of the genome-wide inbreeding coefficient of individuals i, and the mean statistics are averaged over all 3500 markers. The adjustment is to account for the

expected values of unrelated individuals. The definition of IBS here was 0·5 * (IBS of all the four possible 4 allelic comparisons) between two individuals. Then if the allele frequency of allele i is p_i, then $P(\text{IBS}|\text{unrelated}) = 2\sum_{i-1}^{m} p_i^2$. If the alleles are equifrequent, as is the case here, then $P(\text{IBS}|\text{unrelated}) = 2m/m^2 = 2/m$. In order to get an unbiased estimated, the term is then divided by $1-1/m$.

Of the N progeny simulated, $N-1$ had a phenotype. The accuracies of breeding value were for individuals without phenotypes.

The value of $V(\hat{\mathbf{g}})/V(\mathbf{g})$ was the reliability of breeding values.

Note that for the scenario simulated (a single chromosome of 3500 cM in length), the genome-wide SD in identity for full sibs, half sibs and double first cousins are approximately 0·042, 0·030 and 0·033, respectively.

This simulation is based on the PEV expected from theory but no QTLs were included in the simulation. The second simulation method simulates QTL and markers to investigate the effect of the number of markers, the number of QTLs and the effective population size on the increase in accuracy as a result of using the realized relationship matrix and to further verify the analytical approach above.

To create a population in equilibrium between mutation, drift and recombination, we simulated a population of $N_e = 1000$ individuals with random mating for 6000 generations. Each individual in the population consisted of 29 pairs of chromosomes, and was either male or female (probability 0·5). Each chromosome was 1 M long, and had 345 marker loci and 340 QTLs. To create an offspring, a pair of parents of different sex was randomly chosen from the population. For each parent in a mating pair, a gamete was formed from its chromosome pairs by sampling the number of crossovers for each chromosome pair from a Poisson distribution, with a mean of 1. Crossover points were randomly positioned along chromosome pairs. The haploid gametes were mutated at a rate that gave an average final marker heterozygosity of 0·32. The mutation rate was adjusted to ensure this heterozygosity using the formula $u = H/4N_e$ where H was the desired heterozygosity (chosen as 0·32, as this is similar to the heterozygosity of Single Nucleotide Polymorphism (SNP) from whole genome association experiments in some species, e.g. Hayes *et al*., 2007) and N_e was the effective population size simulated. The mutation rate at the QTL was altered to give 10, 100 or 1000 segregating QTLs. If a locus was mutated, a new allele was added. If the locus was a QTL, the effect of the new QTL allele on the quantitative trait following mutation was sampled from a gamma distribution (scale = 5·4 and shape = 0·42), and with an equal probability of favourable or unfavourable effect, as described by Hayes & Goddard (2001). The genetic value of individual i was $g_i = \sum_{j=1}^{q} u_{sij} + \sum_{i=1}^{q} u_{mij}$, where u_{sij} is the effect of the paternal allele inherited by progeny i at QTL j, and u_{mij} is the effect of the maternal allele inherited by progeny i at QTL j.

To measure reliability in full sib families, in generation 6000, 10 males and 10 females were mated to generate 10 full sib families, each of 200 individuals. Phenotypes were created for 100 individuals in each full sib family by adding a random normal variate to g_i, to give the desired level of heritability.

Breeding values for both phenotyped and non-phenotyped individuals were then predicted by solving the equations for model (1) above:

$$[\hat{\mathbf{g}}] = [\mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\sigma_g^2]^{-1}[\mathbf{Z}'\mathbf{y}], \quad (5)$$

where $\mathbf{Z}$ is matrix allocating records to phenotypes, $\mathbf{y}$ is a vector of phenotypic records, $\mathbf{G}$ is the realized relationship matrix calculated as above, and σ_g^2 is the additive genetic variance. $\mathbf{G}$ was calculated using approximately 100, 1000, 2500, 5000 or 10 000 equally spaced markers. Reliabilities of breeding values for the un-phenotyped progeny were the square of the correlation between their true and predicted breeding values. Results are averages of 10 replicates.

A second set of simulations were performed to assess the effect of changing N_e on the accuracy of predicting breeding values for un-phenotyped individuals drawn at random from a population. In these simulations, the N_e for 6000 generations was 25, 100 or 250. Heterozygosity of markers was maintained at approximately 0·32 for each value of N_e by adjusting the marker mutation rate. The 1000 individuals were assigned phenotypes as above. A further eight generations of breeding with random mating was then performed, so that the differences between individuals in ancestry eight generations ago, when phenotypes were collected, were small. The idea here was to separate the contribution of linkage and linkage disequilibrium to the reliability of the breeding values. Breeding values were predicted with (5) for individuals in generation 6009 using phenotypes and genotypes from individuals in generation 6001. Breeding values were also calculated for individuals in generation 6002 in some simulations.

3. Results

(i) *Prediction of GEBV when the pedigree is known and there are an infinite number of QTLs*

The agreement between accuracy of GEBV calculated by the analytical method and from simulation was generally excellent (Table 1). Excluding cases where $h^2 = 1$, the average difference between the predicted and observed reliabilities was 0·004, and the maximum difference was 0·023. Reliability increased with

Table 1. *Comparison of analytical* (*in bold*) *and simulated results for reliability of GEBVs calculated using either predicted or realized relationship matrices*

Design	h^2	N progeny	Reliability		Ratio (SE)
			A-matrix	G-matrix	
Full sibs	1·0	10	0·450	0·480	1·067 (0·006)
			0·454	**0·466**	
		100	0·495	0·741	1·496 (0·013)
			0·495	**0·625**	
		1000	0·499	0·968	1·939 (0·002)
			0·499	**0·890**	
	0·50	10	0·374	0·384	1·025 (0·005)
			0·385	**0·383**	
		100	0·486	0·582	1·201 (0·009)
			0·485	**0·560**	
		1000	0·498	0·816	1·535 (0·009)
			0·499	**0·819**	
	0·10	10	0·161	0·163	1·011 (0·006)
			0·161	**0·162**	
		100	0·419	0·438	1·042 (0·002)
			0·419	**0·437**	
		1000	0·491	0·619	1·261 (0·008)
			0·491	**0·622**	
Half sibs	1·0	10	0·188	0·194	1·036 (0·008)
			0·189	**0·195**	
		100	0·243	0·320	1·320 (0·004)
			0·243	**0·308**	
		1000	0·249	0·433	1·736 (0·014)
			0·249	**0·444**	
	0·50	10	0·140	0·146	1·039 (0·008)
			0·147	**0·145**	
		100	0·234	0·272	1·168 (0·003)
			0·234	**0·271**	
		1000	0·248	0·392	1·577 (0·010)
			0·248	**0·408**	
	0·10	10	0·047	0·048	1·021 (0·008)
			0·047	**0·048**	
		100	0·179	0·188	1·046 (0·003)
			0·179	**0·188**	
		1000	0·241	0·300	1·249 (0·007)
			0·241	**0·306**	
Double first cousins	1·0	10	0·188	0·200	1·068 (0·010)
			0·189	**0·200**	
		100	0·243	0·356	1·469 (0·006)
			0·243	**0·336**	
		1000	0·249	0·783	3·141 (0·015)
			0·249	**0·691**	
	0·50	10	0·140	0·146	1·038 (0·008)
			0·140	**0·145**	
		100	0·234	0·284	1·216 (0·004)
			0·234	**0·283**	
		1000	0·248	0·538	2·169 (0·020)
			0·248	**0·561**	
	0·10	10	0·047	0·048	1·019 (0·008)
			0·047	**0·048**	
		100	0·179	0·189	1·054 (0·003)
			0·179	**0·190**	
		1000	**0·240**	**0·335**	1·396 (0·008)

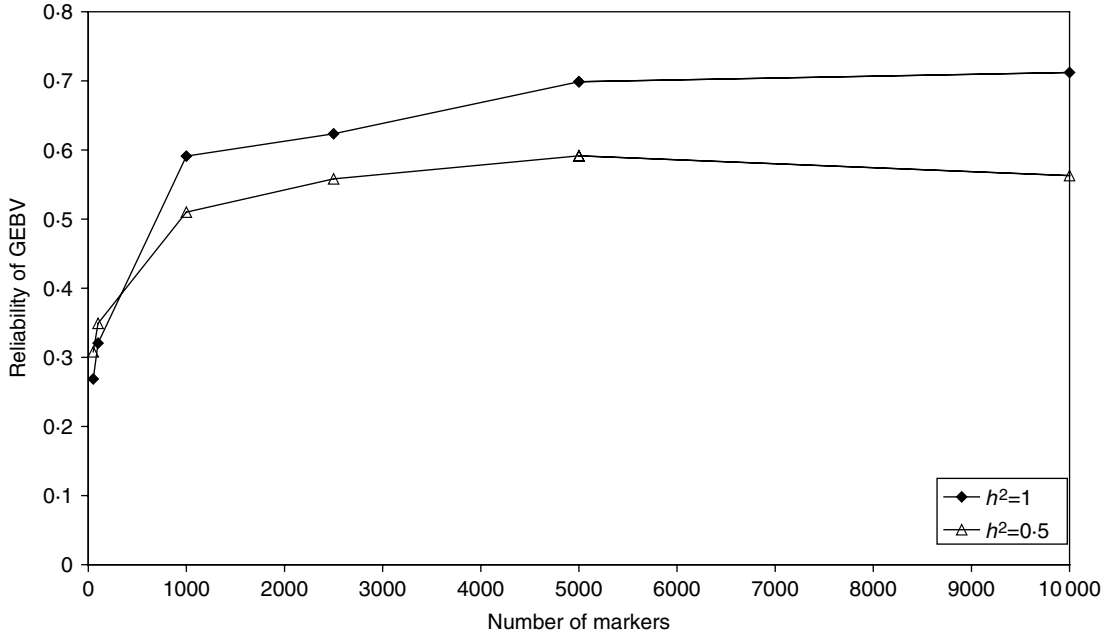

Fig. 1. Reliabilities with increasing number of markers in simulations for a full sib family of 100 individuals.

increasing number of family members and heritability within each type of family. This is expected because with more family members the effects of alleles within that family are estimated more accurately and as heritability increases there is less environmental noise affecting the phenotypic value. For full-sibs and double first cousins, the reliabilities approach 1·0 if family size is high enough. For example, when the heritability of the trait was 0·5, and there were 1000 full sibs per family, the predicted reliability of GEBV was 0·819 and the reliability observed in the simulations was 0·816.

The reliabilities are high in these situations because there are only two paternal and two maternal alleles segregating within a full-sib family and four of each within a double first cousin family. Consequently, if there are enough records, the effect of these alleles can be estimated accurately. For half-sib families, the effect of paternal alleles can be estimated accurately but the effect of the maternal alleles cannot be estimated at all because each new half sib carries a new maternal allele. Therefore, the reliabilities for half-sibs approach are 0·5.

When the heritability of the trait was one, the agreement between the analytical predictions and simulation results was less consistent. Particularly as the number of individuals per family became larger, the analytical prediction tended to be lower than the simulated results.

(ii) *Prediction of GEBV when the pedigree is known and there are a finite number of QTLs*

In the second set of simulations, the markers simulated were similar to SNPs in the number of alleles and level of polymorphism, segregating in full sib families with 100 individuals per family with phenotypic records. A large number of markers were required to achieve the reliabilities predicted by the analytical method, regardless of heritability (Fig. 1). With 5000 markers, reliability of GEBV for unphenotyped individuals was very close to that predicted by the analytical method.

Changing the number of QTLs affecting the quantitative trait did not alter the reliabilities of GEBV that could be predicted for un-phenotyped full sibs, provided the number of markers was large (Fig. 2). However, reliabilities were high in the case of 10 QTLs even when relatively few markers were used. With low numbers of QTLs, the effects of large segments of chromosome will be the same even with some level of recombination, as many recombination events will not add new QTL alleles to the chromosome segment. So even if these recombination events cannot be detected as a result of using the limited number of markers, provided the number of QTLs is very low the effects of many IBS segments that are not IBD will actually be zero, and this will have little impact on accuracy.

(iii) *Prediction of GEBV in random populations*

In the random mating population, the deterministic prediction of reliability was always higher than that from the simulations (Table 2). In part this is because the simulation measured the reliability of eight generations after the phenotypic data were collected. Consequently, there have been eight generations of recombination between markers and the QTL before the prediction was tested. This would reduce the

Table 2. *Effect of N_e on reliability of GEBV in random mating population*

Effective population size (N_e)	Number of records	Effective number of chromosome segments	h^2	Reliability	
				Simulated	Predicted
25	1000	1450	0·5	0·24	0·29
			0·2	0·075	0·14
50	1000	2900	0·5	0·16	0·21
			0·2	0·058	0·09
100	1000	5800	0·5	0·083	0·13
			0·2	0·042	0·06
250	2500	14 500	0·5	0·09	0·15

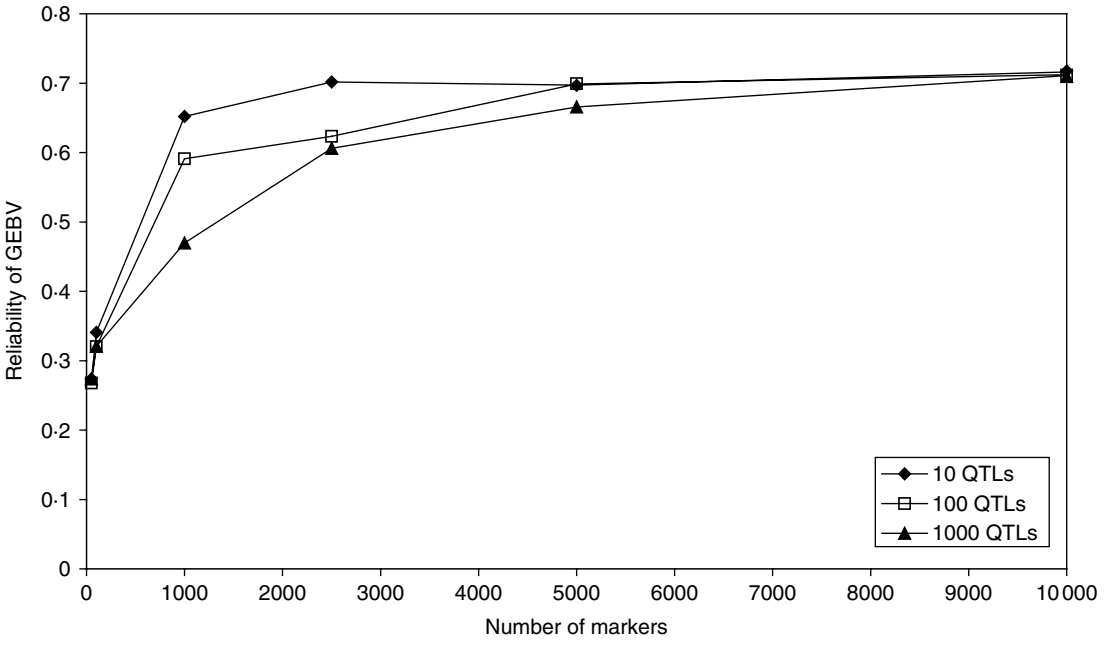

Fig. 2. Effect of different numbers of QTLs on the reliability of GEBV in full sib families.

reliability by 1–2 % per generation and when this is allowed for the simulated and predicted reliabilities agree more closely.

For comparison, when the reliability of GEBV was calculated for individuals in the generation subsequent to collecting phenotypes, the reliabilities were higher than the predictions. For example, with $N_e = 25$, the reliability of breeding values in generation 6002 was 0·44, compared with the predicted value of 0·29. The high reliability in generation 6002 is likely to reflect the differences in pedigree between individuals, the contribution of linkage between the markers with the QTL, as well as linkage disequilibrium between the markers and the QTL.

It is also possible that the reliability in the simulations would have been greater had more than 10 000 markers been used since 5000 markers were needed for maximum reliability even in full-sib families. However, the derivation of the reliability used an inexact value for the number of independent chromosome segments and this may be the reason why the simulated reliability is less than that predicted. Despite the over-prediction of reliability by the prediction method, it is still useful because it explains the variables affecting reliability and gives a good guide when N_e is large and simulation becomes very slow.

The results show that increasing N_e decreases reliability because it increases the number of effective chromosome segments whose effects must be estimated. The more effective the segments, the smaller are their effects and the more records are required to estimate them accurately. The formula and the results in Table 2 show that the accuracy remains almost constant if N/N_e remains constant.

Decreasing the heritability also decreases the reliability because the ratio of phenotypic variance to QTL variance increases, just as it does if the number of effective chromosome segments is increased.

4. Discussion

In this paper, we have presented a model equivalent to the BLUP model suggested by Meuwissen *et al.* (2001) that uses the realized relationship matrix rather than a design matrix to allocate chromosome segment effects to individuals. If the number of markers being used to estimate breeding values is much larger than the number of individuals, a major advantage of our equivalent model is the reduction in the number of parameters that must be estimated from the number of markers to the number of individuals. As the models are equivalent there is no reduction in the accuracy of predicting breeding values.

One extension of our equivalent model leads to an analytical prediction of the accuracy of breeding values that can be achieved for individuals without phenotypes but belonging to kinships such as full sibs, half sibs or double first cousins. Provided the heritability of the trait was less than one, the agreement between the analytical predictions and results from simulations were very good.

The analytic method underestimates the reliability when $h^2=1$ because in this situation the approximation used for the residual variance fails. In the appendix we assumed that, when the effects at one QTL are being estimated, the residual variance is not decreased by the estimated effects of all other QTLs. That is, we assumed that the residual variance was that which would occur if the effect of each QTL was estimated by itself without fitting all other QTLs in the model. This approximation works well except when $h^2=1$. In this situation there is in theory no residual variance and the estimated effects are accurate enough to decrease the residual variance and consequently increase the reliability above that calculated. Since traits with $h^2=1$ are rare, this should not be a severe limitation in practice.

The analytical method predicts that breeding values for individuals within families can be predicted with close to 100% accuracy. However, large numbers of phenotypic records are required to achieve this. For example with $h^2=1$, 1000 phenotypes from individuals in a full sib family would allow the GEBV for another member of the family without a phenotype to be predicted with an accuracy of 0·94. Villanueva *et al.* (2005), using simulation, also demonstrated using the realized relationship matrix rather than the predicted relationship matrix in the calculation of breeding values could lead to higher accuracies of selection. However, the gains they achieved as a result of using the realized relationship matrix were modest compared with those achieved here, because they used relatively small family sizes and a limited number of markers. In practice, the ability to exploit the variation in the proportion of the genome, which is IBD between individuals in kinships such as full sibs to increase the accuracy of breeding values, will be limited by the number of full sibs or kindreds with phenotypes required in the prediction. The number of markers required is also large, in the order of 2500. While such large full or even half sib families are rare in most species, they do occur in aquaculture and plant species.

A second extension of our model leads to an analytical prediction of the accuracy of GEBV in randomly mating populations. Here, the accuracy of breeding values depends on the effective population size, the number of phenotypic records, the heritability and the number of markers. If N_e is large, the number of independent chromosome segments is also large. This means that the extent of LD will be limited in the population, so a very large number of markers are required to capture the effects of the QTL. Further, a very large number of markers and phenotypic records will be required to predict breeding values for unphenotyped individuals with any accuracy, particularly if the heritability is low (Fig. 3). In humans, the effective population size is very large (approximately 10 000) and the extent of LD is very limited (e.g. Dunning *et al.*, 2000; Reich *et al.*, 2001; Tenesa *et al.*, 2007). Our analytical approach suggests that very large numbers of phenotypic records would be required to predict breeding values for unphenotyped individuals with any accuracy in the human population.

In livestock populations, the effective population sizes can be as low as 100 (e.g. Holstein Friesian cattle, Riquet *et al.*, 1999). In this situation, the analytical approach predicts that accurate breeding values can be predicted with thousands of records rather than hundreds of thousands, provided the heritability of the trait is high. However, in some livestock populations, the number of independent chromosome segments is likely to be larger than what the current effective population would predict because the effective population size has been larger in the past (e.g. Hayes *et al.*, 2003).

An alternative to predicting the number of independent chromosome segments from effective population size would be to use very dense SNP data to infer the number of independent segments directly. For example, The International HapMap Consortium (2007) genotyped three human populations for 3·3 million SNPs. They estimated the number of 'haplotype blocks' in the human genome, where a haplotype block contains SNPs in very high LD within the block, but reduced LD between blocks. These haplotype blocks are similar in concept to our

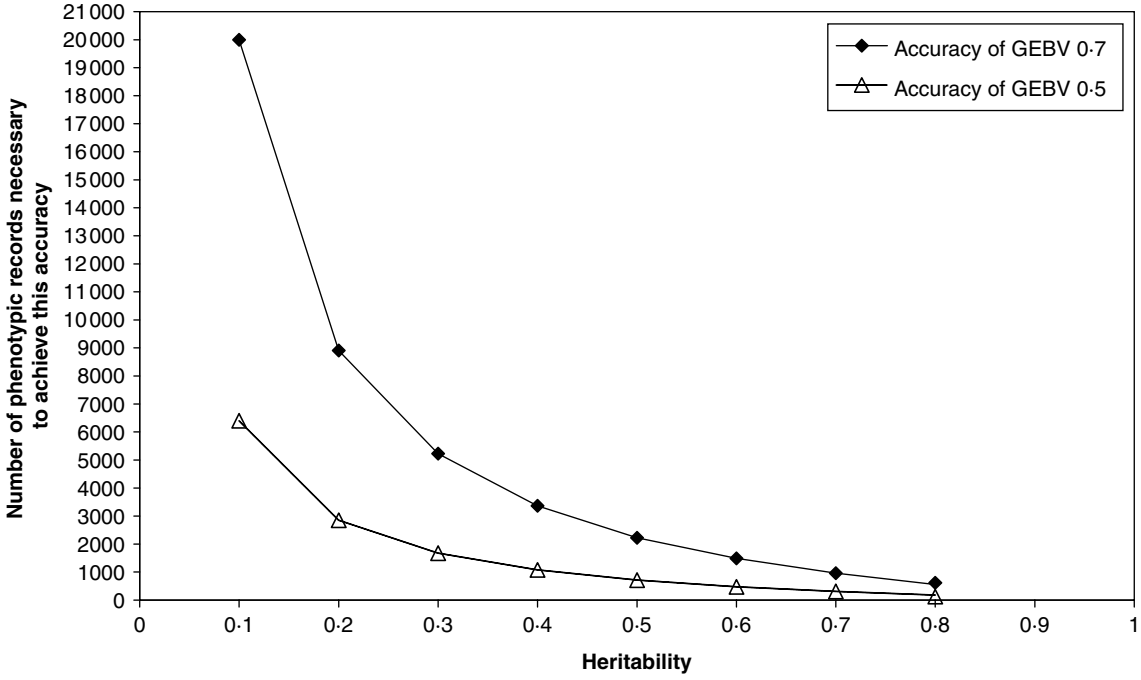

Fig. 3. Number of phenotypic records required to achieve a desired accuracy of GEBV, 0·5 or 0·7, given the heritability of the trait. Effective population size (N_e) = 1000 and a normal distribution of QTL effects assumed.

independent chromosome segments. In the European and Chinese plus Japanese population, the estimate of the number of 'haplotype blocks' was approximately 300 000, while in the Nigerian population there were more than 600 000 haplotype blocks. Again, such large numbers imply that very large numbers of phenotypic records would be required to predict breeding values for unphenotyped individuals with any accuracy in the human population.

In many species, there are some moderately large families and so the use of the realized relationship matrix would use both linkage based on families and LD across the whole population. This can also be achieved by the 'LDLA' method of Meuwissen & Goddard (2004). As the pedigree is traced further and further back in time, the linkage analysis becomes indistinguishable from the LDLA method.

Our analytical method assumes that all independent chromosome segments have an effect on the trait. If the number of QTLs affecting a trait is substantially less than the effective number of chromosome segments, then EBVs of higher accuracy can be obtained by methods that assume that only some markers have QTLs associated with them (Meuwissen *et al.*, 2001; Habier *et al.*, 2007; Wray *et al.*, 2007; Goddard, 2008). These methods of analysis can be described in terms of model selection (which markers are included in the model) or in terms of a prior distribution of marker effects, which contains a large number of markers with zero effect. If the assumption that there are many chromosome segments with zero effect is true, utilizing this assumption leads to higher EBV reliabilities (Goddard, 2008). Our analytical predictions will under-estimate the reliabilities achievable in this case. However, if the number of QTLs approaches the number of chromosome segments, then it is best to fit models such as those used in this paper and to acknowledge that very large datasets will be needed.

The analytical methods described here will be a useful tool for designing experiments where the aim is to predict either GEBVs or phenotypes where gene action is additive, from dense marker data. Given a desired level of accuracy of predicting GEBV for individuals with genotypes only, our analytical method determines the number of individuals that must be phenotyped and genotyped in order to achieve this level of accuracy. For example, if the heritability of the trait is 0·3, and $N_e = 1000$, approximately 5750 individuals with genotypes and phenotypes are required in order to predict GEBVs of un-phenotyped individuals in the same population with an accuracy of 0·7. Further work could extend our analytical method to simultaneously consider linkage and LD in the prediction of accuracy of GEBV, and account for imperfect marker coverage.

Appendix. Variance of estimated QTL effects within family

The reliability of estimating the effect of individual alleles can be derived from the BLUP equations.

Ignoring fixed effects, the model for an individual record is

$$y_i = f + \sum_{j=1}^{q} \left(w_{sij} u_{sij} + w_{dij} u_{dij} \right) + e_i,$$

where, as in the main text, f = the family mean, w_j indicates the allele at the jth QTL with sub-script s or d for sire and dam alleles and $u_{s\_}$ and $u_{d\_}$ are the effects of the sire and dam alleles.

This results in a large set of equations but many of the terms are approximately independent. For instance, the alleles at one effective QTL (w_j) are inherited independently of the alleles at other effective QTL (e.g. w_{j+1}), given our definition of QTL as independent loci. This means that the equations are approximately block diagonal. In the case of a family of half-sibs, for instance, there are two paternal alleles and approximately half the offspring (N/n, where $n=2$) in the family will receive each one. Therefore, we will approximate the complete set of equations with the equations for one QTL and assuming that all alleles are equally represented. In matrix notation, the model becomes

$$\mathbf{y} = f\mathbf{1_n} + \mathbf{Wu} + \mathbf{e} \qquad \text{(A1)}$$

and the mixed model equations, treating f as fixed temporarily, are

$$\begin{bmatrix} \mathbf{1_n}'\mathbf{1_n} & \mathbf{1_n}'\mathbf{W} \\ \mathbf{W}'\mathbf{1_n} & \mathbf{W}'\mathbf{W} + \lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} f \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}.$$

When the terms are evaluated, the left-hand side matrix becomes

$$\begin{bmatrix} N & \mathbf{1_n}'N/n \\ N/n\mathbf{1_n} & (N/n + \lambda)\mathbf{I} \end{bmatrix}.$$

The inverse of this matrix is

$$\begin{bmatrix} 1/N + 1/(n\lambda) & -\mathbf{1_n}'/(n\lambda) \\ -1/(n\lambda) & (\mathbf{I} + J(N/n^2\lambda)(n/N + n\lambda) \end{bmatrix}.$$

The PEV of $\hat{u}$ can be obtained from this inverse matrix. The estimates of u are used for selection within family, so we require the PEV of $u - \bar{u}$. This is

$$\lambda V(u)\ (n-1)/(N + n\lambda).$$

Using the variance of the true $u - \bar{u} = \sigma_u^2(n-1)/n$, the variance of $\hat{u} - \bar{u}$ is

$$\sigma_u^2\ (n-1)/n * N/(N + n\lambda).$$

Note that except for $(n-1)/n$, which corrects for selection within only n possible alleles, this is the normal formula for reliability of a BLUP solution based on N/n records for each effect to be estimated, i.e. $(N/n)/(N/n + \lambda)$.

In the full equation set, the residual variance is $V(y)(1-h^2)$ because all the genetic variances are included in the model. However, the full equations will never be exactly balanced across all terms, and so the PEV will be greater than that calculated above if λ were based on this residual. We have found that the PEV are approximated better if we use the error in eqn (A1), which is the phenotypic variance minus the variance explained by one QTL allele. The variance explained by one QTL is very small and $2qV(u) = V(g)$, so $\lambda = V(y)/V(u) = 1/(h^2/(2q)) = 2q/h^2$.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C. F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D. & Ponder, B. A. J. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *American Journal of Human Genetics* **67**, 1544–1554.

Franklin, I. R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology* **11**, 60–80.

Fernando, R. L. (1998). Some theoretical aspects of finite locus models. Proceedings of the 6th World Congress of Genetics Applied to Livestock Production, 11–16 January 1998, University of New England, Armidale, Australia, volume 26 (1998), pp. 329–336.

Goddard, M. E. (2008). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* Epub ahead of print. PMID: 18704696.

Guo, S. W. (1996). Variation in genetic identity among relatives. *Human Heredity* **46**, 61–70.

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

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., Visscher, P. M., McPartlan, H. & Goddard, M. E. (2003). A novel multi-locus measure of linkage disequilibrium and it use to estimate past effective population size. *Genome Research* **13**, 635.

Hayes, B. J., Chamberlain, A. C., McPartlan, H., McLeod, I., Sethuraman, L. & Goddard, M. E. (2007). Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. *Genetics Research* **89**, 215–220.

Hill, W. G. (1993). Variation in genetic identity within kinships. *Heredity* **71**, 652–653.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**(2), e1000008. doi: 10.1371/journal.pgen.1000008.

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. & Goddard, M. E. (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* **36**(3), 261–279.

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

Rasmusson, M. (1993). Variation in genetic identity within kinships. *Heredity* **70**, 266–268.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjlan, R., Farhadian, S. F., Ward, R. & Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.

Riquet, J., Coppieters, W., Cambisano, N., Arranz, J. J., Berzi, P., Davis, S. K., Grisart, B., Farnir, F., Karim, L., Mni, M., Simon, P., Taylor, J. F., Vanmanshoven, P., Wagenaar, D., Womack, J. E. & Georges, M. (1999). Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. *Genetics* **96**, 9252–9257.

Sanna, S., Jackson, A. U., Nagaraja, R., Willer, C. J., Chen, W. M., Bonnycastle, L. L., Shen, H., Timpson, N., Lettre, G., Usala, G., Chines, P. S., Stringham, H. M., Scott, L. J., Dei, M., Lai, S., Albai, G., Crisponi, L., Naitza, S., Doheny, K. F., Pugh, E. W., Ben-Shlomo, Y., Ebrahim, S., Lawlor, D. A., Bergman, R. N., Watanabe, R. M., Uda, M., Tuomilehto, J., Coresh, J., Hirschhorn, J. N., Shuldiner, A. R., Schlessinger, D., Collins, F. S., Davey Smith, G., Boerwinkle, E., Cao, A., Boehnke, M., Abecasis, G. R. & Mohlke, K. L. (2008). Common variants in the GDF5-UQCC region are associated with variation in human height. *Nature Genetics* **40**, 198–203.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research* **35**, 131–155.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. & Visscher, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Research* **17**, 520–526.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3·1 million SNPs. *Nature* **449**(7164), 851–861.

Van Raden, P. M. (2007). Efficient estimation of breeding values from dense genomic data. *Journal of Dairy Science* **90**(Suppl. 1), 374–375.

Villanueva, B., Pong-Wong, R., Fernández, J. & Toro, M. A. (2005). Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of Animal Science* **83**, 1747–1752.

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics* **40**, 489–490.

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2**, e41.

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