Estimation of Genomic Breeding Values

Izel Fourie Sørensen, Palle Duun Rohde & Peter Sørensen

2022-03-10

Contents

Learning objective:

This section introduces the basic concepts of estimating breeding values such as:

- basic principle behind estimating genomic breeding values
- accuracy of estimated genomic breeding values
- use of genomic relationships for estimating breeding values
- different methods, data sources and experimental designs for estimating genomic breeding values

1 Introduction

A new technology called **genomic selection** has revolutionized animal and plant breeding. Genomic selection refers to selection decisions based on genomic breeding values (GEBVs). We have previously learned how phenotypic records and genetic relationships computed from pedigree information can be used to estimate breeding values (EBVs). Genome-wide DNA markers can replace or supplement pedigree information for this purpose. The first ideas of genomic prediction were presented by (Meuwissen2001a). They showed that information from genotypes of very many marker loci evenly spread over the complete genome can successfully be used to estimate genomic breeding values. Because the information of the genotypes is spread over the complete genome it is often referred to as **genomic information** and from the use of this information for selection purposes the term of **genomic selection** was invented. The early results on genomic selection were not considered until the paper by (Schaeffer2006) showed that in a cattle breeding program the introduction of genomic selection could lead to savings in about 90% of the total costs, provided that the accuracies computed by (Meuwissen2001a) can really be achieved. After the publication of (Schaeffer2006) many animal and plant breeding organisation started to introduce procedures of genomic selection.

2 Genomic information

In recent years much attention has been given to genomic information due to the dramatic development in genotyping technologies. Today dense genetic maps are available for most of the most important animal and plant species (Table 1). It is, however, still lacking for several species, but they circumvent this by using a so-called RAD-sequencing (Restriction site associated DNA sequencing) which enable dividing the entire genome into smaller segments just like if a genetic map had been available. Ultimately the entire DNA sequence may be genotyped. This is possible, but still very expensive, so only a few founder individual have been fully sequenced (mostly bulls, but also some horses and dogs). Lower resolution maps are also used and especially in cattle to save costs (e.g. females).

Table 1. Number of markers in currently available SNP chips (SHOULD BE UPDATED)

Species	No. SNPs (in thousands)	Genome size (x10^9)
Cattle	778	2,67
Pig	64	2,81
Chicken	581	1,05
Horse	70	2,47
Sheep	54	2,62
Dog	170	2,41

The genetic maps are based on DNA markers in the form of single nucleotide polymorphisms (SNP) and they enable us to divide the entire genome into thousands of relatively small chromosome segments.

2.1 Genetic markers

The single location in the genome that are considered in GS are called **markers**. When looking at the complete set of markers consisting the genomic information in a population, the so-called **Single Nucleotide Polymorphisms** (SNP) have been shown to be the most useful types of markers. These SNP correspond to differences of single bases at a given position in the genome. Based on empirical analyses of very many SNP-loci, almost all SNP just take two different states. Furthermore it is important that these SNPs are more or less evenly spread over the complete genome. Some SNPs are in coding regions and some my be placed in regions of unknown functionality. Figure 1 shows the distribution of SNP over the genome.

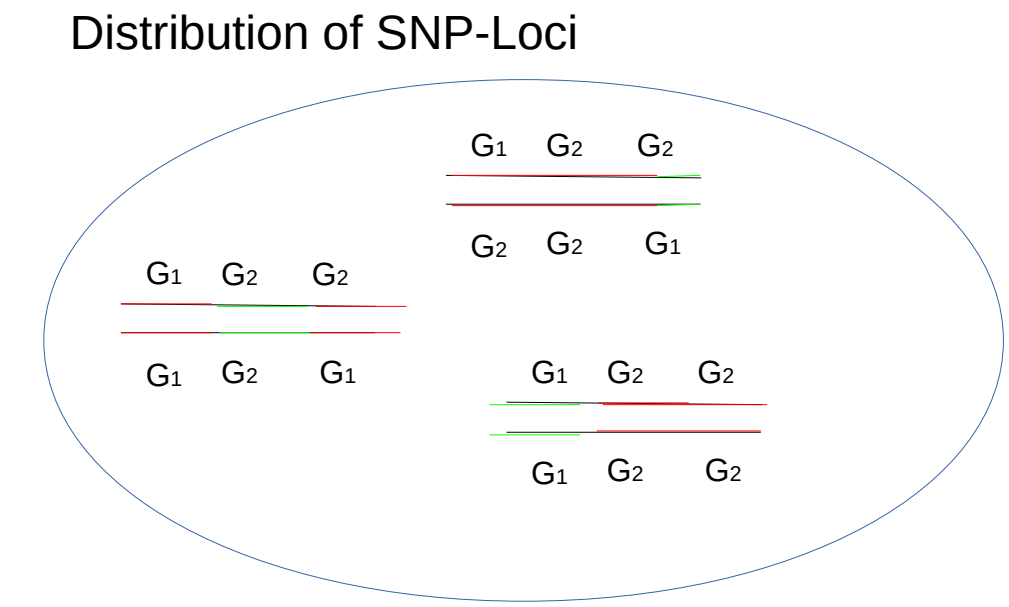

Figure 1: Distribution of SNP-Loci Across A Genome

2.2 Quantitative Trait Loci and linkage disequilibrium

The loci that are relevant for a quantitative traits are called **Quantitative Trait Loci** (QTL). Any given SNP-Marker can only be informative for a given QTL, if a certain **linkage disequilibrium** between the QTL and the marker locus exists. The idea behind this linkage disequilibrium is that a certain positive QTL-allele evolved in a certain genetic neighborhood of a number of SNP loci. As a result of that the positive QTL-allele is very often inherited with the same SNP-allele. Over the generations, recombination between the QTL and the neighboring SNP-loci can happen and thereby weaken the association between the positive QTL-allele and the given SNP-allele. This recombination effect is smaller when the QTL and the SNP-loci are physically closer together on the chromosome. The non-random association between QTL and SNP-markers is called linkage disequilibrium.

The marker locus is called M and the QTL is called Q, then the LD can be measured by

$$D = p(M_1Q_1) * p(M_2Q_2) - p(M_1Q_2) * p(M_2Q_1) \tag{1}$$

where $p(M_xQ_y)$ corresponds to the frequency of the combination of marker allele M_x and QTL allele Q_y. Very often the LD measure shown in (1) is re-scaled to the interval between 0 and 1 which leads to

$$r^2 = \frac{D^2}{p(M_1) * p(M_2) * p(Q_1) * p(Q_2)} \tag{2}$$

In (2) r^2 describes the proportion of the variance at the QTL which is explained by the marker M. Hence the LD must be high such that the marker can explain a large part of the variance at the QTL. For the length of most livestock species, about $50'000$ SNP markers are required to get a sufficient coverage of the complete genome.

3 Basic principles for estimating genomic breeding values (GEBV)

The GEBV are calculated as the sum of the effects of dense genetic markers across the entire genome, thereby potentially capturing all the quantitative trait loci (QTL) that contribute to variation in a trait. The QTL effects inferred from individual single nucleotide polymorphism (SNP) markers, are first estimated in a large reference population with phenotypic information. In subsequent generations, only marker information is required to calculate GEBV. Breeding values are conceptually simple to calculate from marker information. First, the entire genome is divided into small chromosome segments by dense markers. Second, the additive effects of each chromosome segment are estimated simultaneously. Finally, the genomic EBV equals the sum of all chromosome segment effects. The chromosome segment effects are estimated for a group of individuals (i.e. a reference population). For any remaining individual, only a blood or tissue sample is needed to determine its genome-wide (or genomic) EBV (Figure 1). For breeding purposes, it is desirable that the EBV can be estimated accurately early in the individuals life. The effect of each of these small chromosome segments can be estimated if we have phenotypes and genotypes from a number of individuals. With sufficiently dense marker maps, the chromosome segment effects apply to all individuals in the population in which they were estimated, because markers are in linkage disequilibrium with the causal gene that they bracket.

3.1 A linear mixed model for estimating marker effects (MBLUP)

A linear model to estimate marker effects based on the data from the reference population in the two-step procedure can be defined as follows

$$y = X\beta + Mg + e \tag{3}$$

where
- m number of SNP markers
- y vector of observations
- β vector of fixed effects
- X design matrix linking fixed effects to observations
- g random genetic effect of SNP-genotypes
- M design matrix linking SNP-genotype effects to observations
- e vector of random residuals

The observations y used in (3) are in most evaluations not phenotypes but traditionally predicted breeding values with an accuracy above a certain threshold. As a consequence of that the variance-covariance matrix (R) of the residuals e is not just an identity matrix (I) times a residual variance component (σ_e^2) but R is a diagonal matrix with elements $(R)_{ii} = \frac{1}{B_m} - 1$ where B_m is the accuracy of the traditionally predicted breeding value from an individual from the reference population, corrected for the parental contributions. In effect, B_m corresponds to the accuracy of the mendelian sampling term.

The mixed-model equations resulting from models given in (3) have the following structure

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} M \\ M^T R^{-1} X & M^T R^{-1} M + I * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ M^T R^{-1} y \end{bmatrix} \tag{4}$$

where

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} \sum_{i=1}^{m} 2 * p_i * (1 - p_i) \tag{5}$$

In (5) σ_a^2 is the total genetic variance and p_i is the frequency of the SNP-allele that is associated with the positive QTL-allele.

The solutions for $\hat{g}$ from (4) correspond to the SNP-genotype effects. The predicted breeding value $\hat{a}$ for any selection candidate with genomic information is then computed as

$$\hat{a} = \sum_{i=1}^{m} M_i \hat{g}_i \tag{6}$$

where M_i corresponds to the vector of SNP-genotypes of the selection candidate. This model is sometimes referred to the SNP-BLUP model.

3.1.1 Matrix M

The elements in matrix M can be encoded in different ways. The results from the genotyping laboratory sends a code representing the nucleotide that can be found at a given position. For the use in the linear model we have to use a different encoding. Let us assume that at a given SNP-position, the bases G or C are observed and G corresponds to the allele with the positive effect on our trait of interest. Based on the two observed alleles, the possible genotypes are GG, GC or CC. One possible code for this SNP in the matrix M might be the number of G-Alleles which corresponds to 2, 1 and 0. Alternatively, it is also possible to use the codes 1, 0 and -1 instead which corresponds to the factors with which a is multiplied to get the genotypic values in the single locus model.

3.2 A linear mixed model for estimating genomic values (GBLUP)

The term `GBLUP` stands for genomic BLUP and is the most widely used single-step procedure. In GBLUP genomic breed values are directly predicted without the prediction of marker effects. This can be done by including the genomic breeding values a which corresponds to the sum of all SNP-allele effects directly as a random effect in the model.

$$y = X\beta + Za + e \tag{7}$$

where Z is the design matrix linking genomic breeding values to observations. The mixed model equations are the defined as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{8}$$

where G is defined as in (9) and λ is the same as in equation (5). Several authors have shown that both procedures (two-step and single step) are equivalent. From (7) we can see that the GBLUP model looks very similar to the linear mixed model used in pedigree based breeding value estimation (PBLUP), except that the covariances between random effects based on the numerator relationship matrix and in GBLUP they are modeled via the genomic relationship matrix G. This means in the PBLUP model the covariance between random breeding values is based on the concept of common ancestry and identity by descent. This is replaced in GBLUP by the concept of sharing the same alleles based on identity by state which is assumed to be the cause of the covariance between random genomic breeding values.

The predicted genomic breeding values $\hat{a}$ coming out of (8) are referred to as `genomic estimated breeding values` (GEBV).

3.3 Genomic Relationship Matrix G

Multiplying the matrix M with its transpose M^T results in a $n \times n$ square matrix MM^T. On the diagonal of this matrix we get counts of how many alleles in each individual have a positive effect. The off-diagonal elements count how many individual share the same alleles across all SNP-positions. In contrast to the additive genetic relationship matrix A, the counts here are based on identity by state and not on identity by descent.

The problem with matrix MM^T is its dependence on the number SNP-markers. Therefore the matrix MM^T is proportional to the relationship A but it does not correspond to A directly. As a solution to that problem (VanRaden2008) proposed to re-scale such that allele frequencies on a given locus are expressed as to times the deviation from 0.5. This re-scaling is done with an $n \times m$ matrix P where each of the m columns corresponds to a SNP-Locus. Elements in column i of matrix P have all the same value corresponding to $2p_i - 0.5$ where p_i corresponds to the frequency of the SNP-allele associated to the positive QTL-allele at locus i. The difference between matrices M and P is assigned to a new matrix W

$$W = M - P$$

Finally the matrix WW^T must be scaled with the sum of $2p_i(1-p_i)$ over all SNP-loci to get to the genomic relationship matrix G.

$$G = \frac{WW^T}{\sum_{i=1}^{m} 2p_i(1-p_i)} \tag{9}$$

The matrix G has similar properties as the numerator relationship matrix A. The genomic inbreeding coefficient F_j is defined as $F_j = (G)_{jj} - 1$. The genomic relationship a_{ij} between two individuals i and j corresponds to the element in matrix G divided by the square root of the diagonal elements

$$a_{ij} = \frac{G_{ij}}{\sqrt{G_{ii}G_{jj}}}$$

3.4 How Does GBLUP Work?

The genomic relationship matrix G allows to predict genomic breeding values for individuals with marker genotypes without any observation in the dataset. This fact is the basis of the large benefit of genomic selection. As soon as a DNA is available fpr an individual, its marker genotypes can be determined and a genomic breeding value can be predicted. This genomic breeding value is much more accurate then the traditional breeding value based only on ancestral information.

The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (10). In this form the Inverse G^{-1} of G and the vector $\hat{g}$ of predicted genotypic breeding values are split into one part corresponding to the individuals with observations and a second part for the individuals without phenotypic information.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{10}$$

The matrix $G^{(11)}$ denotes the part of G^{-1} corresponding to the individuals with phenotypic observations. Similarly, $G^{(22)}$ stands for the part of the individuals without genotypic observations. The matrices $G^{(12)}$ and $G^{(21)}$ are the parts of G^{-1} which link the two groups of individuals. The same partitioning holds for the vector of predicted breeding values. The vector $\hat{g}_1$ contains the predicted breeding values for the individuals with observations and the vector $\hat{g}_2$ contains the predicted breeding values of all individuals without phenotypic observations.

Based on the last line of (10) the predicted breeding values $\hat{g}_2$ of all individuals without phenotypic observations can be computed from the predicted breeding values $\hat{g}_1$ from the individuals with observations.

$$\hat{g}_2 = -\left(G^{22}\right)^{-1} G^{21}\hat{g}_1 \tag{11}$$

Equation (11) is referred to as genomic regression of predicted breeding values of individuals without observation on the predicted genomic breeding values of individuals with observations.

3.5 Accuracy of GEBV

The accuracy of GEBV is trait dependent and depends on heritability and the number of phenotypic records. The accuracy of GEBV increases when the size of the reference population increases, when the reference population represents as much of the relevant genetic variation in the population as possible, when selection candidates are closely related to the reference population, when genetic diversity in the population is low (i.e. low effective population size) and with better statistical models.

A common finding is that a straightforward BLUP method for estimating the marker effects gave reliabilities of GEBV almost as high as more complex methods. The BLUP method is attractive because the only prior information required is the additive genetic variance of the trait.

More informative marker maps also increase accuracy, although the increase here is marginal when the marker density is already high (i.e. 50,000 markers for within breed selection in dairy cattle; advantageous with more markers for very heterogeneous populations, across-breed analyses).

3.6 Practical Problems

The model equations (7) look very straight-forward, but the practical implementation can be quite complicated. The reason for these problems is the fact that compared to the total size of a population only a small fraction of all individuals are genotyped and hence contribute the the genomic evaluation. On the other hand DGV do not contain all information that occur in conventional breeding values.

Because all non-genotyped offsprings of parents are ignored by GBLUP, this loss of information is even more dramatic. For the two step-procedure as long as the reference population has a reasonable size and is not too heterogeneous, this is not a problem, we can still come up with reasonable estimates of SNP-effects. Due to the in-balanced availability of genotypic information, a procedure to combine DGV with traditional predicted breeding values was adopted. This procedure starts with predicting DGV and combining them with traditionally predicted breeding values from parents which are termed as parent averages (PA). This procedure of combining predicted breeding values from different sources is called **blending**. The problem with blending one has to be aware of is that there is a covariance between DGV and PA which must be accounted for.

A further problem is that there are different techniques to generate genotyping results. The different results also have different densities which means that they give different numbers of SNP-loci per genome. The different techniques also vary in price which is the reason that genotyping results from different technologies must be combined. Combining genotyping results with different densities of SNP-markers per genome is done with a process that is called **imputing**. This basically comes done to inferring missing SNP-genotypes on marker panels with less density based on results from denser marker panels.

4 Impact of genomic selection on breeding programmes

The impact of using genomic information depends on the efficiency of traditional breeding that does not use genomic information. If all selection candidates already have accurate EBV at the time of selection then genomic information will not add much, if anything. Hence, genome-wide marker information is most useful when phenotypic recording is restricted – for instance when phenotypes are expressed late in the animal's life (e.g. meat quality, longevity), are expressed only in one sex (e.g. milk yield, egg production) or are expensive to measure (e.g. feed efficiency, bacteriological samples, progesterone profiles or other physiological measures). Furthermore, genome-wide marker information is useful for traits with low heritability provided a sufficient amount of phenotypes can be recorded. It should therefore be clear that the extra benefits of genome-wide information vary across traits and across species although it can in principle be used for all species and traits.

Dairy cattle breeding is characterised by the main traits only being measurable in females while very intense selection is only possible in males. Thus genome-wide markers are very useful in dairy cattle breeding.

In pig breeding, most traits are measured on all selection candidates before sexual maturity. Therefore genomic information gives less extra value for pig breeding compared with dairy cattle. However, exceptions for pigs include litter size (only measurable in females and after sexual maturity), feed efficiency (only measured on few animals because it is expensive), longevity and carcass traits (expressed late). Another potential benefit for pigs is that traits can be recorded on crossbreed production animals which may be housed in different production environments and have different effects of single genes compared with purebred animals.

GEBV can be used to enhance the screening of potential breeding individuals for testing (pre-selection) which is especially attractive in situations when the costs of genotyping are relatively inexpensive compared to the costs of recording phenotypes.

GEBV are also useful in intensifying selection for young animals thereby facilitating a reduced overall generation interval. For instance, with the availability of accurate genome-wide breeding values for young bulls it is more attractive to use the best young bulls widely rather than having to wait for the results of progeny

group testing. Here the substantial reduction in generation interval offset the slightly lower accuracy of genome-wide breeding values compared with breeding values based on progeny results.

Another benefit of using GEBV rather than traditional EBV is that it results in less inbreeding if the same selection intensities are maintained. This happens because breeding based on traditional EBV puts more emphasis than GEBV on parent information, especially for traits with low heritability.

A potential danger with GEBV is that it does not capture the effect of new non-recurrent mutations in selection candidates without phenotypic information (relating to self or progeny). Thus if selection and mating decisions are made before there is phenotypic information available from progeny, or the individual itself, it becomes impossible to estimate the effect of a new mutation. This means that new unfavourable mutations may be easier spread in the population and that favourable mutations may be missed if selection is based entirely on GEBV with negative consequences for long term genetic progress.

Also, often some traits that should be in the breeding goal are not systematically recorded (e.g. many diseases). But such traits are still influenced by selection on other traits and frequently the combined correlated effect on such non-recorded traits is negative. With selection based on GEBV there is a risk that the negative pressure on non-recorded traits increases. So, although genomic breeding values offers exiting opportunities for enhancing genetic progress by allowing for accurate selection of individuals then they should be used with appropriate care.