Estimation of Genomic Breeding Values

Guillaume Ramstein & Peter Sørensen

2022-03-18

Contents

Learning objective:

This section introduces the basic concepts of estimating genomic 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 estimated 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 genomic marker data 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 coined. 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 of 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 for 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 this lack of genomic information can be circumvented by using RAD-sequencing (Restriction site associated DNA sequencing) or Genotyping-by-Sequencing (GBS), which enable dividing the entire genome into smaller segments. Ultimately the entire genome may be sequenced. This is possible, but still very expensive, so only a few founder individual have been fully sequenced (in animals, mostly bulls, but also some horses and dogs; in plants, dozens of accessions in major crops like maize, rice, barley, and wheat). Lower resolution maps are also used and especially in cattle to save costs (e.g. females).

Table 1. Number of markers used for genomic selection in different species

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
Maize	*	2.3
Wheat	*	17.1
Rice	*	0.38
Barley	*	5.3
Tomato	*	0.83

*: Number of SNPs vary greatly depending on the assay: from about 6,000 to 900,000 in SNP arrays, to several millions in whole-genome sequencing data.

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 Genomic markers

The different locations in the genome that are considered in genomic selection are called **markers**. When looking at the complete set of markers making up the genomic information in a population, the so-called **Single Nucleotide Polymorphisms** (SNPs) have been shown to be the most useful types of markers. These SNPs 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 may be located in coding regions and some my be placed in regions of unknown function. Figure 1 illustrates the distribution of SNP over the genome.

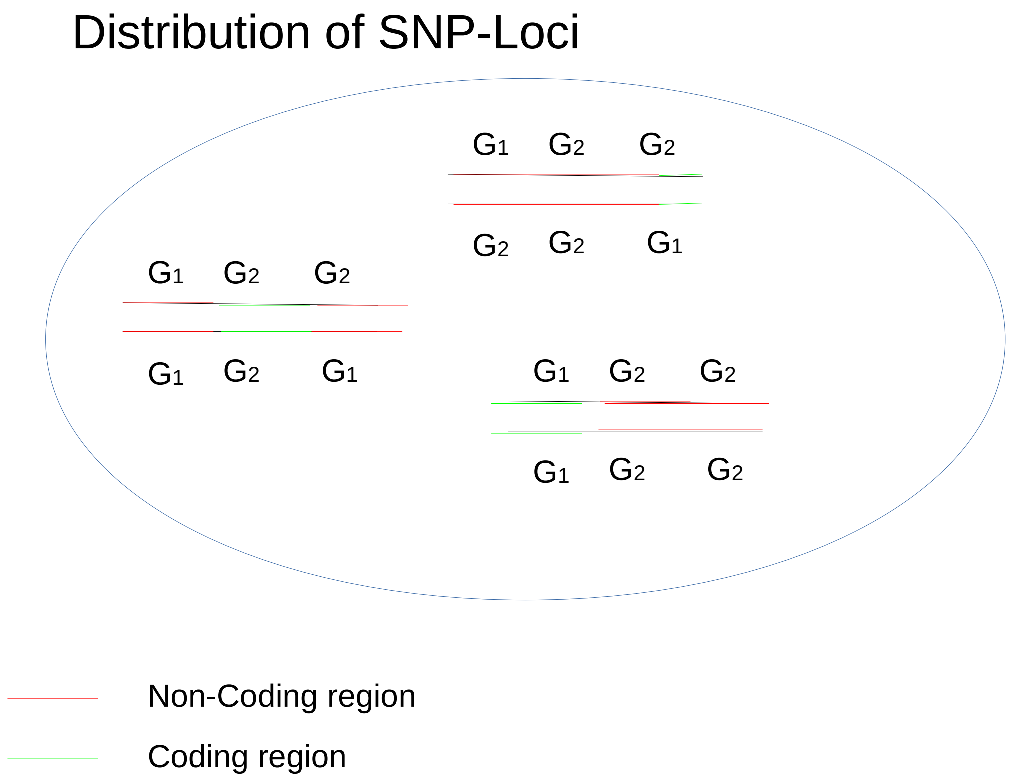

Figure 1: Distribution of SNP-Loci Across A Genome

2.2 Quantitative Trait Loci and linkage disequilibrium

The loci that are relevant for a quantitative trait 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 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 statistical 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 genome size of most animal and plant 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

Genomic 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 estimated breeding value (GEBV) is calculated as 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) breeding values. For breeding purposes, it is desirable that the GEBV can be estimated accurately, and early in the individual's life. The effect of each of these small chromosome segments can be estimated if we have phenotypes and genotypes from many individuals (from several hundreds to hundreds of thousands or even millions). With sufficiently dense marker maps, the chromosome segment effects capture the genomic variability in the population in which they were estimated, because markers are in linkage disequilibrium with the causal gene that they bracket.

We will present two approaches that are commonly used to estimate genomic breeding values. The first approach is referred to as MBLUP (or SNP-BLUP). In this approach marker effects are estimated from observed phenotypic and genomic marker data recorded in the reference (or training) population. Genomic breeding values are estimated from the marker effects and genomic marker data for potential selection candidates in the breeding population. The second approach is referred to as GBLUP. In this approach genomic breeding values are estimated from observed phenotypic and genomic marker data for individuals in the reference (or training) population. Genomic breeding values for selection candidates are estimated based on their **genomic relationship** to individuals in the reference population. Both approaches allow for estimation of genomic breeding values for individuals without phenotypes and close relationships. This is one of the main advantages the genomic prediction and selection. As soon as DNA is available for an individual, its marker genotypes can be determined and a genomic breeding value can be estimated. Furthermore, GEBVs are generally more accurate than the traditional estimated breeding values (EBVs) based only on pedigree information.

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

The linear mixed model for estimating marker effects contains the observation vector for the trait(s) of interest (y), the fixed effects b, which explain systematic differences in y, the random marker effects s, and the random residual effects e. A matrix formulation of a general linear mixed model for estimating marker effects is:

$$y = Xb + Ms + e \tag{3}$$

where

y : is the vector of observed values of the trait,

X : is a known design matrix that relates the elements of b to their corresponding element in y.

b : is a vector of fixed effects,

M : is a known design matrix that relates the elements of s to their corresponding element in y.

s : is a vector of random marker effects,

e : is a vector of random residual effects,

In the linear mixed model above the marker and residual effects (s and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal (MVN) distribution. In general terms the distributions of these random variables are:

$$\begin{aligned} s &\sim MVN(0, S) \\ e &\sim MVN(0, R) \\ y &\sim MVN(Xb, V) \end{aligned}$$

where $S = I_s\sigma_s^2$ is a square matrix of (co)variances among marker effects (usually assumed independent), and $R = I\sigma_e^2$ is a square matrix of residual (co)variances among residuals (also assumed independent, most of the times), and $V = MM'\sigma_s^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix.

The marker variance σ_s^2 is defined as:

$$\sigma_s^2 = \frac{\sigma_a^2}{\sum_{j=1}^{m} 2p_j(1-p_j)} \tag{4}$$

where σ_a^2 is the total additive genetic variance, m is the number of markers, and p_j is the frequency of the marker-allele that is associated with the positive QTL-allele.

3.1.1 Estimation of marker effects in MBLUP

Estimates of the fixed effect b, and random marker effects, s, in the linear mixed model specified above can be done using the BLUE and BLUP equations shown below.

The best linear unbiased estimator (BLUE) of the fixed effects $\hat{b}$ is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{5}$$

The best linear unbiased prediction (BLUP) of the marker effects $\hat{s}$ is:

$$\hat{s} = SM'V^{-1}(y - X\hat{b}) \tag{6}$$

The BLUP equation for the estimate of the marker effects consists of three parts; The term, $y - X\hat{b}$, shows that the observed phenotypic values are corrected for the fixed effects represented by $X\hat{b}$. The covariance between the true marker effects (s) and phenotypes (y) is $Cov(s, y) = SM' = M'\sigma_s^2$. The inverse of the phenotypic covariance matrix is $[Var(y)]^{-1} = V^{-1}$. Alternatively, estimates of the (fixed and random) effects in the model can be obtained by solving the mixed model equations. The mixed-model equations for the model given in (3) have the following structure:

$$\left[\begin{array}{cc} X^TR^{-1}X & X^TR^{-1}M \\ M^TR^{-1}X & M^TR^{-1}M + S^{-1} \end{array}\right] \left[\begin{array}{c} \hat{b} \\ \hat{s} \end{array}\right] = \left[\begin{array}{c} X^TR^{-1}y \\ M^TR^{-1}y \end{array}\right] \tag{7}$$

3.1.2 Estimation of genomic breeding values in MBLUP

The estimates of the marker effects $\hat{s}$ in (7) can be used to estimate genomic breeding values $\hat{a}$ for any individual with genomic information (i.e. genotypes for the same set of markers used in the reference population) by:

$$\hat{a} = \sum_{j=1}^{m} M_j \hat{s}_j \tag{8}$$

where M_j corresponds to the vector of observed marker genotypes of an individual.

3.1.3 Encoding of the marker genotype matrix M

The elements in the matrix M can be encoded in different ways. The results from the genotyping laboratory represents the nucleotides found at a given genome position. To be used in the linear model the nucleotides (genotypes) at each position (marker locus) must be encoded numerically. 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 breeding values (GBLUP)

The linear mixed model for estimating genomic breeding values (GBLUP) contains the observation vector for the trait(s) of interest (y), the fixed effects b that explain systematic differences in y, and the random genomic effects a and random residual effects e. A matrix formulation of a general GBLUP model is:

$$y = Xb + Za + e \tag{9}$$

where

$$
\begin{aligned}
y &: \text{is the vector of observed values of the trait,}\\
b &: \text{is a vector of fixed effects,}\\
a &: \text{is a vector of random genomic effects,}\\
e &: \text{is a vector of random residual effects,}\\
X &: \text{is a known design matrix that relates the elements of } b \text{ to their corresponding element in } y.\\
Z &: \text{is a known design matrix that relates the elements of } a \text{ to their corresponding element in } y.
\end{aligned}
$$

In the linear mixed model (specified above) the genomic and residual effects (a and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal (MVN) distribution. In general terms the distributions of these random variables are:

$$
\begin{aligned}
a &\sim MVN(0, \tilde{G})\\
e &\sim MVN(0, R)\\
y &\sim MVN(Xb, V)
\end{aligned}
$$

where $\tilde{G} = G\sigma_a^2$, and $R = I\sigma_e^2$ are square matrices of genomic and residual (co)variances among the individuals, respectively, and $V = G\sigma_a^2 + I\sigma_e^2$ is the overall phenotypic covariance matrix. The genomic relationship matrix G is estimated from genomic marker data instead of pedigree information.

3.2.1 Estimation of genomic effects in GBLUP

Estimates of the fixed effect b, and random genomic effects, a, in the linear mixed model specified above can be done using the BLUE and BLUP equations shown below.

The best linear unbiased estimator (BLUE) of the fixed effects $\hat{b}$ is:

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y \tag{10}$$

The best linear unbiased prediction (BLUP) of the genomic effects $\hat{a}$ is:

$$\hat{a} = \tilde{G}Z'V^{-1}(y - X\hat{b}) \tag{11}$$

The BLUP equation for the estimate of the genomic effects consists of three parts; The term, $y - X\hat{b}$, shows that the observed phenotypic values are corrected for the fixed effects represented by $X\hat{b}$. The covariance between the true genomic effects (a) and phenotypes (y) is $Cov(a, y) = \tilde{G}Z' = GZ'\sigma_a^2$. The inverse of the phenotypic covariance matrix is $[Var(y)]^{-1} = V^{-1}$. Alternatively, estimates of the (fixed and random) effects in the model can be obtained by solving the mixed model equations which have the following structure:

$$\left[\begin{array}{lr} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + \tilde{G}^{-1} \end{array}\right] \left[\begin{array}{c} \hat{b} \\ \hat{a} \end{array}\right] = \left[\begin{array}{c} X^TR^{-1}y \\ Z^TR^{-1}y \end{array}\right] \tag{12}$$

From (11) we can see that the GBLUP estimation procedure looks very similar to the pedigree based breeding value estimation procedure (PBLUP). In GLUP the covariances between breeding values is based on the genomic relationship matrix G which is computed from genomic markers whereas in PBLUP it is based on the numerator relationship matrix A computed from pedigree information.

3.2.2 Estimation of genomic breeding values for individuals without phenotypes in GBLUP

The estimates of the genomic breeding values $\hat{a}_1$ for individuals with phenotypes can be used to estimate genomic breeding values $\hat{a}_2$ for individuals with only genomic information. Under the assumption of multivariate normality for the true genomic breeding values ($a \sim MVN(0, G\sigma_a^2)$), the expected value of the estimated genomic breeding value for individuals without phenotypes (a_2) conditional on the genomic breeding values for individuals with phenotypes $\hat{a}_1$ can be written as:

$$\hat{a}_2 = \tilde{G}_{12}\tilde{G}_{11}^{-1}\hat{a}_1 \tag{13}$$

The equation given in (13) consists of three parts; The term, $\hat{a}_1$, represent the genomic breeding values for individuals with phenotypes in the reference population. The covariance between the true breeding values for invididuals without phenotypes (a_2) and individuals with phenotypes (a_1) is $Cov(a_1, a_2) = \tilde{G}_{12} = G_{12}\sigma_a^2$. The inverse of the genomic covariance matrix for individuals with phenotypes is $\tilde{G}_{11}^{-1} = \sigma_a^{-2}G_{11}^{-1}$.

3.2.3 Genomic Relationship Matrix G

The additive genomic relationship matrix G is constructed using all genomic markers as follows:

$$G = \frac{WW^T}{\sum_{i=1}^{m} 2p_i(1-p_i)} \tag{14}$$

where W is the centered and scaled genotype matrix, and m is the total number of markers. Each column vector of W was calculated as follows: $w_i = M_i - 2p_i - 0.5$, where p_i is the minor allele frequency of the i'th genomic marker and M_i is the i'th column vector of the allele count matrix, M, which contains the genotypes coded as 0, 1 or 2 counting the number of minor allele. The centering of the allele counts and scaling factor $\sum_{i=1}^{m} 2p_i(1-p_i)$ ensures that the genomic relationship matrix G has similar properties as the numerator relationship matrix A.

The main difference between the two types of genetic relationship matrices (A and G) is that A is based on the concept of identity by descent (sharing of the same alleles, transmitted from common ancestors) whereas G is based on the concept of identity by state (sharing of the same alleles, regardless of their origin).

4 Impact of genomic prediction and selection on breeding programmes

The relative benefit of using genomic information depends on the efficiency of traditional breeding that does not use genomic information. If all selection candidates already have accurate EBVs 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 in animals; grain yield in plants) or are expensive to measure (e.g., feed efficiency, bacteriological samples, progesterone profiles or other physiological measures in animals; metabolic and physiologic measures in plants). 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 genomic 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.

Genomic breeding values 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.

Genomic breeding values 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 GEBVs compared with EBVs based on progeny results.

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