Chapter 9

Genomic Selection

Similarly to BLUP, the principles of **Genomic Selection** (GS) was proposed quite a while before its introduction in 2008. The first ideas of GS were presented by (Meuwissen et al., 2001a). They showed that information from genotypes of very many loci evenly spread over the complete genome can successfully be used for the purposes of livestock breeding. 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 GS were not considered until the paper by (Schaeffer, 2006) showed that in a cattle breeding program the introduction of GS could lead to savings in about 90% of the total costs, provided that the accuracies computed by (Meuwissen et al., 2001a) can really be achieved. After the publication of (Schaeffer, 2006) many livestock breeding organisation started to introduce procedures of GS.

9.1 Background

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 9.1 shows the distribution of SNP over the genome.

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

$$D = p(M_1Q_1) * p(M_2Q_2) - p(M_1Q_2) * p(M_2Q_1) \tag{9.1}$$

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 (9.1) is re-scaled to the interval between 0 and 1 which leads to

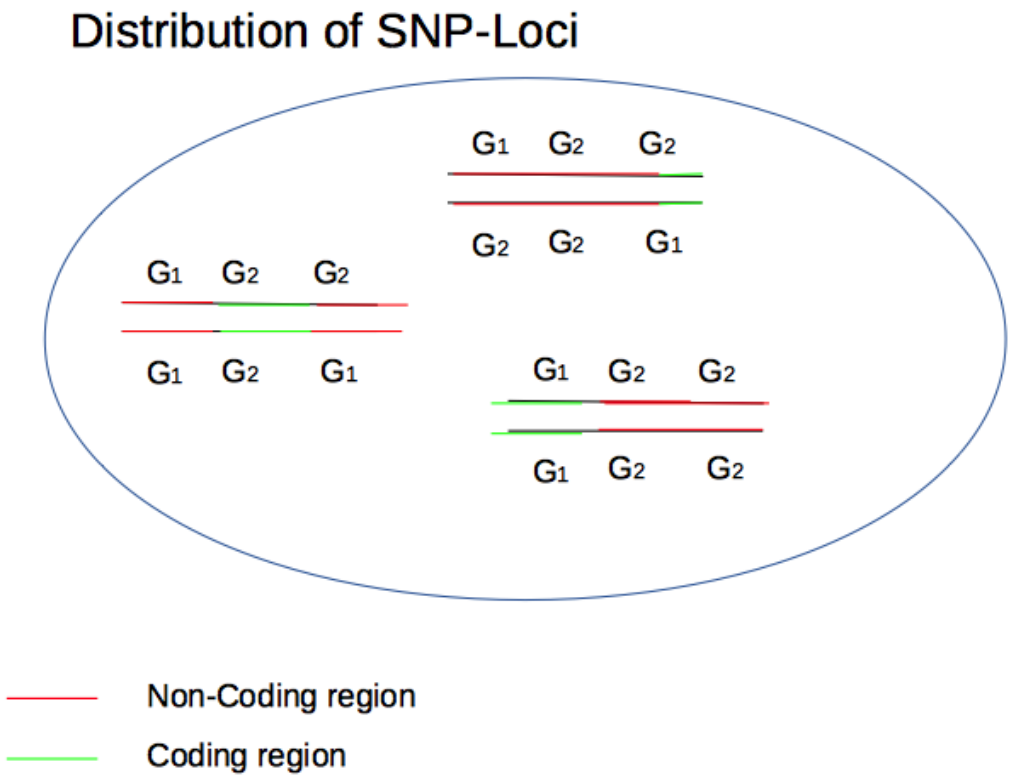

Figure 9.1: Distribution of SNP-Loci Across A Genome

$$r^2 = \frac{D^2}{p(M_1) * p(M_2) * p(Q_1) * p(Q_2)} \tag{9.2}$$

In (9.2) 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.

Nowadays the term `genomic selection` is often used ambiguously. What most people mean when they are talking about GS should better be called **genomic prediction** of breeding values. This prediction can be done in different ways which are listed below

- Two-step procedure: Effects of SNPs are predicted using single locus models in a reference population which corresponds of mainly male breeding animals with predicted traditional BLUP-breeding values with an accuracy above a certain threshold. Predictions of genomic breeding values for all animals in the population with genomic information are computed by summing up all previously estimated SNP-effects. This procedure is currently applied in the Swiss dairy cattle populations
- Single-step procedures try to predict genomic breeding values and traditional breeding values in a single evaluation.

9.2 A Linear Model To Predict Genomic Breeding Values

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

$$y = X\beta + Mg + e \tag{9.3}$$

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 (9.3) 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 animal 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 (9.3) have the following structure

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}M \\ M^TR^{-1}X & M^TR^{-1}M + I * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ M^TR^{-1}y \end{bmatrix} \tag{9.4}$$

where

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2} \sum_{i=1}^{m} 2 * p_i * (1 - p_i) \tag{9.5}$$

In (9.5) 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 (9.4) correspond to the SNP-genotype effects. The predicted breeding value $\hat{a}$ for any selection candidate with genomic information is then computed as

$$\hat{a} = \sum_{i=1}^{m} M_i \hat{g}_i \tag{9.6}$$

where M_i corresponds to the vector of SNP-genotypes of the selection candidate.

9.2.1 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.

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 (VanRaden, 2008) 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 Z

$$Z = M - p$$

Finally the matrix ZZ^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.

$$G = \frac{ZZ^T}{\sum_{i=1}^{m} 2p_i(1-p_i)} \tag{9.7}$$

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

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

9.3 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 u which corresponds to the sum of all SNP-allele effects directly as a random effect in the model.

$$y = X\beta + Wu + e \tag{9.8}$$

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

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}W \\ W^TR^{-1}X & W^TR^{-1}W + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ W^TR^{-1}y \end{bmatrix} \tag{9.9}$$

where G is defined as in (9.7) and λ is the same as in equation (9.5). Several authors have shown that both procedures (two-step and single step) are equivalent. From (9.8) we can see that the GBLUP model looks very similar to the animal model, except that the covariances between random effects in the animal model are based on the numerator relationship matrix and in GBLUP they are modeled via the genomic relationship matrix G. This means in the animal 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 $\hat{u}$ coming out of (9.9) are referred to as `direct genomic breeding values` (DGV).

9.4 Practical Problems

The model equations (9.8) 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 animals 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.