4.3 Genomic BLUP

The term **genomic BLUP** is used for the use of genomic information together with pedigree data and phenotypic observations to predict breeding values. Hence the goal is the same as with the pedigree-based BLUP animal model. The main difference is just in the information that goes into the model. But otherwise, the internal modelling mechanisms are the same as before.

The prediction of genomic breeding values which consists of objective of genomic BLUP can be done in two ways. The two ways are

1. marker effect model
2. breeding value based model

4.3.1 Marker Effect Model

When using marker effect models to predict genomic breeding values, this is done in two steps. In a first step marker effects are estimated from a reference population. In that reference population all animals have a complete set of marker genotypes as well as phenotypic observations of the trait of intertest. In a second step the estimated marker effects ($\hat{\mathbf{q}}^T = [\ \hat{q}_1 \quad \hat{q}_2 \quad ... \quad \hat{q}_k\ ]$) are used to predict genomic breeding values for any animal that has genomic information in the form of marker genotypes available.

In Figure 4.3 the principle of the two step procedure to predict genomic breeding values is shown. A possible linear model to estimate SNP-marker-effects based on the data from the reference population can be defined as follows

$$y = Xb + Mq + e \tag{4.20}$$

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

The mixed-model equations resulting from models given in (4.20) have the following structure

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} M \\ M^T R^{-1} X & M^T R^{-1} M + I * \lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{q} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ M^T R^{-1} y \end{bmatrix} \tag{4.21}$$

where

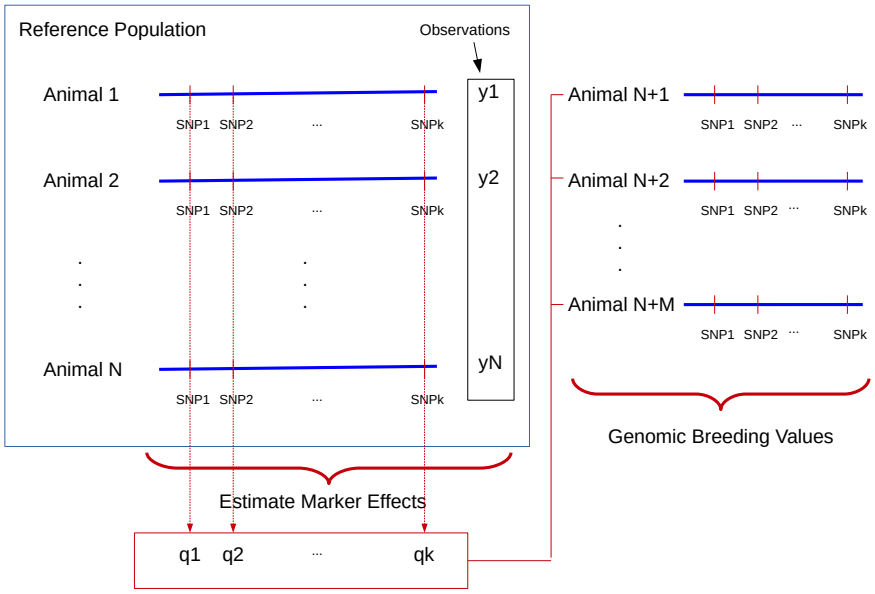

Figure 4.3: Principle of Two-Step Genomic Prediction of Breeding Values

$$\lambda = \frac{\sigma_e^2}{\sigma_q^2} \tag{4.22}$$

In (4.22) σ_q^2 is the total genetic variance explained by the given markers in the dataset

The solutions for $\hat{q}$ from (4.21) correspond to the SNP-genotype effects. The predicted breeding value $\hat{u}$ for any selection candidate i with genomic information is then computed as

$$\hat{u}_i = M_i \cdot \hat{q} \tag{4.23}$$

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

4.3.2 Breeding Value Based Model

The use of breeding value based models to predict genomic breeding values is also known as **single-step** prediction of genomic breeding values. As the term single-step already alludes to, with this method genomic breeding values are predicted directly from the data. This is done by directly integrating genomic

breeding values into the mixed linear effects model where the random effects in the model are the genomic breeding values.

When only looking at the model for predicting genomic breeding values, it looks similar to the pedigree-based animal model as shown below.

$$y = Xb + Zu + e \tag{4.24}$$

The mixed model equations to get solutions used for estimates of fixed effects and predicted genomic breeding values can be written as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + D^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{4.25}$$

The difference to the pedigree-based animal model is the matrix D which depended on the numerator realtionship matrix A for the animal model. In single-step genomic BlUP, the matrix D corresponds to

$$D = G * \sigma_u^2$$

where G corresponds to the genomic relationship matrix and σ_u^2 is taken to be the genetic-additive variance. How the matrix G is constructed is shown in the next section.

4.4 Genomic Relationship Matrix

The variance-covariance matrix between the genetic effects u in model (4.24) is proportional to the genomic relationship matrix G. Analogously to the traditional BLUP animal model where the variance-covariance matrix of the random breeding values is proportional to the numerator relationship matrix A.

4.4.1 Derivation of G

Because the traditional pedigree-based BLUP animal model is very well respected in animal breeding and the defined model (4.24) produces an analogy of the genomic evaluation model to the already known animal model the following properties of u and the genomic relationship matrix G are essential.

1. The genomic breeding values u should correspond to a linear combination of the single SNP-effects q
2. The genomic breeding values u should be defined as deviations from a common mean, leading to the expected value $E[u] = 0$.

3. The variance-covariance matrix of the vector u corresponds to the product of G times a common variance component σ_u^2.
4. The genomic relationship matrix G should be similar to the numerator relationship matrix A. The diagonal elements should be close to 1 and off-diagonal elements of animals that are related should have higher values than elements between unrelated animals.

The matrix G can be computed based on SNP genotypes. In what follows the material of [VanRaden, 2008] and [Gianola et al., 2009] is used to derive the genomic relationship matrix.

4.4.2 Linear Combination of SNP Effects

Based on the SNP marker information the marker effects in the vector q can be estimated. Hence, we assume that the vector q is known. The property that u should be a linear combination of the effects in q means that there exists a matrix U for which we can write

$$u = U \cdot q \tag{4.26}$$

The matrix U is determined based on the desired properties described above.

4.4.3 Deviation

The genomic breeding values u should be defined as deviation from a common basis. Due to this definition the expected value of the genetic effect is determined by $E[u] = 0$. This requirement has the following consequences for the matrix U.

Let us have a look at the random variable w which takes the SNP-genotype codes in the matrix M in the marker effect model. Let us further assume that the SNP loci are in Hardy-Weinberg equilibrium. Then w can take the following values

$$w = \begin{cases} -1 & \text{with probability} \quad (1-p)^2 \\ 0 & \text{with probability} \quad 2p(1-p) \\ 1 & \text{with probability} \quad p^2 \end{cases} \tag{4.27}$$

The expected value of w corresponds to

$$E[w] = (-1)*(1-p)^2+0*2p(1-p)+1*p^2 = -1+2p-p^2+p^2 = 2p-1 \tag{4.28}$$

The matrix U is computed as the difference between the matrix M and the matrix P where the matrix P corresponds to column vectors which have elements corresponding to $2p_j - 1$ where p_j corresponds to the allele frequency of the positive allele at SNP locus j. The following table gives an overview of the elements of matrix U for the different genotypes at SNP locus j.

Genotype	Genotypic Value	Coding in Matrix U
$(G_2G_2)_j$	$-2p_jq_j$	$-1 - 2(p_j - 0.5) = -2p_j$
$(G_1G_2)_j$	$(1 - 2p_j)q_j$	$-2(p_j - 0.5) = 1 - 2p_j$
$(G_1G_1)_j$	$(2 - 2p_j)q_j$	$1 - 2(p_j - 0.5) = 2 - 2p_j$

Here we assume that for a locus G_j, the allele $(G_1)_j$ has a positive effect and occurs with frequency p_j. We can now verify that with this definition of U, the expected value for a genetic effect determined by the locus j corresponds to

$$\begin{aligned} E\left[u\right]_j &= \left[(1-p_j)^2 * (-2p_j) + 2p_j(1-p_j)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= 0 \end{aligned} \tag{4.29}$$

4.4.4 Variance of Genomic Breeding Values

As already postulated the variance-covariance matrix of the genomic breeding values should be proportional to the genomic relationship matrix G.

$$var(u) = G * \sigma_u^2 \tag{4.30}$$

Computing the same variance-covariance matrix based on equation (4.26)

$$var(u) = U \cdot var(q) \cdot U^T \tag{4.31}$$

The variance-covariance matrix of the SNP effects is $var(q) = I * \sigma_q^2$. Inserting this into (4.31) we get $var(u) = UU^T\sigma_q^2$.

In [Gianola et al., 2009] the variance component σ_u^2 was derived from σ_q^2 leading to

$$\sigma_u^2 = 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 \tag{4.32}$$

Now we combine all relationships for $var(u)$ leading to

$$var(u) = G * \sigma_u^2 = UU^T \sigma_q^2 \tag{4.33}$$

In (4.33), σ_u^2 is replaced by the result of (4.32).

$$G * 2 \sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 = UU^T \sigma_q^2 \tag{4.34}$$

Dividing both sides of (4.34) by σ_q^2 and solving for G gives us a formula for the genomic relationship matrix G

$$G = \frac{UU^T}{2\sum_{j=1}^{m} p_j(1-p_j)} \tag{4.35}$$

4.5 How Does GBLUP Work

The genomic relationship matrix G allows to predict genomic breeding values for animals with SNP-Genotypes without any observation in the dataset. This fact is the basis of the large benefit of genomic selection. As soon as a young animal is born, its SNP 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 BVM model given in (4.24) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (4.36). In this form the Inverse G^{-1} of G and the vector $\hat{u}$ of predicted genotypic breeding values are split into one part corresponding to the animals with observations and a second part for the animals without phenotypic information.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{4.36}$$

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

Based on the last line of (4.36) the predicted breeding values $\hat{u}_2$ of all animals without phenotypic observations can be computed from the predicted breeding values $\hat{u}_1$ from the animals with observations.

$$\hat{u}_2 = -\left(G^{22}\right)^{-1} G^{21} \hat{u}_1 \tag{4.37}$$

Equation (4.37) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations.

4.6 Single Step Genomic BLUP With Real-World Datasets

In real-world livestock breeding datasets not all animals are genotyped. But we want to have predicted breeding values for all animals in a population. Futhermore, the genomic information of the genotyped animals should also give more accurate predicted breeding values for related animals without genomic information.

The single step genomic BLUP model can be specified as

$$y = Xb + Zu + e \tag{4.38}$$

with $var(u) = H * \sigma_u^2$ and $var(e) = I * \sigma_e^2$. At this point it is important to note that the vector u of genomic breeding values can be split into two parts

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

where u_1 is the vector of breeding values for non-genotyped animals and u_2 is the vector of genotyped animals.

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda * H^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix} \tag{4.39}$$

where here $\lambda = \sigma_e^2 / \sigma_u^2$.

The above required inverse matrix H^{-1} can be shown (e.g. in [Legarra et al., 2014]) to correspond to

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix}$$

where A^{-1} is the inverse numerator relationship matrix and $A_2 2$ corresponds to the part of the numerator relationship matrix containing all genotyped animals.