Chapter 3

Genomic Best Linear Unbiased Prediction (GBLUP)

In chapter 2 we introduced the fixed linear effects model to estimate additive genotypic values for SNP-Loci. In most real-world genomic datasets the number of SNP loci is larger than the number of observations. But from the point of view of quantitative genetics, we still assume that only a subset of the observed SNP-Loci is linked to a QTL and could therefore have an estimable effect on our trait of interest. Hence the original problem of estimating SNP-effect parameters is extended by a new problem of determining which SNPs are important for the expression of a given trait of interest.

3.1 Finding Relevant SNP Loci

Unfortunately it is not as easy as it may have seamed when we were looking at the monogenic model in Figure 1.5. When there are many SNPs that are observed and that are potentially influencing a trait, the different loci are interacting with each other and the distribution of the different trait values across the different genotypes is much more blurry. Furthermore when we use real-world observations of livestock animals these are phenotypic values which are influenced by many different environmental factors for which the phenotypic measurements all have to be corrected for.

This new problem of determining which SNP locus is linked to a QTL may sound like a not so difficult problem. But the number of possible SNP combinations is quite large. For a given number of k SNP loci the possible number of SNP combinations that might affect a trait is determined by the cardinality of the

powerset of k elements which is in the order of 2^k. Typical values of k might be $1.5 * 10^5$ and hence the number of possible combinations of any number SNP loci is a very large number. As a consequence of that a brute force approach where all possible combinations of SNP loci are tried cannot be used. Figure 3.1 tries to illustrate the problem of selecting important SNPs for a given trait.

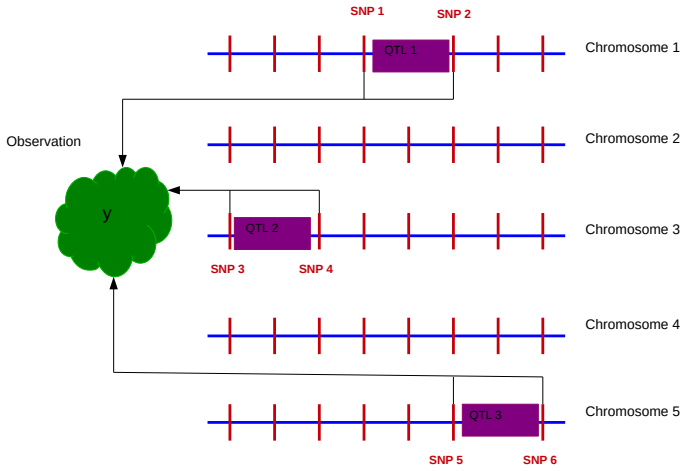

Figure 3.1: Finding SNP Loci Important For The Expression Of A Quantitative Trait

3.2 Stepwise Approach

In fixed linear effects model when the number of predictors is excessively large it is often desirable to find the subset of predictors that have a relevant effect on the response variable. Having too many predictors in a model decreases the power to predict future values of responses. To find a subset of relevant predictor variables out of a large set of predictors can be done with two stepwise approaches.

1. Forward selection
2. Backward elimination

In Statistics textbooks, the procedure of finding relevant predictor variables in a fixed linear model is called `model selection`. In what follows these two approaches are only described very shortly. Due to some practical problems with these techniques they did not find their way into the practical analyses of genomic data.

3.2.1 Forward Selection

The forward selection approach is described by the following step-wise procedure.

1. Start with an empty model $\mathcal{M}_0$ that contains no predictor variables but only an intercept.
2. Include the predictor variable that reduces the residual sum of squares the most.
3. Continue with step 2 until all predictor variables are included. This results in a series of models $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq ....$
4. From the series of models select the one that optimizes a previously defined selection criterion.

As model quality criterion, different quantities can be used. Possible quantities are the Mallow C_p criterion, the Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC).

3.2.2 Backward Elimination

Backward elimination can be viewed as the reverse process of forward selection. The following steps constitute the backward elimination algorithm

1. Start with the full model $\mathcal{M}_0$ containing all available predictor variables.
2. Drop the predictor variable that increases the residual sum of squares the least.
3. Continue with step 2 until all predictor variables are dropped. This results in a series of models $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq ....$
4. From the series of models select the one that optimizes a defined selection criterion.

Backward elimination yields often better results compared to forward selection but it is more computationally expensive. In the case where the number of predictor variables (p) is larger than the number of observations (n), the full model cannot be fit to the data. Hence in that case forward selection would be a possible way to select an optimal model. Although if the number of predictors is very large and therefore many different predictor variables cause a similar reduction of the residual sum of squares, no unique series of models can be generated in the process of forward selection. The result of the forward selection depends on the order of the inclusion of the predictor variables.

3.3 Model Selection With Genomic Data

In real-world genomic data analyses, the number of predictors (p) can be as high as $1.5 * 10^5$. Therefore the backward elimination approach as described in 3.2.2 cannot be used because of the problem of $n << p$. Also the forward selection

approach does not yield a stable procedure for finding the subset of relevant SNPs. This has several reasons which are shortly described in the following subsections.

3.3.1 Fitting The Full Model

Fitting the full model with such a high number of predictors leads to the problem that the design matrix X will not have full column rank. The solution of the least squares normal equation then depends on a generalized inverse $(X^T X)^-$ of $X^T X$. Generalized inverses are not unique and furthermore for a given generalized inverse, there are infinitely many solutions that satisfy the normal equation coming out of least squares. Instead of the non-unique solutions, we have to focus on estimable functions (see section @ref(#asm-flem-estimable-functions)) of the solutions which are independent of the choice of a concrete solution. Although, even if it is possible to fit the full model of a genomic dataset, applying the backward elimination procedure is very time consuming and due to its greediness is not expected to result in a stable subset of relevant SNPs that has an influence on a given trait of interest.

3.4 Mixed Linear Effects Model

Based on the above described problems with the use of the fixed linear effects model for analyzing genomic data, animal breeders were looking for an alternative. In traditional genetic evaluation in animal breeding the BLUP animal model was used world-wide. The term `traditional genetic evaluation` refers to the prediction of breeding values based on phenotypic observation and pedigree relationships between animals in a given population. When looking at Figure 1.7 the traditional genetic evaluation is shown on the left side. The BLUP animal model is a mixed linear effects model where the breeding values of all animals in the populations are taken as random effects. In most of these traditional genetic evaluations the number of predicted breeding values exceeds the number of observations. This is possible due to the BLUP methodology which uses the variance-covariance matrix between the random effects to distribute the information of the observations also to predicted breeding values of animals which do not have any observations. In a BLUP animal model the variance-covariance matrix is proportional to the numerator relationship matrix A. We will see later that when using the genomic version of BLUP the matrix A will be replaced by its equivalent which is called the **genomic relationship matrix**.

Mixed linear effects models can be applied to genomic data using two different parametrisations. At this point, we are using the terminology proposed by [Fernando et al., 2016]. In the first approach SNP loci also referred to as markers are modeled as random effects. These models are called `marker effect models`

(MEM). In a second parametrisation, breeding values of animals corresponding to a linear combination of marker effects are modeled as random effects. This second type of models are referred to as `breeding value models` (BVM). Figure 3.2 illustrates the difference between the two types of models.

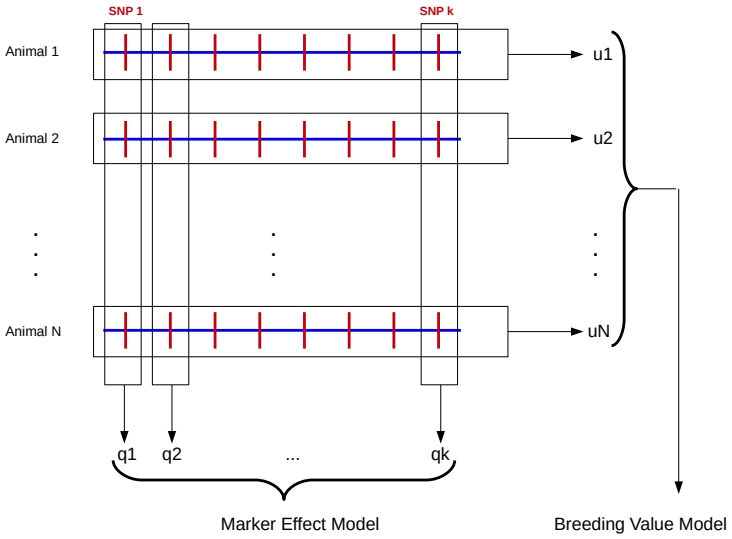

Figure 3.2: Two Types Of Mixed Linear Effects Models For Genomic Data

3.4.1 Marker Effect Models

In MEM random effects of markers are directly included in the model. For an idealized data set we can write

$$y = 1_n \mu + Wq + e \tag{3.1}$$

where

y	vector of length n with observations
μ	general mean denoting fixed effects
1_n	vector of length n of all ones
q	vector of length m of random SNP effects
W	design matrix relating SNP-genotypes to observations
e	vector of length n of random error terms

The vector q contains a separate random effect for each SNP. Because the SNP effects are random, the expected value $E[q]$ and the variance $var(q)$ must be specified. In general, the random effects are defined as deviations and hence their expected value is 0. This means $E[q] = 0$. The variance $var(q)$ can be computed as $var(q) = WW^T\sigma_q^2$. The variance explained by each SNP corresponds to σ_q^2 and is assumed to be constant. The variance $var(e)$ of the random error terms is taken to be $var(e) = I * \sigma_e^2$ where I is the identity matrix and σ_e^2 is the error variance.

3.4.2 Breeding Value Models

In a breeding value model a linear combination of all SNP effects are combined into a random genomic breeding value. This approach is meant when animal breeders are talking about Genomic BLUP (GBLUP). The mixed linear effects model in GBLUP corresponds to

$$y = Xb + Zg + e \tag{3.2}$$

where

y	vector of length n with observations
b	vector of length r with fixed effects
X	incidence matrix linking elements in b to observations
g	vector of length t with random genomic breeding values
Z	incidence matrix linking elements in g to observations
e	vector of length n of random error terms

The vector g contains the genetic effects of all animals that are genotyped which means that they have genomic information based on SNP genotypes available. The expected values of all random effects is assumed to be 0. The variance $var(g)$ of the random genomic breeding values is given by $var(g) = G * \sigma_g^2$. This expression looks very similar to the variance of the breeding values in the

traditional BLUP animal model. The matrix G is called **genomic relationship matrix** (GRM). The variance $var(e)$ of the random error terms is given by $var(e) = I * \sigma_e^2$.

Mostly the older animals for which SNP information is available may have observations (y) in the dataset. The younger animals may have SNP information but in most cases no information is available for them. The goal of GBLUP is to predict genomic breeding values for these animals. Depending on the number of genotyped animals which is in most cases smaller compared to the number of SNP loci, the BVM model has the following advantages over the MEM model

1. The length of the vector g is t which corresponds to the number of genotyped animals which in most cases is smaller than the length of the vector q which is m corresponding to the number of SNPs.
2. Accuracies of genomic breeding values can be computed analogously to the traditional BLUP animal model. This is analogy of accuracies does not exist in MEM.
3. BVM can be combined with pedigree-based animal model analysis which is then referred to as **single step** approach.

More recently with the number of genotyped animals growing very fast, these advantages are no longer as important as they used to be.

3.5 Genomic Relationship Matrix

The variance-covariance matrix between the genetic effects g in model (3.2) 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.

3.5.1 Derivation of G

Because the traditional pedigree-based BLUP animal model is very well respected in animal breeding and the defined model (3.2) produces an analogy of the genomic evaluation model to the already known animal model the following properties of g and the genomic relationship matrix G are essential.

1. The genetic effects g should correspond to a linear combination of the single SNP-effects q
2. The genetic effects g should be defined as deviations from a common mean, leading to the expected value $E[g] = 0$.
3. The variance-covariance matrix of the vector g corresponds to the product of G times a common variance component σ_g^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.

3.5.2 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 g should be a linear combination of the effects in q means that there exists a matrix U for which we can write

$$g = U \cdot q \tag{3.3}$$

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

3.5.3 Deviation

The genetic effects g should be defined as deviation from a common basis. Due to this definition the expected value of the genetic effect is determined by $E[g] = 0$. 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 W in the MEM model given in (3.1). Let us further assume that the SNP loci are in Hardy-Weinberg equilibrium. Then w can take the following values

$$w = \begin{cases} -1 & \text{with probability} \quad (1-p)^2 \\ 0 & \text{with probability} \quad 2p(1-p) \\ 1 & \text{with probability} \quad p^2 \end{cases} \tag{3.4}$$

The expected value of w corresponds to

$$E[w] = (-1)*(1-p)^2 + 0*2p(1-p) + 1*p^2 = -1 + 2p - p^2 + p^2 = 2p - 1 \tag{3.5}$$

The matrix U is computed as the difference between the matrix W 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

$$\begin{aligned} E\left[g\right]_j &= \left[(1-p_j)^2 * (-2p_j) + 2p_j(1-p_j)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= 0 \end{aligned} \tag{3.6}$$

3.5.4 Variance of Genetic Effects

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

$$var(g) = G * \sigma_g^2 \tag{3.7}$$

Computing the same variance-covariance matrix based on equation (3.3)

$$var(g) = U \cdot var(q) \cdot U^T \tag{3.8}$$

The variance-covariance matrix of the SNP effects is $var(q) = I * \sigma_q^2$. Inserting this into (3.8) we get $var(g) = UU^T\sigma_q^2$.

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

$$\sigma_g^2 = 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 \tag{3.9}$$

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

$$var(g) = G * \sigma_g^2 = UU^T\sigma_q^2 \tag{3.10}$$

In (3.10), σ_g^2 is replaced by the result of (3.9).

$$G * 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 = UU^T\sigma_q^2 \tag{3.11}$$

Dividing both sides of (3.11) by σ_q^2 and solving for G gives us a formula for the genomic relationship matrix G

$$G = \frac{UU^T}{2\sum_{j=1}^{m} p_j(1-p_j)} \tag{3.12}$$

3.6 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 (3.2) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (3.13). 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 animals with observations and a second part for the animals without phenotypic information.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{3.13}$$

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 $\hat{g}_1$ contains the predicted breeding values for the animals with observations and the vector $\hat{g}_2$ contains the predicted breeding values of all animals without phenotypic observations.

Based on the last line of (3.13) the predicted breeding values $\hat{g}_2$ of all animals without phenotypic observations can be computed from the predicted breeding values $\hat{g}_1$ from the animals with observations.

$$\hat{g}_2 = -\left(G^{22}\right)^{-1} G^{21}\hat{g}_1 \tag{3.14}$$

Equation (3.14) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations.