**Estimating Breeding Values from Markers**

Jeffrey Endelman

Agron/Hort 812, Spring 2020

The traditional approach to using genetic markers for selection involves two steps. In step one, QTL mapping is conducted to identify significant markers. In step two, marker effects are estimated based on multiple regression (Lande and Thompson 1990). Because of limited statistical power, however, not all of the QTL will be detected, and thus the effects of the significant markers tend to be overestimated (Beavis 1998).

A close up of a map

Description automatically generated

Meuwissen et al. (2001) popularized the idea of **genomic selection**, in which breeding value predictions are made *without* the initial step of QTL mapping. The statistical paradigm underlying this approach inverts classical ideas about which quantities are random vs. model parameters. In classical quantitative genetics, the genotypes are random variables and the additive effects are model parameters. For genomic selection, the marker genotypes are model parameters, and the marker effects are i.i.d. random effects. On a practical level, many studies have shown that genomic selection leads to more accurate predictions compared to the two-step approach (e.g., Heffner et al. 2011). (This result mirrors similar developments in the broader field of statistics.) From a theoretical point of view, however, interpreting the variances and predictions in a genomic selection model, especially in relation to classical quantitative genetic concepts, is not straightforward.

**A screenshot of a cell phone

Description automatically generated**

**Genomic Selection by BLUP**

Let **Z** denote the *n* x *m* matrix of centered genotypes for *n* individuals, *m* markers. Let **a** denote the *n* x 1 vector of additive (i.e., breeding) values, and is the *m* x 1 vector of marker effects.

|  |  |
| --- | --- |
|  | [1] |

From Eq. 1, the variance of the breeding values is multivariate normal with covariance matrix

|  |  |
| --- | --- |
|  | [2] |

By analogy with the classical additive relationship model, we can define a marker-based relationship matrix **A***m* as

|  |  |
| --- | --- |
|  | [3] |

In scalar notation, the relationship coefficient between individuals *i* and *j* is

In the literature, has been called the realized relationship or genomic relationship matrix, and the symbol **G** is also commonly used (which I dislike). Unlike the pedigree relationship matrix, its entries are not nonnegative, and the individual diagonal elements cannot be interpreted as 1 plus the inbreeding coefficient. However, the mean of the diagonal elements provides information about the average level of inbreeding (Endelman and Jannink 2012).

The final expression in Eq. 3 implies the following definition for additive variance:

|  |  |
| --- | --- |
|  | [4] |

This result can be derived by taking the expectation of the breeding value with respect to the sampling of alleles (i.e., expectation in the classical quantitative genetic sense):

|  |  |
| --- | --- |
|  | [5] |

Note that Eq. 4 expresses the additive variance as the sum of locus-specific variances. As discussed in the previous lecture, in the classic quantitative genetic framework, this is only possible when the loci are in linkage equilibrium, which is not true with genome-wide markers. As a result, the connection between and the traditional additive variance (i.e., for the pedigree relationship model) is unclear.



de los Campos et al. (2015)

When the additive values are included as random effects in the model, the methodology is known as **G-BLUP**. When the marker effects are included as random effects, the methodology is known as **RR-BLUP**. As shown above, the two methods lead to the same predictions of **genomic-estimated breeding values** (GEBV).

It is useful to differentiate between making predictions for individuals with both phenotype and marker data vs. individuals with only marker data. The former is known as genome-wide **marker-assisted** selection, while the latter is genome-wide **marker-based** selection. Both types of predictions can be made at the same time. For the G-BLUP method, GEBV are calculated for any individual in the relationship matrix, regardless of whether phenotype data are available. For the RR-BLUP method, the estimated marker effects are multiplied by the centered marker genotypes to predict GEBV.

|  |  |
| --- | --- |
| G-BLUP:  RR-BLUP: | [6] |

**Prediction accuracy vs. predictive ability**

In a strict sense, **prediction accuracy** refers to the correlation between the predicted and true value = . Because the true value is unknown for real data, there are two commonly used approaches to estimate accuracy:

1. Based on the inverse of the coefficient matrix of the MME (see Lecture 1). The main limitation of this approach is that it assumes the model is correct.

2. For marker-based prediction, cross-validation can be used by masking the phenotypes for some individuals. The correlation between the GEBVs for the masked individuals and their estimated true values (calculated without masking) is called **predictive ability**. Dividing the predictive ability by the square-root of the reliability of the validation data (e.g., based on point #1) approximates the accuracy of marker-based selection (Dekkers 2007). This approach is less sensitive to the correctness of the model than #1 but suffers from high mean-squared error when the validation data are less reliable (Estaghvirou et al. 2013).

**References (optional reading)**

Beavis WD (1998) QTL analyses: power, precision and accuracy, pp. 145–162 in *Molecular Dissection of Complex Traits*, edited by AH Paterson. CRC Press, Boca, Raton, FL.

de los Campos G, Sorensen D, Gianola D (2015) Genomic Heritability: What is it? PloS Genetics 11(5):e1005048.

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331­-341.

Endelman JB, Jannink J-L (2012) Shrinkage estimation of the realized relationship matrix. G3 2:1405-1413.

Estaghvirou SBO, Ogutu JO, Schulz-Streeck T, Knaak C, Ouzunova M, Gordillo A, Piepho H-P (2013) Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genomics 14:860.

Heffner EL, Jannink J-L, Sorrells ME (2011). Genomic selection using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75.

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743­-756.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.