File S5: prediction error variance in the training- and validation set.

We assume a balanced and completely random design, with n genotypes and r replicates. Given the model $y_{i,j} = \mu + G_i + E_{i,j}$, the best linear unbiased predictor (BLUP) of $G = (G_1, \ldots, G_n)^t$ and the best linear unbiased estimator (BLUE) of μ are given by

$$\hat{G} = \delta K Z^t (\delta Z K Z^t + I_N)^{-1}(y - \hat{\mu} 1_N), \quad \hat{\mu} = \frac{1_N^t (\delta Z K Z^t + I_N)^{-1} y}{1_N^t (\delta Z K Z^t + I_N)^{-1} 1_N}, \tag{2}$$

where $\delta = \sigma_A^2/\sigma_E^2$ is the shrinkage parameter, N is the total number of individuals and Z is the $N \times n$ incidence matrix assigning individuals to genotypes. See e.g. [6] or [7], or equation (23) in the present work (Appendix C). The parameter $\delta = h^2/(1-h^2)$ is a function of the heritability, and determines the extent to which the phenotypic data y are 'shrunk' towards zero. When the heritability is high, δ is large, and there is little shrinkage, i.e. $\hat{G}$ will be close to the observed phenotypic observations y. For low heritability, δ is small, and y will be shrunk towards the vector of zeros. When BLUPs are based on the genotypic means the same expressions hold, with $N = n$ and $Z = I_n$, and $\hat{G} = \delta_r K (\delta_r K + I_n)^{-1}((\bar{y}_1, \ldots, \bar{y}_n)^t - \hat{\mu} 1_n)$. Since the noise level is reduced from σ_E^2 to $r^{-1}\sigma_E^2$, the shrinkage parameter δ becomes $\sigma_A^2/(r^{-1}\sigma_E^2)$.

The preceding expressions assume the shrinkage parameter to be known, while it is usually estimated from the data. As a consequence, the standard error of $\hat{\mu}$ and prediction error variance of $\hat{G}$ obtained by setting $\delta = \hat{\delta} = \hat{h}^2/(1-\hat{h}^2)$ in (2) are larger than what would be obtained when δ is known ([8], [9]). Before we give examples of too much or too little shrinkage (section), we first give expressions for the prediction error variance for the training and validation set, for the case when heritability is known ($\hat{\delta} = \delta$). These can be derived as a special case of the more general expressions in e.g. [6] or [7].

Prediction error variance when $\delta = \hat{\delta}$

First we consider the genetic effects $G = (G_1, \ldots, G_n)^t$ of the genotypes in the training sample. If we assume that $G \sim N(0, \sigma_A^2 K)$ (i.e. in equation (21) in the main text (Appendix B), γ and the QTL-effects α_m are zero), the prediction error variance is given by the diagonal elements of

$$E(\hat{G} - G)(\hat{G} - G)^t = (Z^t Z + \delta^{-1} K^{-1} - J_n)^{-1}, \tag{3}$$

where Z is the $N \times n$ incidence matrix assigning plants to genotypes, and J_n is the $n \times n$ matrix with identical elements $1/n$. In case the phenotypic data consists of genotypic means, $N = n$. For efficient computation, see [10] [11].

The genetic effects $G_{\text{pred}} = (G_{n+1}, \ldots, G_{n+m})^t$ of m unobserved (but genotyped) genotypes can be predicted with the conditional mean

$$\hat{G}_{\text{pred}} := E[G_{\text{pred}}|y] = \hat{\delta} K_{\text{pred.obs}} Z^t (\hat{\delta} Z K Z^t + I_N)^{-1}(y - \hat{\mu} 1_N), \tag{4}$$

where $K_{\text{pred.obs}}$ is the $m \times n$ matrix of kinship coefficients for the unobserved versus observed genotypes. To give expressions for the prediction error variance $E(\hat{G}_{\text{pred}} - G_{\text{pred}})_{i'}^2$ $(i' = 1, \ldots, m)$ we assume again that $\gamma = 0$, all genetic signal being polygenic. Writing $K_{\text{pred.pred}}$ for the $m \times m$ kinship matrix of the unobserved genotypes, it is assumed that the kinship matrix is the $(n+m) \times (n+m)$ block matrix with K and $K_{\text{pred.pred}}$ on the diagonal and off-diagonal blocks $K_{\text{pred.obs}}$ and $K_{\text{pred.obs}}^t$. Then the conditional distribution of $G_{\text{pred}}|G$ is

$$G_{\text{pred}}|G \sim N\left(K_{\text{pred.obs}} K^{-1} G, \sigma_A^2 \left(K_{\text{pred.pred}} - K_{\text{pred.obs}} K^{-1} K_{\text{pred.obs}}^t\right)\right).$$

Since $\hat{G}_{\text{pred}} = K_{\text{pred.obs}} K^{-1} \hat{G}$ (by comparing (2) and (4)), it follows that

$$\begin{aligned} (\hat{G}_{\text{pred}} - G_{\text{pred}})|(\hat{G} - G) &= K_{\text{pred.obs}} K^{-1}(\hat{G} - G) - Y, \\ \text{where} \quad Y &\sim N\left(0, \sigma_A^2 (K_{\text{pred.pred}} - K_{\text{pred.obs}} K^{-1} K_{\text{pred.obs}}^t)\right). \end{aligned}$$

Consequently, the prediction error variances $E(\hat{G}_{\text{pred}} - G_{\text{pred}})_i^2$ are the diagonal elements of

$$\begin{aligned} &E(\hat{G}_{\text{pred}} - G_{\text{pred}})(\hat{G}_{\text{pred}} - G_{\text{pred}})^t = E\left[E(\hat{G}_{\text{pred}} - G_{\text{pred}})(\hat{G}_{\text{pred}} - G_{\text{pred}})^t \mid (\hat{G} - G)\right] \\ &= (K_{\text{pred.obs}} K^{-1}) \left[E(\hat{G} - G)(\hat{G} - G)^t\right] K^{-1} K_{\text{pred.obs}}^t \\ &\quad + \sigma_A^2 (K_{\text{pred.pred}} - K_{\text{pred.obs}} K^{-1} K_{\text{pred.obs}}^t). \end{aligned} \tag{5}$$

Hence, the prediction error variance for the validation set contains a term depending on $\delta^{-1} = \sigma_E^2/\sigma_A^2$ (see (3)), as well as a term which depends only on the genetic variance σ_A.

Prediction error variance with incorrect shrinkage ($\delta \neq \hat{\delta}$)

For the case that the amount of shrinkage is not chosen correctly ($\hat{\delta} \neq \delta = \sigma_A^2/(r^{-1}\sigma_E^2)$), we now give an expression for the prediction error variance for the training set based on genotypic means, under the additional assumption that μ is known to be zero. The BLUP for G then simplifies to

$$\hat{G} = \hat{\delta}K(\hat{\delta}K + I_n)^{-1}\bar{y}, \tag{6}$$

where we recall that we still assume a balanced and completely random design. Hence $\bar{y}_i = G_i + \bar{E}_i$, with $\bar{E}_i \sim N(0, r^{-1}\sigma_E^2)$ and $G = (G_1, \ldots, G_n)^t \sim N(0, \sigma_A^2 K)$. Since $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_n)^t \sim N(0, \sigma_A^2 K + r^{-1}\sigma_E^2 I_n) = N(0, \sigma_E^2(\delta K + r^{-1} I_n)$, the variance-covariance matrix of $\hat{G} - G$ equals

$$\mathrm{Var}(\hat{G} - G) = \sigma_A^2 K - 2\hat{\delta}K(\hat{\delta}K + I_n)^{-1}\sigma_A^2 K + \hat{\delta}K(\hat{\delta}K + I_n)^{-1}(\delta K + r^{-1} I_n)(\hat{\delta}K + I_n)^{-1}\hat{\delta}K\sigma_E^2,$$

where we used that (by the independence of G and E)

$$\mathrm{Cov}(G, \hat{G}) = \mathrm{Cov}(G, \hat{\delta}K(\hat{\delta}K + I_n)^{-1}G) = \hat{\delta}K(\hat{\delta}K + I_n)^{-1}\sigma_A^2 K$$

and that (using $\bar{y} \sim N(0, \sigma_E^2(\delta K + r^{-1} I_n)$ and the symmetry of K and I_n)

$$\hat{G} = \hat{\delta}K(\hat{\delta}K + I_n)^{-1}\bar{y} \sim N(0, \hat{\delta}K(\hat{\delta}K + I_n)^{-1}(\delta K + r^{-1} I_n)(\hat{\delta}K + I_n)^{-1}\hat{\delta}K\sigma_E^2).$$

In particular, when $\hat{\delta} = \infty$ (i.e. $\hat{h}^2 = 1$), there is no shrinkage, and $\hat{G} = \bar{y}$. The prediction error variance is then completely determined by the residual variance, since $\hat{G} - G = \bar{y} - G = \bar{E}$, and

$$E(\hat{G} - G)(\hat{G} - G)^t = r^{-1}\sigma_E^2 I_n.$$

On the other hand, when $\hat{\delta} = 0$ (i.e. $\hat{h}^2 = 0$), there is 'total' shrinkage towards zero, i.e. $\hat{G} = 0$, and

$$E(\hat{G} - G)(\hat{G} - G)^t = E(GG^t) = \sigma_A^2 K.$$

This explains the asymmetry in the observed accuracy in our simulations, in particular when $h^2 = 0.5$: when the number of replicates r is sufficiently large, overestimating the heritability will have less impact on the prediction error variance (and hence accuracy) than underestimating it.