*Genet. Res., Camb.* (2010), **92**, *pp.* 423–441. © Cambridge University Press 2011
doi:10.1017/S0016672310000583


# Methods of plant breeding in the genome era

SHIZHONG XU* AND ZHIQIU HU
*Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA*

(*Received 5 October 2010 and in revised form 17 October 2010*)

**Summary**

Methods of genomic value prediction are reviewed. The majority of the methods are related to mixed model methodology, either explicitly or implicitly, by treating systematic environmental effects as fixed and quantitative trait locus (QTL) effects as random. Six different methods are reviewed, including least squares (LS), ridge regression, Bayesian shrinkage, least absolute shrinkage and selection operator (Lasso), empirical Bayes and partial least squares (PLS). The LS and PLS methods are non-Bayesian because they do not require probability distributions for the data. The PLS method is introduced as a special dimension reduction scheme to handle high-density marker information. Theory and methods of cross-validation are described. The leave-one-out cross-validation approach is recommended for model validation. A working example is used to demonstrate the utility of genome selection (GS) in barley. The data set contained 150 double haploid lines and 495 DNA markers covering the entire barley genome, with an average marker interval of 2·23 cM. Eight quantitative traits were included in the analysis. GS using the empirical Bayesian method showed high predictability of the markers for all eight traits with a mean accuracy of prediction of 0·70. With traditional marker-assisted selection (MAS), the average accuracy of prediction was 0·59, giving an average gain of GS over MAS of 0·11. This study provided strong evidence that GS using marker information alone can be an efficient tool for plant breeding.

## 1. Introduction

The purpose of plant breeding is to improve the productivity of agricultural crops per unit of farmland by manipulating the genetic compositions of the target populations. Since the origin of life, natural selection has been constantly acting to produce the current diversity of living organisms on earth. Due to natural selection, all species have more or less adapted to their own local conditions. Agricultural crops have also adapted to their own niches, but in this case the adaptation has been directed by artificial selection imposed by humans. The key difference between natural and artificial selection is that natural selection acts on phenotypes of traits related to organismal fitness, while artificial selection is based on specific phenotypes of agronomic importance. More recently, the efficiency of artificial selection can be increased by incorporating genotypic information. This review will focus on methods incorporating genotypic information for plant breeding. Unlike phenotypic data, genotypic data are not subject to changes due to environmental errors and more directly related to the target genes for the traits of interest. With high-density marker information, the genotypes of all quantitative trait loci (QTLs) are partially or fully observed. Selection of plants using genome-wide QTLs can be more efficient than most other methods of selection. Such a genome selection (GS) approach is the state-of-art breeding technology in all plant species. The following few sections will briefly review the history of selection procedures and the application of marker information for facilitating plant breeding.

### (i) *Phenotypic selection*

Phenotypic selection is the simplest way to improve the productivity of crops. By definition, the criterion of phenotypic selection is the phenotypic value of a

* Corresponding author: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA. e-mail: shxu@ucr.edu

desired trait. The selection response is proportional to the heritability of the target trait and the selection intensity, as indicated by the breeders' equation (Falconer & Mackay, 1996; Lynch & Walsh, 1998)

$$R = h^2 i \sigma_P, \quad (1)$$

where $h^2$, $i$ and $\sigma_P$ are the heritability, the selection intensity and the phenotypic standard deviation of the trait, respectively. With the advent of modern technologies, other information has also been used to select superior plants for breeding. However, phenotypic selection remains an important criterion in plant breeding. Experienced breeders always choose plants with desired morphological characters to breed, provided the plants have passed all other criteria of selection.

(ii) *Best linear unbiased prediction* (*BLUP*)

Any information related to the genetic constitution of the candidate plants may be used to predict the breeding values of candidates. The heritability of the trait is an indication of the extent to which the individual's own phenotypic value predicts its breeding value, the prediction being more accurate for traits with higher heritability than traits with lower heritability. Therefore, heritability of a trait is the accuracy of breeding value prediction using phenotypic information as the selection criterion. Phenotypic values of other plants who are genetically related to the candidates (e.g. parents, progeny and siblings), can also be used to predict the breeding values of the candidate plants because relatives and the candidate plant share common genetic material (Fisher, 1918). This method of selection is called pedigree selection (Chahal & Gosal, 2002). Prior to the genome era, the combination of pedigree analysis with phenotypic selection was the most widely utilized plant breeding method (Moose & Mumm, 2008).

Sib analysis and progeny testing are special forms of pedigree analysis because both take advantage of information from relatives. Plants are related in many different ways and the above two types of pedigree analyses only count for a subset of these relatives. The optimal way of pedigree analysis is to include all relatives. This requires a way to handle heterogeneous genetic relationship among plants in the breeding population. Data collected from experiments are subject to unexpected environmental and human errors, leading to missing values for some plots. Therefore, even a well-balanced experimental design may produce unbalanced data. The ordinary least squares method is incapable of dealing with the heterogeneous genetic relationship and the unbalanced data. The BLUP (Henderson, 1950) emerged as the ideal tool for plant and animal breeders to solve both problems.

Let $y$ be a vector of phenotypic values for $n$ individuals in a population. The linear mixed model for $y$ can be described as

$$y = X\beta + Z\gamma + \varepsilon, \quad (2)$$

where $\beta$ and $X$ are the fixed effects and the design matrix of the fixed effects, $\gamma$ and $Z$ are the random effects and the corresponding design matrix for the random effects and $\varepsilon$ is the residual error vector with an assumed $N(0,R)$ distribution, where $R$ is an $n \times n$ covariance matrix for the residual errors. The fixed effects represent systematic effects that should be controlled to reduce the residual errors. The random effects are defined as the genetic effects for plants included in the data and/or plants not included in the data but which contribute to the population as ancestors. Let us assume $\gamma \sim N(0, A\sigma_A^2)$, where $\sigma_A^2$ is the additive genetic variance and $A$ is the additive relationship matrix for all the plants contributing to the random effect $\gamma$. The predicted breeding value for a candidate plant is defined as a linear combination of $\hat{\gamma}$, say $\hat{\xi} = L^T\hat{\gamma}$, where $\hat{\gamma}$ is the BLUP of $\gamma$ estimated from the mixed model equation (Henderson, 1975)

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + A^{-1}/\sigma_A^2 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}. \quad (3)$$

This mixed model equation explains why BLUP can handle unbalanced data and deal with heterogeneous relationships. The mixed model uses the general linear model (dummy variable) notation to represent the analysis of variance (ANOVA) model and connects the model effects with the data through the design matrices, $X$ and $Z$. These design matrices represent the characteristics of the experimental design and apply to all experiments, regardless of whether or not the data are balanced. The additive relationship matrix ($A$) is twice the kinship matrix ($\Theta$), $A = 2\Theta$, and it represents any arbitrary kinships among the plants contributing $\gamma$. Note that the kinship matrix contains all pair-wise co-ancestry coefficients among the plants (see next section for the definition of co-ancestry coefficient). If the plants are classified into full-sib or half-sib families, the $A$ matrix has a clear pattern that allows simplified algorithms to be used to handle calculations involving it. If all plants are independent, $A = I$, and the mixed model remains valid. Efficient algorithms are available to estimate the variance components and predict $\gamma$ (Calvin, 1993). Note that the information required to perform BLUP includes the phenotypic values ($y$) and the kinships of the plants ($\Theta$).

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

(iii) *Selection using realized kinship*

What does the kinship matrix $\Theta$ mean? How useful is it in predicting the breeding values of plants? For any two individuals, $i$ and $j$, the kinship (also called co-ancestry coefficient) is denoted by $\Theta_{ij}$ and thus $A_{ij}=2\Theta_{ij}$. Technically, there are two explanations for $\Theta_{ij}$: (1) At any given locus, $\Theta_{ij}$ represents the probability that a randomly sampled allele from $i$ is identical-by-descent (IBD) to a randomly sampled allele from $j$. Therefore, it is the 'average' IBD value at this locus for all pairs of individuals with the same relationship. (2) Considering the same pair of individuals, $\Theta_{ij}$ also represents the 'average' IBD values for individual $i$ and $j$ across all loci in the genome. The second explanation is more useful in the mixed model equation because it complies with the infinitesimal model of quantitative traits (Fisher, 1918). In the genome era, genome-wide markers can be used to estimate $\Theta$. The estimated $\Theta$ using molecular markers is called the realized kinship, denoted by $\tilde{\Theta}$. When the marker density is infinitely high and the genome size is infinitely large, the realized kinship is identical to the theoretical kinship $\tilde{\Theta}=\Theta$. Therefore, if $\Theta$ is known for all individuals (pedigree information is available), the realized kinship does not help at all in plant breeding.

Three special situations make the realized kinship useful. (1) If pedigree information is unavailable, then we can use genome-wide markers to infer the realized kinship and use $\tilde{\Theta}$ in the mixed model equation to perform BLUP (Queller & Goodnight, 1989; Lynch & Ritland, 1999). (2) If a quantitative trait is controlled by loci with heterogeneous effects (violating the infinitesimal model), the realized kinship estimated only from these loci with large effects can improve the accuracy of the BLUP prediction. This condition is rarely, if ever met because it depends on knowledge of the genetic architecture of the trait. (3) For plants with small genomes, the realized kinship should be different from the expected kinship. Sib analysis using the realized kinship may allow the separation of the genetic variance from the common environmental variance because the realized kinship varies across sibling pairs whereas the expected kinship is a constant for all sibling pairs (Visscher *et al*., 2006). Common environmental effects in plants are not as common as in animals and humans, but they may exist to some extent.

(iv) *Marker-assisted selection* (*MAS*)

MAS emerged as an efficient method to improve the accuracy of selection in plant breeding programs (Dudley, 1993). Contrary to BLUP, the success of MAS relies on monogenic or oligogenic models (Lamkeya & Lee, 1993). The monogenic model means that the variance of a trait is controlled by the segregation of a single major gene. If the trait is controlled by a few major genes, the model is said to be oligogenic. MAS depends heavily on the result of QTL mapping, especially interval mapping or single marker analysis. The genetic model for interval mapping is

$$y=X\beta+W_k\alpha_k+\varepsilon, \tag{4}$$

where $\alpha_k$ is the effect of QTL $k$ and $W_k$ is a genotype indicator variable. The subscript $k$ means that all markers have been evaluated and the $k$th marker happens to have the largest effect. The molecular breeding value for individual $j$ is

$$\hat{y}_j=X_j\hat{\beta}+W_{jk}\hat{\alpha}_k, \tag{5}$$

where $\hat{\alpha}_k$ is the estimated QTL effect. Selection can be performed using this molecular breeding value. The MAS scheme is effective under the monogenic model. Lande & Thompson (1990) realized that if the trait is controlled by one major gene plus numerous genes with small effects, the observed phenotype for individual $j$ should also be used to predict the breeding value for the candidate. Therefore, they proposed a new MAS scheme through an index that combines the molecular breeding value with the observed phenotype. The index is

$$I_j=b_1y_j+b_2\hat{y}_j, \tag{6}$$

where weights, $b_1$ and $b_2$, are obtained using the standard procedure of index selection (Smith, 1936; Hazel, 1943). The polygenic information is contained in $y$, although not explicitly expressed. The molecular breeding value under the oligogenic model is

$$\hat{y}_j=X_j\hat{\beta}+\sum_{k=1}^{p}W_{jk}\hat{\alpha}_k, \tag{7}$$

where $p$ is the number of major genes detected via QTL mapping and is usually a small number, typically less than five. Again, the Lande & Thompson (1990) index is a better choice due to its ability to capture the polygenic effect through $y_j$.

(v) *Genome prediction under the $Q+K$ model*

Association mapping deals with randomly sampled individuals from a target population. Such a population usually has a heterogeneous background, e.g. a population with explicit or hidden structures. Pritchard *et al*. (2000*a*, *b*) used the following model to describe the phenotype:

$$y=X\beta+Q\delta+W_k\alpha_k+\varepsilon. \tag{8}$$

The additional term $\delta$ represents the structural effects and $Q$ is determined by the population structure.

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

Such a model is called the Q model (Thornsberry *et al.*, 2001; Camus-Kulandaivelu *et al.*, 2006). Yu *et al.* (2006) extended the Q model by adding a polygenic component using the realized kinship $\tilde{\Theta}$. The modified model is

$$y = X\beta + Q\delta + W_k\alpha_k + \gamma + \varepsilon, \tag{9}$$

where $\gamma$ is the polygenic effect with an assumed $N(0, \tilde{A}\sigma_A^2)$ distribution and $\tilde{A} = 2\tilde{\Theta}$ is estimated from genome-wide marker information. This model is called the Q+K model, where Q stands for the population structure and K stands for the kinship (Yu *et al.*, 2006). Since $A$ is not available in wild populations and populations without recorded history, $\tilde{A}$ is always needed. In the case where pedigree information is known, the true $A$ should be used. Using $\tilde{A}$ while $A$ is already available will do more harm than good to the association study.

(vi) *Whole GS*

Whole GS is a method using genome-wide high-density markers in a different way from the marker-based analyses described above (Meuwissen *et al.*, 2001; Xu, 2003). Different markers are treated separately rather than pooled together as a polygene. The GS model is

$$y = X\beta + \sum_{k=1}^{p} Z_k\gamma_k + \varepsilon, \tag{10}$$

where the number of markers $p$ can be extremely large to cover the entire genome rather than a few detected from interval mapping. Each marker effect is assumed to be $N(0,\sigma_k^2)$ distributed with a marker-specific variance. Putting all marker effects in a single vector $\gamma = \{\gamma_k\}_{k=1}^{p}$ and letting $Z = \{Z_k\}_{k=1}^{p}$ be the genotype indicator variable array for all markers, the above model can be rewritten as

$$y = X\beta + Z\gamma + \varepsilon, \tag{11}$$

where $\gamma \sim N(0,G)$ and $G = \mathrm{diag}\{\sigma_k^2\}$. The corresponding mixed model equation is

$$\begin{bmatrix} X^{\mathrm{T}}R^{-1}X & X^{\mathrm{T}}R^{-1}Z \\ Z^{\mathrm{T}}R^{-1}X & Z^{\mathrm{T}}R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}R^{-1}y \\ Z^{\mathrm{T}}R^{-1}y \end{bmatrix}. \tag{12}$$

The estimated breeding value for individual $j$ is

$$\hat{y}_j = X_j\hat{\beta} + \sum_{k=1}^{p} Z_{jk}\hat{\gamma}_k. \tag{13}$$

The difference between this model and the models described previously is that no polygenic effects are included because they have been absorbed by the genome-wide markers. As a result, pedigree information is no longer relevant.

## 2. Methods of genome selection

(i) *LS method*

Let us reintroduce the model here for the LS method:

$$y = X\beta + \sum_{k=1}^{p} Z_k\gamma_k + \varepsilon, \tag{14}$$

where $p$ is the number of markers included in the model. This time $p < n$, where $n$ is the sample size. This constraint is needed because LS solution of the parameters requires the existence of $(Z^{\mathrm{T}}R^{-1}Z)^{-1}$. When $p \geqslant n$, the $Z^{\mathrm{T}}R^{-1}Z$ matrix is uninvertable. The LS solution for the parameters is through solving the following normal equation:

$$\begin{bmatrix} X^{\mathrm{T}}R^{-1}X & X^{\mathrm{T}}R^{-1}Z \\ Z^{\mathrm{T}}R^{-1}X & Z^{\mathrm{T}}R^{-1}Z \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}R^{-1}y \\ Z^{\mathrm{T}}R^{-1}y \end{bmatrix}. \tag{15}$$

Note that $\gamma_k$ is treated as a fixed effect and no distribution is assigned to it. The question is how to deal with the high-density markers of the whole genome. The answer is to adopt a variable selection scheme to eliminate markers with small effects, so that $p$ is small enough to be manageable. Forward selection is a reasonable approach for selecting the markers. The criterion for marker inclusion is quite arbitrary but should not be too stringent. Professional computer software is available to perform forward selection, such as the REG and GLMSELECT procedures in SAS. The latter is a more efficient version of model selection and can deal with classification variables. The program also provides an option to evaluate the model through cross-validation (described later).

The LS method alone (without variable selection) is not a choice for genome prediction because the constraint of $p < n$ is rarely met for a population with a high-density marker map. It can be useful to re-evaluate the effects of markers that are selected using other variable selection procedures, e.g. Lasso. In fact, if $p$ (the number of selected markers) is much smaller than $n$, the LS method is preferred due to the best linear unbiased property of the method.

(ii) *Ridge regression*

Ridge regression (Hoerl & Kennard, 1970) can relax the restriction of $p < n$ to a certain degree. It uses the same linear model given in eqn (14), but adds a small positive number $\lambda$ to the corresponding diagonal elements of the normal equation:

$$\begin{bmatrix} X^{\mathrm{T}}R^{-1}X & X^{\mathrm{T}}R^{-1}Z \\ Z^{\mathrm{T}}R^{-1}X & Z^{\mathrm{T}}R^{-1}Z + I\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}R^{-1}y \\ Z^{\mathrm{T}}R^{-1}y \end{bmatrix}, \tag{16}$$

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

where $I$ is an identity matrix with the same dimension as $Z^TR^{-1}Z$. Ridge regression is also called shrinkage analysis with a common shrinkage factor $\lambda$ for all regression coefficients. This shrinkage factor will cause biased estimation of the regression coefficients, but the small bias is paid off with reduced estimation errors for the parameters. The shrinkage factor must be provided by the investigators *a priori* or inferred empirically from the data (Draper & Smith, 1998). The REG procedure in SAS has a Ridge option to perform ridge regression analysis.

A more rigorous way to determine $\lambda$ is through Bayesian analysis because ridge regression has a Bayesian analogy. If we assign each regression coefficient to a normal prior, $\gamma_k \sim N(0,\varphi^2), \forall k = 1, \ldots, p$, where $\varphi^2$ is a common prior variance, we obtain $\lambda = 1/\varphi^2$. Since $\varphi^2$ can be estimated from the data under the mixed model framework, we have $\hat{\lambda} = 1/\hat{\varphi}^2$, provided that $\hat{\varphi}^2$ is the estimated variance component. The MIXED procedure in SAS is perhaps the most efficient program to perform variance component analysis.

The common shrinkage prior can increase $p$ indefinitely by increasing $\lambda$. However, as $\lambda$ grows, the degree of shrinkage becomes stronger and eventually all regression coefficients are shrunk to zero. Although the model can handle an arbitrarily large $p$ by further shrinking the regression coefficients, a model with all regression coefficients infinitely small is not desirable, as such a model will not have any ability to predict the phenotype. An optimal shrinkage scheme is one that can selectively shrink the regression coefficients. Markers with small or no effects should be severely penalized whereas those with large effects should not be shrunk at all. The following sections provide a few common procedures with the selective characteristics. Any one of them is acceptable as a tool for GS.

(iii) *Bayesian shrinkage*

The Bayesian shrinkage analysis uses prior $\gamma \sim N(0,G)$, where $G = \text{diag}\{\sigma_k^2\}$ are marker-specific prior variances. The mixed model equation (Henderson, 1975),

$$\begin{bmatrix} X^{\mathrm{T}}R^{-1}X & X^{\mathrm{T}}R^{-1}Z \\ Z^{\mathrm{T}}R^{-1}X & Z^{\mathrm{T}}R^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} X^{\mathrm{T}}R^{-1}y \\ Z^{\mathrm{T}}R^{-1}y \end{bmatrix} \quad (17)$$

is used to derive the posterior distribution for $\beta$ and $\gamma$. Robinson (1991) showed that the posterior distribution for $\beta$ is

$$p(\beta|\cdots) = N(\beta|\mu_\beta, V_\beta), \quad (18)$$

where

$$\mu_\beta = [X^{\mathrm{T}}(ZGZ^{\mathrm{T}}+R)^{-1}X]^{-1}X^{\mathrm{T}}(ZGZ^{\mathrm{T}}+R)^{-1}y \quad (19)$$

and

$$V_\beta = [X^{\mathrm{T}}(ZGZ^{\mathrm{T}}+R)^{-1}X]^{-1}. \quad (20)$$

Conditional on $\beta$, the posterior distribution for $\gamma$ is

$$p(\gamma|\cdots) = N\left(\gamma|\mu_\gamma, V_\gamma\right), \quad (21)$$

where

$$\mu_\gamma = (G^{-1}+Z^{\mathrm{T}}R^{-1}Z)^{-1}Z^{\mathrm{T}}R^{-1}(y-X\beta) \quad (22)$$

and

$$V_\gamma = (G^{-1}+Z^{\mathrm{T}}R^{-1}Z)^{-1}. \quad (23)$$

Therefore, both $\beta$ and $\gamma$ are sampled from their perspective normal posterior distributions in the Markov chain Monte Carlo (MCMC)-implemented Bayesian shrinkage analysis.

Let us assume that the covariance matrix of the residual errors takes the simplified form $R = I\sigma^2$, where $\sigma^2$ is the error variance. Assigning a scaled inverse chi-square distribution to $\sigma^2$, i.e. $\sigma^2 \sim \text{Inv}-\chi^2(\tau,\omega)$, the posterior distribution for $\sigma^2$ remains scaled inverse chi-square,

$$p(\sigma^2|\cdots) = \text{Inv}-\chi^2(\sigma^2|\tau+n, \omega+SS), \quad (24)$$

where

$$SS = (y-X\beta-Z\gamma)^{\mathrm{T}}(y-X\beta-Z\gamma). \quad (25)$$

The error variance $\sigma^2$ can be sampled from $p(\sigma^2|\cdots)$ in the MCMC sampling process.

The Bayesian shrinkage analysis differs from the mixed model variance component analysis in that $\sigma_k^2$ is further assigned by a scaled inverse chi-square distribution $\sigma_k^2 \sim \text{Inv}-\chi^2(\tau,\omega)$, where $\tau$ and $\omega$ are hyper parameters provided by the investigators. Conditional on $\gamma_k$, the posterior distribution for $\sigma_k^2$ is

$$p(\sigma_k^2|\cdots) = \text{Inv}-\chi^2(\sigma_k^2|\tau+1, \omega+\gamma_k^2). \quad (26)$$

When $(\tau,\omega) = (0,0)$, the prior for $\sigma_k^2$ becomes $1/\sigma_k^2$, an improper prior. Theoretically, this prior may cause poor mixing of the Markov chain (ter Braak *et al.*, 2005), but in reality, it usually produces satisfactory results. The uniform prior of $\sigma_k^2$ corresponds to $(\tau,\omega) = (-2,0)$, which generates results similar to the maximum-likelihood analysis.

The Bayesian shrinkage method has been incorporated into the QTL procedure in SAS (Hu & Xu, 2009). PROC QTL is a user-defined SAS procedure. Users need a regular SAS license and the PROC QTL software (separate from the Base SAS) to run the QTL procedure. Once the QTL software is installed, users can call the QTL procedure

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

the same way as calling other built-in SAS procedures.

(iv) *Lasso and related methods*

(a) *Lasso*

The least absolute shrinkage and selection operator (Lasso) was proposed by Tibshirani (1996) to handle oversaturated regression models. It is a penalized regression analysis with solution of the parameters obtained via

$$\gamma^{\text{Lasso}} = \arg\min_{\gamma} \left[ \left( y - \sum_{k=1}^{p} Z_k \gamma_k \right)^{\text{T}} \times \left( y - \sum_{k=1}^{p} Z_k \gamma_k \right) + \lambda \sum_{k=1}^{p} |\gamma_k| \right], \tag{27}$$

where $\lambda > 0$ is the shrinkage or penalization factor. Note that $\beta$ disappeared from the model, which is accomplished via centralization and rescaling of $y$ and $Z$. If there is only one intercept in the model ($\beta$ is a scalar), the standardization is obtained by $y_j = (y_j^* - \bar{y}^*)/s^*$ and $Z_{jk} = (Z_{jk}^* - \bar{Z}_k^*)/S_k^*$, where the variables with a superscript $*$ are the original variables before the standardization. Such a simple standardization scheme is not available in general. A special treatment is required for a general $X\beta$. One can adopt the procedure for removal of $X\beta$ in the restricted maximum likelihood REML method (Patterson & Thompson, 1971). In this procedure, we find a matrix $T = \text{subset}[I - X(X^{\text{T}}X)^{-1}X^{\text{T}}]$ so that

$$\begin{aligned} TX &= \text{subset}[I - X(X^{\text{T}}X)^{-1}X^{\text{T}}]X \\ &= \text{subset}[X - X(X^{\text{T}}X)^{-1}X^{\text{T}}X] = 0. \end{aligned} \tag{28}$$

We define the model with the original data (before standardization) by

$$y^* = X\beta + \sum_{k=1}^{p} Z_k^* \gamma_k + \varepsilon^* \tag{29}$$

and multiply both sides of the equation by matrix $T$ to obtain

$$Ty^* = TX\beta + \sum_{k=1}^{p} TZ_k^* \gamma_k + T\varepsilon^*. \tag{30}$$

Let $y = Ty^*$, $Z_k = TZ_k^*$ and $\varepsilon = T\varepsilon^*$. We now have

$$y = \sum_{k=1}^{p} Z_k \gamma_k + \varepsilon \tag{31}$$

a model with $\beta$ completely removed. The subset can be chosen as any $n-q$ independent rows of matrix $I - X(X^T X)^{-1}X^T$ (Harville, 1977). The current Lasso method does not have such a general treatment. The optimal way to handle the general situation is to modify the Lasso equation using

$$\gamma^{\text{Lasso}} = \arg\min_{\gamma} \left[ \left( y - \sum_{k=1}^{p} Z_k \gamma_k \right)^{\text{T}} \Sigma^{-1} \times \left( y - \sum_{k=1}^{p} Z_k \gamma_k \right) + \lambda \sum_{k=1}^{p} |\gamma_k| \right], \tag{32}$$

where $\Sigma = \text{var}(\varepsilon) = \text{var}(T\varepsilon^*) = \text{TRT}^{\text{T}}$.

Fast algorithms are available and have been implemented in various software packages. The least-angle regression (LARS) program can perform the Lasso analysis (Efron *et al.*, 2004). The GLMSELECT procedure in SAS also has an option to provide Lasso variable selection. The Lasso algorithm estimates all regression coefficients, but at least $n-p$ coefficients will have estimated values of exactly zero. Therefore, Lasso is also a variable selection procedure.

(b) *EM Lasso*

Let us reintroduce the linear model here,

$$y = X\beta + \sum_{k=1}^{p} Z_k \gamma_k + \varepsilon, \tag{33}$$

where $\varepsilon \sim N(0, I\sigma^2)$ and thus $y \sim N(X\beta + \sum_{k=1}^{p} Z_k \gamma_k, I\sigma^2)$. Let $\beta$ be assigned a uniform prior and the prior for each $\gamma_k$ is

$$p(\gamma_k) = \text{Laplace}(\gamma_k | \lambda) = \frac{\lambda}{2} \exp(-\lambda |\gamma_k|). \tag{34}$$

The joint log likelihood function of $\theta = \{\beta, \gamma\}$ is

$$L(\theta) = -\frac{1}{2\sigma^2} \left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right)^{\text{T}} \left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right) - \lambda \sum_{k=1}^{p} |\gamma_k| + \text{const}, \tag{35}$$

where

$$\text{const} = -n \ln(\sigma^2)/2 + n \ln(\lambda/2). \tag{36}$$

Multiplying both sides of the equation by $2\sigma^2$, we have

$$2\sigma^2 L(\theta) = -\left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right)^{\text{T}} \left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right) - 2\sigma^2 \lambda \sum_{k=1}^{p} |\gamma_k| + \text{const}. \tag{37}$$

Redefining the shrinkage factor as $\lambda^* = 2\sigma^2\lambda$, we now have the following function to maximize:

$$\theta^{\text{Lasso}} = \arg\max_{\theta} \left[ -\left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right)^{\text{T}} \times \left( X\beta + \sum_{k=1}^{p} Z_k \gamma_k \right) - \lambda^* \sum_{k=1}^{p} |\gamma_k| \right], \tag{38}$$

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

where $\theta^{\text{Lasso}}=\{\beta^{\text{Lasso}},\gamma^{\text{Lasso}}\}$ and $\gamma^{\text{Lasso}}$ is the Lasso estimate of the QTL effects. Since

$$\begin{aligned} p(\gamma_k) &= \int_{\sigma_k^2} N(\gamma_k|0,\sigma_k^2)\text{Expon}(\sigma_k^2|\lambda) \\ &= \text{Laplace}(\gamma_k|\lambda)=\frac{\lambda}{2}\exp(-\lambda|\gamma_k|), \end{aligned} \tag{39}$$

we have an alternative way to achieve the Lasso estimates of the parameters through the hierarchical model, $\gamma_k \sim N(0,\sigma_k^2)$ and $\sigma_k^2 \sim \text{Expon}(\lambda)$. Details of the hierarchical model can be found in Park & Casella (2008). Therefore, given $G=\text{diag}\{\sigma_k^2\}$, the mixed model equation applies,

$$\begin{bmatrix} X^{\text{T}}R^{-1}X & X^{\text{T}}R^{-1}Z \\ Z^{\text{T}}R^{-1}X & Z^{\text{T}}R^{-1}Z+G^{-1} \end{bmatrix}\begin{bmatrix}\beta\\ \gamma\end{bmatrix}=\begin{bmatrix} X^{\text{T}}R^{-1}y \\ Z^{\text{T}}R^{-1}y\end{bmatrix} \tag{40}$$

from which the posterior mean and posterior variance for each component of $\theta$ can be found. Let $E(\gamma_k|\ldots)$ and $\text{var}(\gamma_k|\ldots)$ be the posterior mean and posterior variance for $\gamma_k$ and $E(\gamma_k^2|\cdots)=E(\gamma_k|\cdots)+\text{var}(\gamma_k^2|\cdots)$, the M-step of the EM estimate for $\gamma_k$ is

$$\sigma_k^2=\frac{\sqrt{1+4\lambda^2E(\gamma_k^2|\cdots)}-1}{2\lambda^2}. \tag{41}$$

The E-step is represented by calculating $E(\gamma_k{}^2|\ldots)$. The residual error variance is obtained by

$$\sigma^2=\frac{1}{n}(y-X\beta)^{\text{T}}\left(y-X\beta-\sum_{k=1}^{p}Z_kE(\gamma_k|\cdots)\right). \tag{42}$$

The EM Lasso algorithm has been coded in a SAS/IML program (Xu, 2010) and can be downloaded from our personal website (http://www.statgen.ucr.edu).

### (c) *Bayesian Lasso*

Bayesian Lasso (Park & Casella, 2008) is an MCMC-implemented Bayesian version of the Lasso procedure. It has been applied to QTL mapping by Yi & Xu (2008). The same hierarchical prior in the EM Lasso is used for the MCMC-implemented Lasso algorithm. In addition, the square of $\lambda$ is further assigned a Gamma prior, $\lambda^2 \sim \text{Gamma}(a,b)$. The sampling process is the same as that described in the Bayesian shrinkage section for $\beta$, $\gamma$ and $\sigma^2$. The conditional posterior distribution for $\nu_k=1/\sigma_k{}^2$ is the inverse Gaussian, i.e.

$$p(\nu_k|\cdots)=\text{Inv-Gauss}\left(\nu_k\left|\sqrt{\lambda^2\sigma^2/\gamma_k^2},\lambda^2\right.\right) \tag{43}$$

and the shrinkage factor $\lambda^2$ has a Gamma posterior distribution,

$$p(\lambda^2|\cdots)=\text{Gamma}\left(\lambda^2\left|p+a,\sum_{k=1}^{p}\sigma_k^2/2+b\right.\right). \tag{44}$$

The main advantage of the Bayesian Lasso over the original Lasso method is that $\lambda$ does not have to be predetermined; rather it can be treated as a variable subject to Monte Carlo sampling. Bayesian Lasso for QTL mapping has been implemented in the R program (Yandell *et al.*, 2007).

### (v) *Empirical Bayes*

Empirical Bayes is a method to incorporate a data-estimated prior distribution (Casella, 1985). Xu (2007) first adopted the empirical Bayesian method to map QTL. The linear model is rewritten as

$$y=X\beta+\zeta, \tag{45}$$

where

$$\zeta=\sum_{k=1}^{p}Z_k\gamma_k+\varepsilon. \tag{46}$$

The expectation and variance–covariance matrix for the data are $E(y)=X\beta$ and

$$V=\sum_{k=1}^{p}Z_kZ_k^{\text{T}}\sigma_k^2+I\sigma^2. \tag{47}$$

With the scaled inverse chi-square distribution for each $\sigma_k^2 \sim \text{Inv}-\chi^2(\tau,\omega)$, the log likelihood function for $\beta$, $G$ and $\sigma^2$ is

$$\begin{aligned} L(G) = &-\frac{1}{2}\ln|V|-\frac{1}{2}(y-X\beta)^TV^{-1}(y-X\beta) \\ &-\frac{1}{2}(\tau+2)\ln|G|-\frac{\omega}{2}\text{tr}(G^{-1}). \end{aligned} \tag{48}$$

Note that this likelihood function does not involve $\gamma$, which is integrated out. An algorithm has been developed to estimate $G$, denoted by $\tilde{G}$. With the $G$ matrix in the prior replaced by the data-estimated value, $\gamma \sim N(0,\tilde{G})$, the mixed model equation,

$$\begin{bmatrix} X^{\text{T}}R^{-1}X & X^{\text{T}}R^{-1}Z \\ Z^{\text{T}}R^{-1}X & Z^{\text{T}}R^{-1}Z+\tilde{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\beta}\\ \hat{\gamma}\end{bmatrix}=\begin{bmatrix} X^{\text{T}}R^{-1}y \\ Z^{\text{T}}R^{-1}y\end{bmatrix} \tag{49}$$

is then used to estimate the QTL effects in a single step without iterations. The empirical Bayesian method for QTL mapping has been incorporated into the QTL procedure in SAS (Hu & Xu, 2009).

### (vi) *PLS*

PLS (Wold, 1973) is used to find the fundamental relationship between two matrices ($Z$ and $Y$), and is a latent variable approach for modelling the covariance structures in the two spaces. The $Z$ matrix in QTL analysis is an $n\times p$ matrix determined by marker genotypes. The $Y$ matrix is an $n\times 1$ vector for the phenotypes of a quantitative trait. In general, PLS can

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

handle multiple traits, i.e., $Y$ can be an $n \times q$ matrix for $q$ traits. The PLS method will try to find the multidimensional direction in the $Z$ space that explains the maximum multidimensional variance in the $Y$ space. PLS regression is particularly suited when the matrix of predictors has more variables than the number of observations, $p > n$, and when there is multicollinearity among the $Z$ values. In contrast, standard regression will fail in these cases. In this section, we review the special case when $Y$ is a vector. Technical details about PLS can be found in numerous articles (Dijkstra, 1983; de Jong, 1993; Lindgren *et al*., 1993) and books (Abdi, 2003). The method described here closely follows Boulesteix & Strimmer (2007).

Like the original Lasso method, the data need to be standardized prior to the analysis. The model for the standardized data is $y = Z\gamma + \varepsilon$. PLS constructs an $n \times c$ latent components matrix as a linear transformation of $Z$, i.e. $T = ZW$, where $W$ is a $p \times c$ matrix of weights and $c < p$ is the number of latent variables provided by the investigators. The $i$th column of matrix $T$ is

$$T_i = W_{1i}Z_1 + W_{2i}Z_2 + \cdots + W_{pi}Z_p, \quad \forall i = 1, \ldots, c. \tag{50}$$

The high-dimensional $Z$ matrix is now projected to the low-dimensional $T$ matrix. We need to find the $W$ matrix using a way satisfying the special constraints described below. Let us assume that we have already found matrix $W$. The next step is to perform linear regression of $Y$ on $T$, as

$$y = T\eta + \varepsilon = ZW\eta + \varepsilon, \tag{51}$$

where $\eta$ is a $c \times 1$ vector of the regression coefficients of $y$ on $T$. We now put the two models together,

$$\begin{cases} y = Z\gamma + \varepsilon \\ y = ZW\eta + \varepsilon \end{cases} \tag{52}$$

and conclude that $\gamma = W\eta$. Therefore, given $W$ and $T = ZW$, we can perform multiple regression for $y = T\eta + \varepsilon$ with the LS solution

$$\eta^{\mathrm{LS}} = (T^{\mathrm{T}}T)^{-1}T^{\mathrm{T}}y = (W^{\mathrm{T}}Z^{\mathrm{T}}ZW)^{-1}W^{\mathrm{T}}Z^{\mathrm{T}}y. \tag{53}$$

The LS solution of $\eta$ is then converted into the PLS estimate of $\gamma$,

$$\gamma^{\mathrm{PLS}} = W\eta^{\mathrm{LS}} = W(W^{\mathrm{T}}Z^{\mathrm{T}}ZW)^{-1}W^{\mathrm{T}}Z^{\mathrm{T}}y. \tag{54}$$

This approach is much like principal component analysis (PCA), but the way to derive the $W$ matrix is different. In PCA, the latent matrix $T$ is found only by maximizing the variances in the predictors $Z$ and ignoring the $Y$ matrix. In the PLS method, the latent matrix $T$ is found by taking into account both $Z$ and $Y$.

Let $W_i$ and $W_j$ be the $i$th and $j$th columns of matrix $W$ for $i < j$. The corresponding latent variables are $T_i = ZW_i$ and $T_j = ZW_j$. Three quantities are required to derive the $W$ matrix, which are

$$\begin{aligned} \mathrm{cov}(T_i, y) &= \mathrm{cov}(ZW_i, y) = \frac{1}{n}W_i^{\mathrm{T}}Z^{\mathrm{T}}y, \\ \mathrm{var}(T_i) &= \mathrm{var}(ZW_i) = \frac{1}{n}W_i^{\mathrm{T}}Z^{\mathrm{T}}ZW_i, \\ \mathrm{cov}(T_i, T_j) &= \mathrm{cov}(ZW_i, ZW_j) = \frac{1}{n}W_i^{\mathrm{T}}Z^{\mathrm{T}}ZW_j. \end{aligned} \tag{55}$$

The columns of matrix $W$ are defined such that the squared sample covariance between $y$ and $T_i$ is maximal, under the restriction that the latent components are mutually uncorrelated. Moreover, the variance of the latent variance is constrained to have a unity value. Mathematically, the solution for $W$ is obtained as

$$\begin{aligned} W_i &= \arg\max_{W_i} [\mathrm{cov}(T_i, y) \times \mathrm{cov}(y, T_i)] \\ &= \arg\max_{W_i} (W_i^{\mathrm{T}}Z^{\mathrm{T}}yy^{\mathrm{T}}ZW_i) \end{aligned} \tag{56}$$

subject to constraints

$$\begin{cases} \mathrm{var}(T_i) = W_i^{\mathrm{T}}Z^{\mathrm{T}}ZW_i = 1, \\ \mathrm{cov}(T_i, T_j) = W_i^{\mathrm{T}}Z^{\mathrm{T}}ZW_j = 0. \end{cases} \tag{57}$$

The maximum number of latent components which have non-zero covariance with $Y$ is $c_{\max} = \min(n,p)$. The weight vector $W_i$ is computed sequentially with the order of $W_1, W_2, \cdots, W_c$, where $c$ is provided by the users. Software packages are available for PLS. The most comprehensive one is the PLS procedure in SAS.

PLS has been applied to GS using simulated data (Solberg *et al*., 2009) and SNP data in dairy cattle (Moser *et al*., 2009). These authors found that the prediction accuracy of PLS is comparable to the Bayesian method. There are fewer reports of PLS application to genomic value prediction than the Bayesian method. More studies on this topic are expected to appear soon in the literature. The earliest report of the application of PLS to quantitative genetics was the updated index selection procedure by Xu & Muir (1992). The authors did not explicitly state that the method is PLS, but the approach they used to find the weights of the updated selection indices is exactly the same as the PLS. Xie & Xu (1997) extended the PLS method to restricted multistage index selection. Both studies gave the expressions of the weights along with the detailed description of the mathematical derivation. The constraints given in eqn (57) are adopted from Xu & Muir (1992) and they are different from the ones used in the SAS/PLS procedure. The PLS in SAS uses $W_i^{\mathrm{T}}W_i = 1$ as a constraint, instead of $W_i^{\mathrm{T}}Z^{\mathrm{T}}ZW_i = 1$. The final $W$ matrices obtained using the two constraint systems are different, but the prediction of the genetic value for any candidate individual remains the same.

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

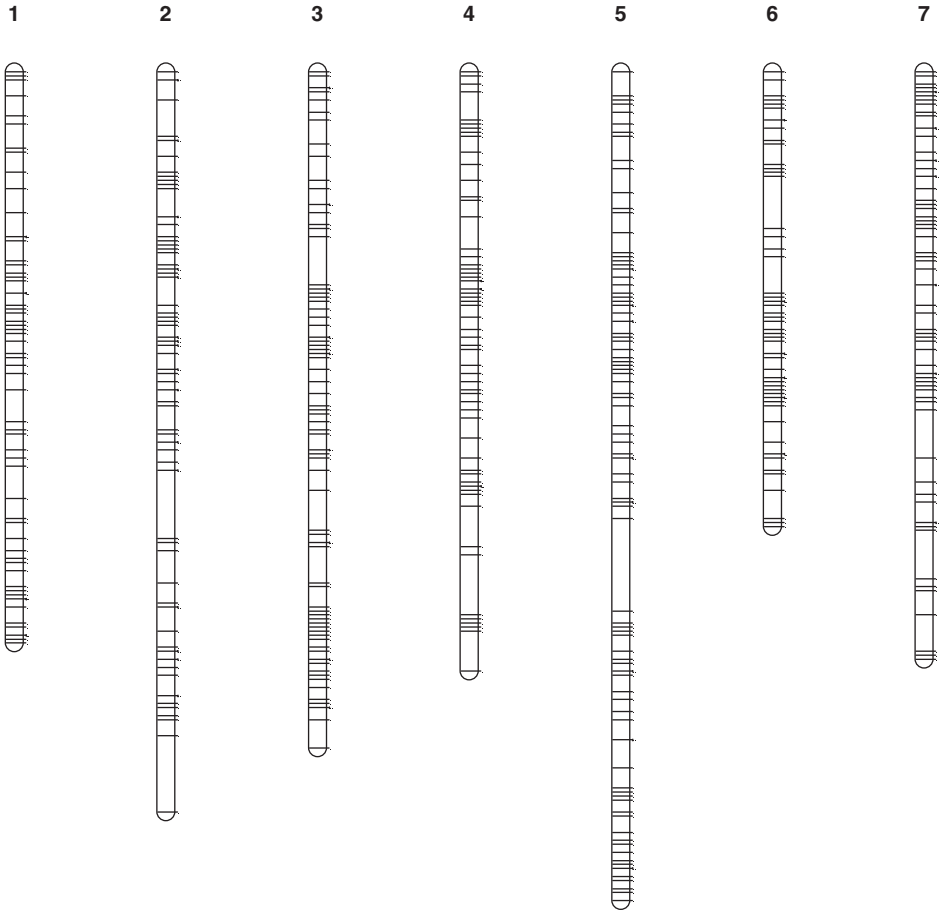


Fig. 1. Linkage map of 495 markers covering the seven chromosomes of the barley genome.

## 3. Cross-validation

The main purpose of GS is to predict the genomic or breeding values of candidate plants. More attention is paid to prediction rather than to hypothesis testing, which, while related, are not necessarily the same. A QTL may pass a threshold for hypothesis testing and be declared as significant, but may have little predictive value. Increasing the sample size can increase the number of detected QTLs (Beavis, 1994), but it does not necessarily increase the predictability of the model. Therefore, model validation is fundamentally important in GS. It is highly recommended that any practical study in GS be accompanied by the result of validation before consideration of publication. Breeding companies will not adopt any new procedures without some forms of validation.

### (i) *Prediction error*

Let $y_j$ be the observed phenotypic value of individual $j$ in the population and

$$\hat{y}_j = X_j\hat{\boldsymbol{\beta}} + \sum_{k=1}^{p} Z_{jk}\hat{\gamma}_k \tag{58}$$

be its predicted value. If individual $j$ has contributed to the estimation of $\beta$ and $\gamma$, the error defined by $y_j - \hat{y}_j$ is a residual error, not a prediction error. The residual error can be arbitrarily small by increasing the number of markers in the model. A prediction error is defined by the difference between the observed phenotypic value and the predicted value for a new individual who has not contributed to the estimation of the parameters that are used to make the prediction. If individual $j$ is a new candidate plant and the phenotypic value has not been observed yet, we can predict the phenotype using the parameters estimated from the current sample. The predicted value is

$$\hat{y}_j^{\text{New}} = X_j^{\text{New}}\hat{\boldsymbol{\beta}} + \sum_{k=1}^{p} Z_{jk}^{\text{New}}\hat{\gamma}_k. \tag{59}$$

Later on the phenotype of this plant is measured with a value $y_j^{\text{New}}$. The error defined by $y_j^{\text{New}} - \hat{y}_j^{\text{New}}$ is called the prediction error. The prediction error can be reduced to some degree but cannot be eliminated. The variance of the prediction errors is defined by

$$\hat{\boldsymbol{\phi}}^2 = \frac{1}{m}\sum_{j=1}^{m}\left(y_j^{\text{New}} - \hat{y}_j^{\text{New}}\right)^2. \tag{60}$$

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

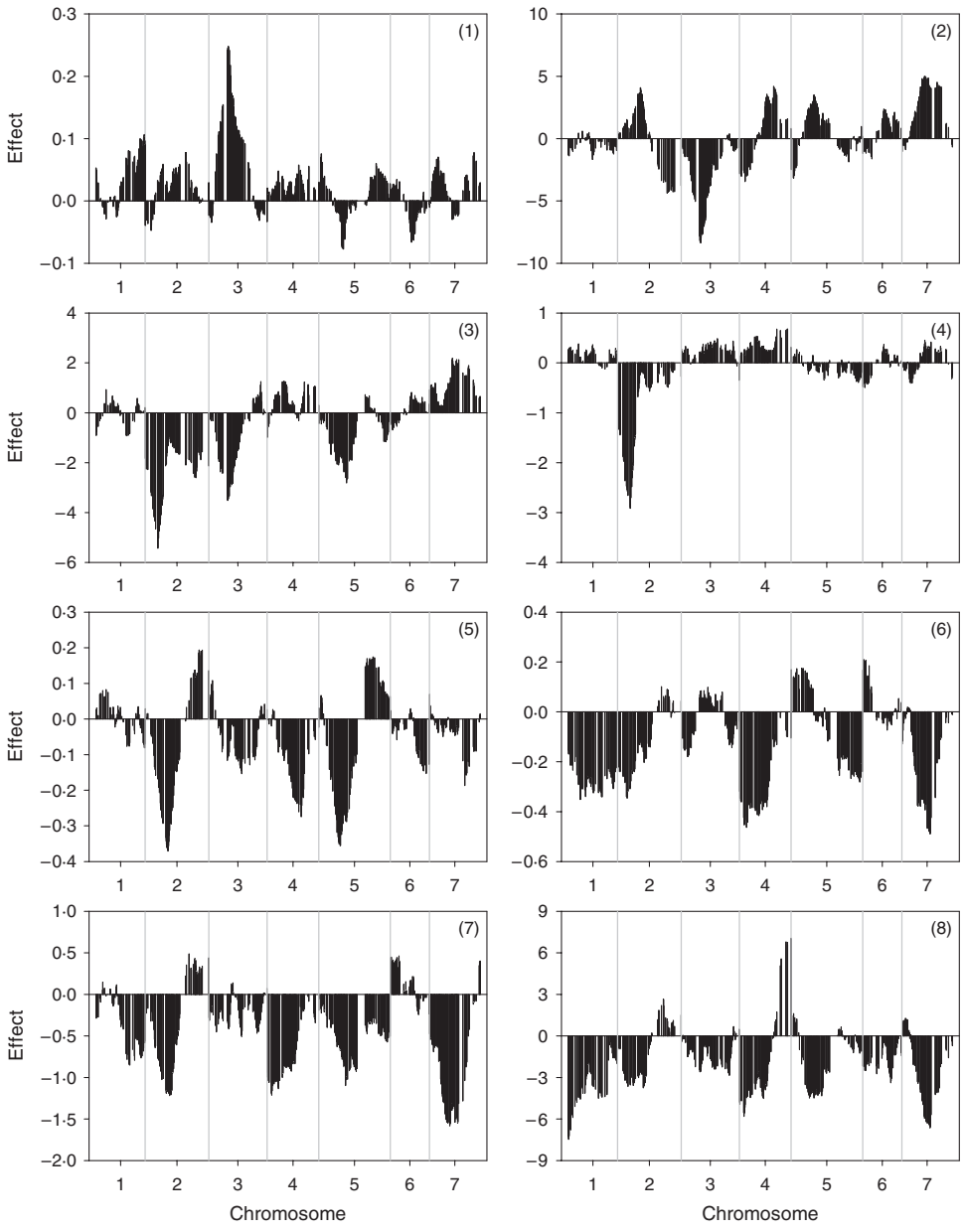


Fig. 2. Estimated QTL effects for eight quantitative traits in the barley experiment using the interval mapping approach. The eight traits correspond to (1) Yield, (2) Lodging, (3) Height, (4) Head, (5) Protein, (6) Extract, (7) Amylase and (8) Power. The seven chromosomes are separated by the vertical reference lines.

All $m$ individuals are new and none of their phenotypes has contributed to the parameter estimation. Assuming that $m \to \infty$, we can write the prediction error variance as

$$\phi_j^2 = \mathrm{var}(y_j^{\mathrm{New}} - \hat{y}_j^{\mathrm{New}}) = \mathrm{var}(y_j^{\mathrm{New}}) + \mathrm{var}(\hat{y}_j^{\mathrm{New}}), \quad (61)$$

where $\mathrm{var}(y_j^{\mathrm{New}}) = \sigma^2$ and

$$\mathrm{var}(\hat{y}_j^{\mathrm{New}}) = X_j^{\mathrm{New}} \mathrm{var}(\hat{\beta}) X_j^{(\mathrm{New})\mathrm{T}} + Z_j^{\mathrm{New}} \mathrm{var}(\hat{\gamma}) Z_j^{(\mathrm{New})\mathrm{T}} + 2X_j^{\mathrm{New}} \mathrm{cov}(\hat{\beta}, \hat{\gamma}) Z_j^{(\mathrm{New})\mathrm{T}}. \quad (62)$$

Let $\hat{\theta} = \left\{\hat{\beta}, \hat{\gamma}\right\}$ and $W_j^{\mathrm{New}} = \{X_j^{\mathrm{New}}, Z_j^{\mathrm{New}}\}$. In ordinary LS analysis,

$$\mathrm{var}(\hat{\theta}) = (W^{\mathrm{T}} W)^{-1} \sigma^2. \quad (63)$$

Therefore, the variance of the predicted value is

$$\mathrm{var}(\hat{y}_j^{\mathrm{New}}) = W_j^{\mathrm{New}} \mathrm{var}(\hat{\theta}) W_j^{(\mathrm{New})\mathrm{T}} = W_j^{\mathrm{New}} (W^{\mathrm{T}} W)^{-1} W_j^{(\mathrm{New})\mathrm{T}} \sigma^2. \quad (64)$$

This leads to

$$\phi_j^2 = \sigma^2 \left[1 + W_j^{\mathrm{New}} (W^{\mathrm{T}} W)^{-1} W_j^{(\mathrm{New})\mathrm{T}}\right]. \quad (65)$$

Therefore,

$$\phi^2 = \frac{1}{m} \sum_{j=1}^{m} \phi_j^2 = \sigma^2 \times \left[1 + \frac{1}{m} \sum_{j=1}^{m} W_j^{\mathrm{New}} (W^{\mathrm{T}} W)^{-1} W_j^{(\mathrm{New})\mathrm{T}}\right] \quad (66)$$

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

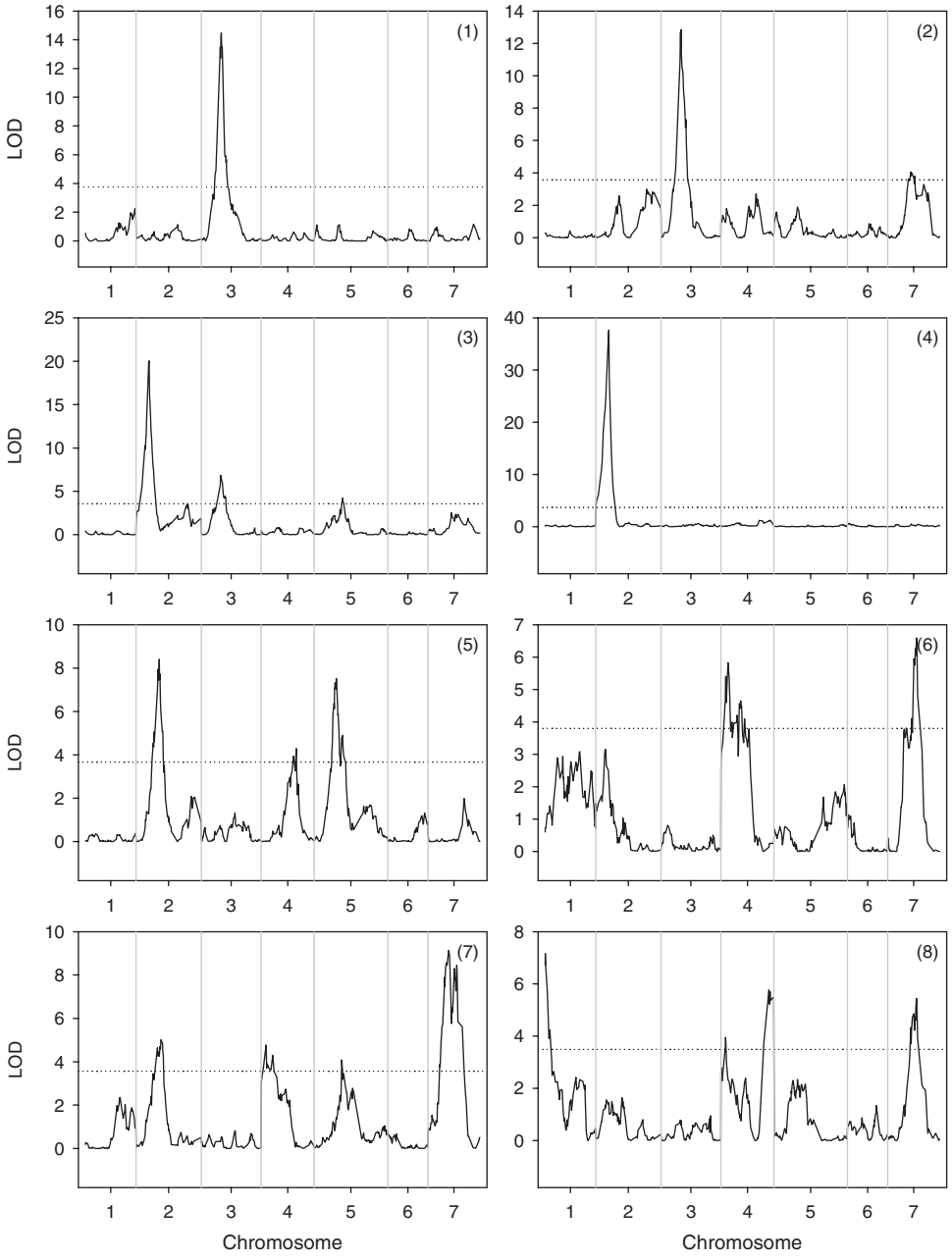


Fig. 3. LOD scores for eight quantitative traits in the barley experiment using the interval mapping approach. The eight traits correspond to (1) Yield, (2) Lodging, (3) Height, (4) Head, (5) Protein, (6) Extract, (7) Amylase and (8) Power. The horizontal reference lines are the permutation (1000 samples) generated critical values for the LOD score test at $\alpha = 0{\cdot}01$.

Clearly $\phi^2 \geqslant \sigma^2$ and the equality holds if and only if $n = \infty$. The prediction error variance is at least as large as the (true, not the estimated) residual error variance.

(ii) *Model validation*

To validate a model, more resources are required. We can divide the sample into a training (learning) sample and a testing sample with approximately equal size. The training sample is used to estimate the parameters. The estimated parameters are then used to calculate the prediction error variance in the testing sample. The prediction error variance $\phi^2$ has a unit depending on the trait (squared unit of the trait). It is usually transformed into a number between 0 and 1 so that different trait analyses can be compared on the same scale. Let us define

$$\hat{\varphi}^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\hat{y}_j^{\text{New}} - \bar{y}^{\text{New}}\right)^2, \tag{67}$$

where $\bar{y}^{\text{New}} = m^{-1}\sum_{j=1}^{m} y_j^{\text{New}}$ is the average value of the phenotypes in the testing sample. The *R*-square

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

Table 1. *Accuracies* (*R-squares*) *of MAS using the least-squares method under two levels of Type I errors* (*alpha values*)

| | Alpha = 0·01 | | Alpha = 0·05 | |
|---|---|---|---|---|
| Trait | Number of markers | $R^2$ | Number of markers | $R^2$ |
| (1) Yield | 20 | 0·5673 | 26 | 0·5511 |
| (2) Lodging | 28 | 0·5593 | 41 | 0·5093 |
| (3) Height | 40 | 0·6489 | 53 | 0·6404 |
| (4) Head | 27 | 0·7213 | 27 | 0·7213 |
| (5) Protein | 47 | 0·5869 | 64 | 0·5510 |
| (6) Extract | 33 | 0·4847 | 71 | 0·4768 |
| (7) Amylase | 59 | 0·5629 | 79 | 0·4556 |
| (8) Power | 31 | 0·5782 | 35 | 0·5739 |

value is defined as (Legates & McCabe, 1999)

$$\hat{R}^2 = \frac{\hat{\varphi}^2}{\hat{\phi}^2 + \hat{\varphi}^2} = 1 - \frac{\hat{\phi}^2}{\hat{\phi}^2 + \hat{\varphi}^2}, \tag{68}$$

which has a domain $0 \leqslant R^2 \leqslant 1$, with zero indicating no predictability and 1 indicating perfect prediction. If $\bar{y}^{\text{New}} = \hat{\bar{y}}^{\text{New}} = m^{-1} \sum_{j=1}^{m} \hat{y}_j^{\text{New}}$, the denominator can be rewritten as

$$\hat{\phi}^2 + \hat{\varphi}^2 = \frac{1}{m} \sum_{j=1}^{m} (y_j^{\text{New}} - \bar{y}^{\text{New}})^2 \tag{69}$$

In this case, the *R*-square is interpreted as the proportion of phenotypic variance contributed by the genomic variance.

(iii) *K-fold cross-validation*

The validation procedure described in the previous section does not optimally use the resources. There are *n* plants in the sample but only half of the *n* plants are used to estimate the parameters and half of the *n* plants used to validate the model. Thus, some resources have been wasted using this true validation procedure. If the training sample and the testing sample are reversed in function, there is another validation scheme and this new scheme will produce a different result. The two results may be combined to calculate a new *R*-square. Such an *R*-square should be more precise because it uses the whole sample. This scheme is called cross-validation (Shao, 1993). There are many different ways to perform cross-validation. The half-half cross-validation is called twofold cross-validation.

With the twofold cross-validation, we have increased the sample size for the *R*-square calculation, but have not increased the sample size for parameter estimation. The parameters are estimated twice but the two sets of estimated parameters are not combined. Each is estimated separately, still using half of the sample. There is no reason not to use a threefold cross-validation, in which the sample is divided into three parts: two parts are used to estimate the parameters and the remaining part is used to validate the parameters. Each of the three parts is eventually validated using parameters estimated from the other parts. This time, the parameters are estimated from 2/3 of the sample. Similarly, a fivefold cross-validation uses 4/5 of the sample to estimate the parameters and validates the prediction of the remaining 1/5 of the sample (Moser *et al.*, 2009). In general, people can choose any K-fold cross-validation, where K is an integer between 2 and *n*.

(iv) *Leave-one-out* (n-*fold cross-validation*)

Leave-one-out cross-validation applies to the case when $K = n$ (Efron, 1983). We use $n-1$ plants to estimate the parameters and predict the value for the remaining plant. The complete cross-validation requires *n* separate analyses, one for each plant. The computation can be intensive for large samples, but it is the optimal way to utilize the current resources, and thus should be the most reliable cross-validation approach.

Compared to other K-fold cross-validations, the *n*-fold cross-validation has the smallest prediction error variance. This is because it has the smallest estimate errors for the parameters due to the maximum possible sample size ($n-1$) used. Theoretically, the *R*-square value should also be the highest for the *n*-fold cross-validation. Is the high *R*-square an over estimate of the predictability? In practice, if we have *n* plants in the current sample, we will never use a subsample to estimate the parameters. Suppose that we now have new plants with available DNA samples but not the phenotypes. We are ready to predict the genetic values of these plants for selection. The optimal approach for predicting the breeding values of the new plants is to use parameters estimated from all *n* plants. The *n*-fold cross-validation uses $n-1$ plants to estimate the parameter and $n-1$ is the nearest integer to *n*. Therefore, the leave-one-out cross-validation mimics most closely to the actual prediction in practice.

## 4. Working example

(i) *Barley experiment*

The original experiment was conducted by |Hayes *et al.* (1993, 1994). The data were retrieved from http://www.genenetwork.org/. The experiment involved 150 double haploids (DH) derived from the cross of two Spring barley varieties, Morex and Steptoe. There were 495 markers distributed along the

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

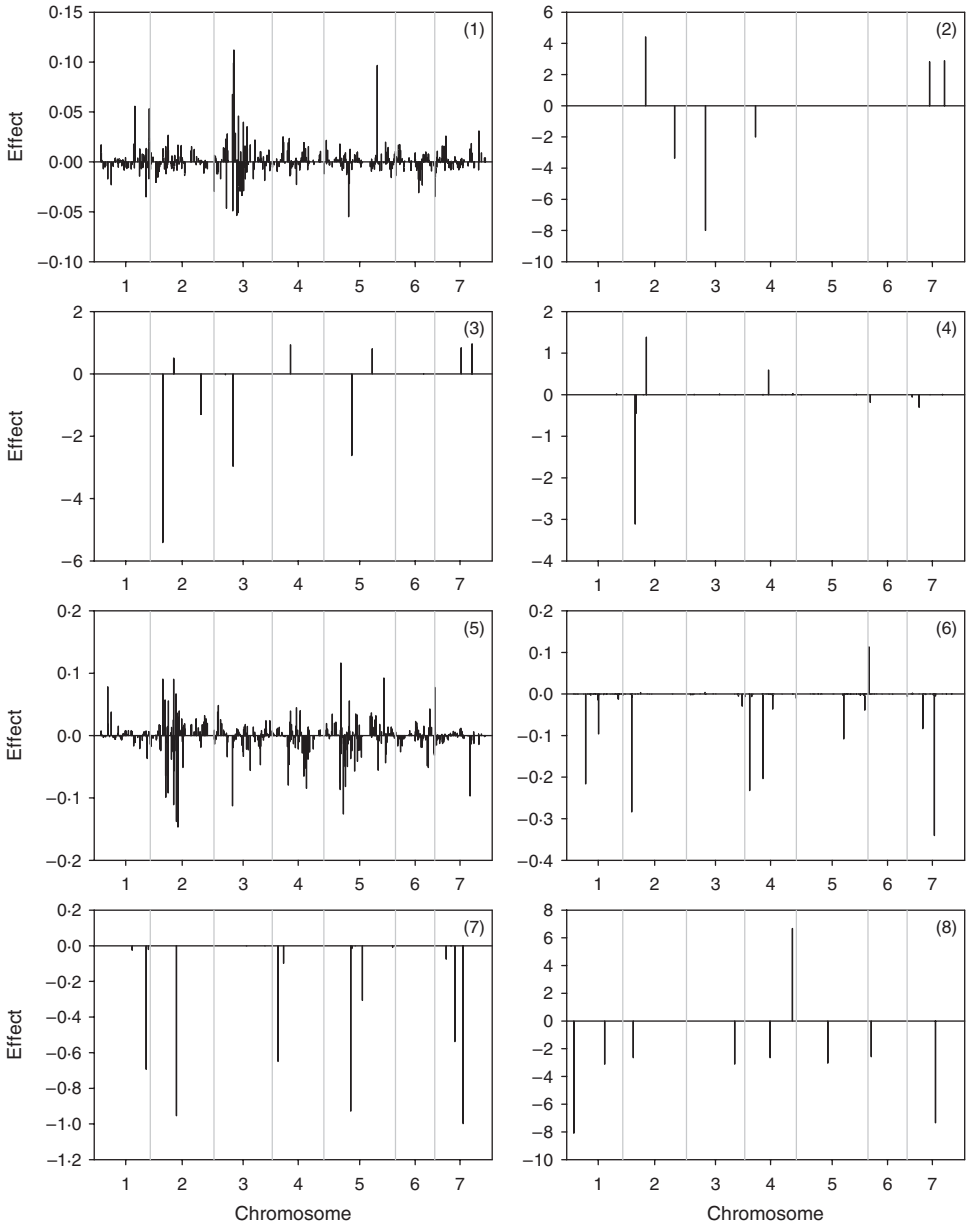


Fig. 4. Estimated QTL effects for eight quantitative traits in the barley experiment using the empirical Bayesian method. The eight traits correspond to (1) Yield, (2) Lodging, (3) Height, (4) Head, (5) Protein, (6) Extract, (7) Amylase and (8) Power. The seven chromosomes are separated by the vertical reference lines.

seven pairs of chromosomes of the barley genome, with an average marker interval of 2·23 cM. The marker map with the seven linkage groups is shown in Fig. 1. Eight quantitative traits were recorded in 16 environments. The eight traits were (1) YIELD (grain yield in MT/ha), (2) LODGING (lodging in %), (3) HEIGHT (plant height in cm), (4) HEAD (heading date after January 1), (5) PROTEIN (grain protein in %), (6) EXTRACT (malt extract in %), (7) AMYLASE (alpha amylase in 20 Deg units) and (8) POWER (diastatic power in Deg). The phenotypic values of the 150 DH lines were the averages of the 16 replications for each trait.

(ii) *MAS*

MAS utilizes results of QTL mapping. Since the marker density in the barley QTL mapping experiment is sufficiently high (2·23 cM per interval), individual marker analysis was performed using the QTL procedure in SAS (Hu & Xu, 2009). We used the permutation test (Churchill & Doerge, 1994) with 1000 permuted samples to draw the critical values for the LOD score profile for each trait. The estimated QTL effects are depicted in Fig. 2 for the eight traits. The corresponding LOD scores are given in Fig. 3 along with the permutation-generated critical values

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

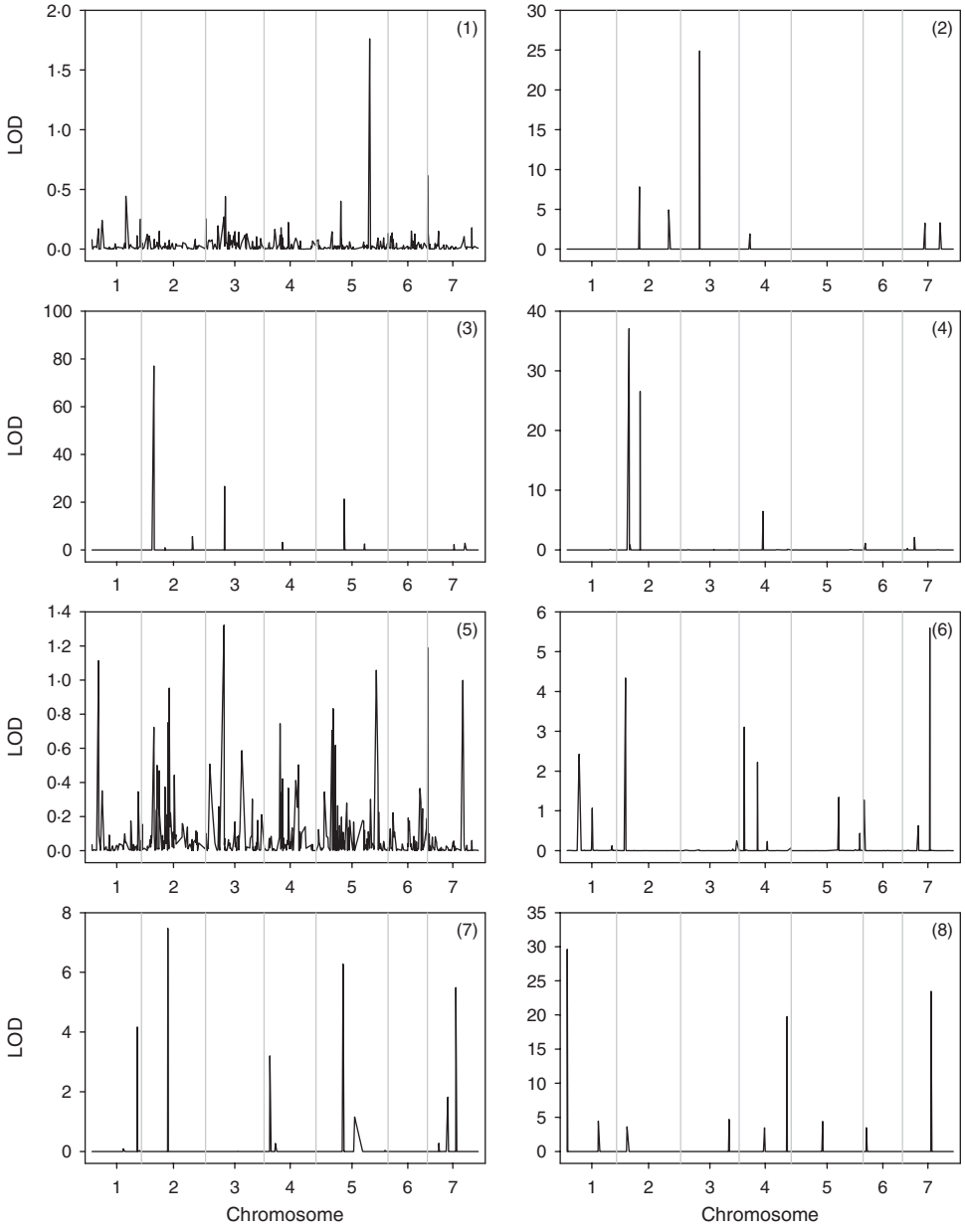


Fig. 5. The LOD scores for eight quantitative traits in the barley experiment using the empirical Bayesian method. The eight traits correspond to (1) Yield, (2) Lodging, (3) Height, (4) Head, (5) Protein, (6) Extract, (7) Amylase and (8) Power. The seven chromosomes are separated by the vertical reference lines.

at the $\alpha = 0{\cdot}01$ level. Large numbers of QTLs were detected for each of the eight traits, with an average number of 35 QTLs per trait (see Table 1). When the LOD critical values were lowered down to $\alpha = 0{\cdot}05$, the average number of markers detected per trait raised to 50.

We then used the results of the interval mapping to select these significant QTLs. Using the multiple regression method (ordinary LS), we re-estimated the QTL effects and performed the leave-one-out cross-validation analysis. The *R*-square values for $\alpha = 0{\cdot}01$ and $\alpha = 0{\cdot}05$ are given in Table 1. The average *R*-square for $\alpha = 0{\cdot}01$ and $\alpha = 0{\cdot}05$ were 0·59 and 0·56, respectively. Therefore, lowering the critical values decreased the predictability. The highest *R*-square occurred for Head with an *R*-square of 0·72. The trait extract had the lowest *R*-square of 0·48. The conclusion was that MAS using the detected QTL will be effective.

### (iii) *GS*

We now use the empirical Bayesian method (Xu, 2007) to perform GS using all markers. The hyper-parameters for each trait were set at $(\tau,\omega) = (0,0)$, corresponding to the Jeffreys' prior (Jeffreys, 1939). With the empirical Bayesian method, each marker had an estimated effect and an LOD score but all

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

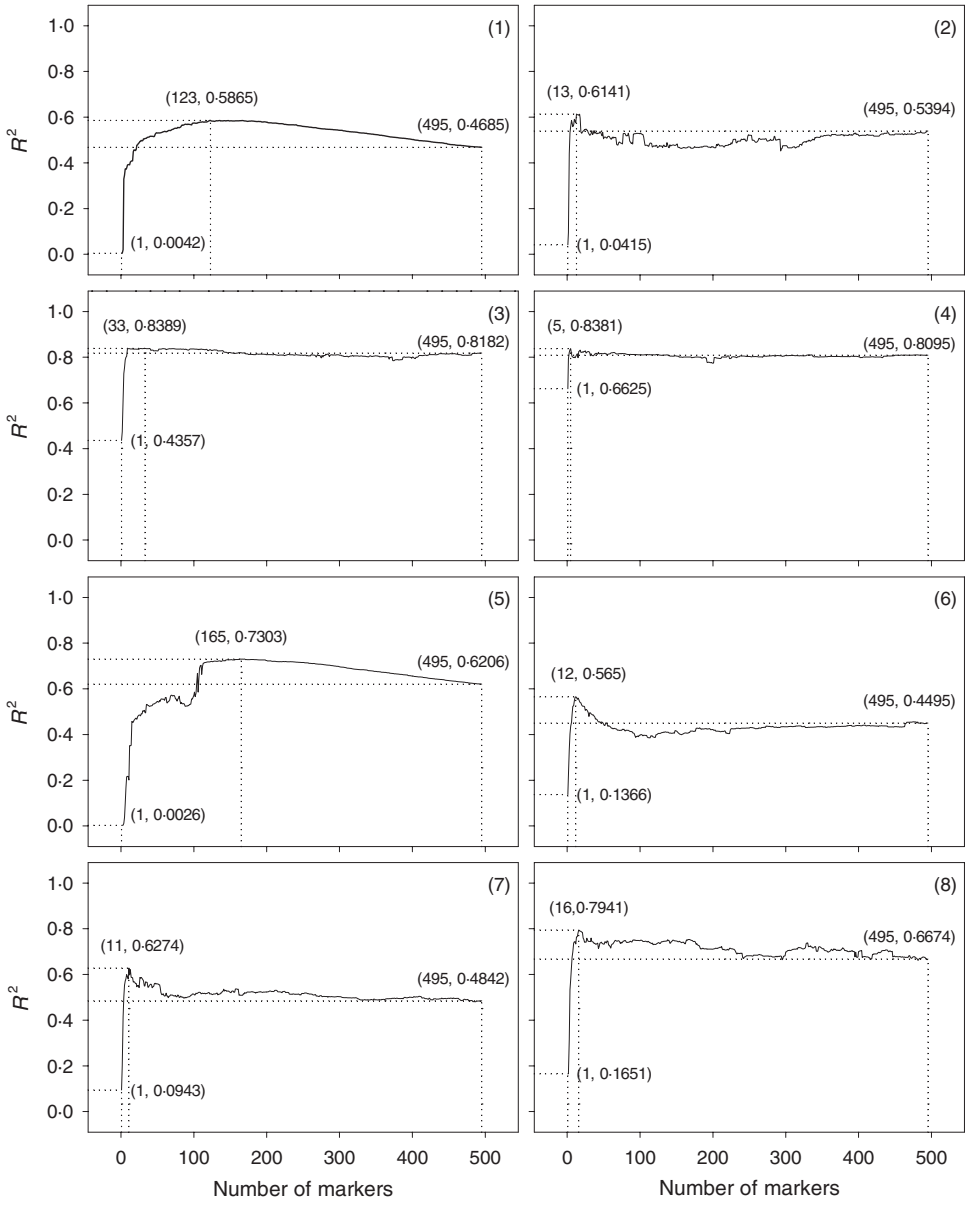


Fig. 6. The *R*-square profiles plotted against the top markers included in the model for genome prediction for eight quantitative traits in the barley experiment. The eight traits correspond to (1) Yield, (2) Lodging, (3) Height, (4) Head, (5) Protein, (6) Extract, (7) Amylase and (8) Power. The co-ordinates of three interesting points are marked for each trait. The three co-ordinates correspond to the *R*-square values for the top one marker, the optimal number of markers and all markers.

effects were estimated in a single model. The estimated effect profiles are depicted in Fig. 4. The corresponding LOD score profiles are given in Fig. 5. From these two figures, we can see clearly that the eight traits are divided into two different types, polygenic traits (Yield and Protein) and oligogenic traits (the remaining six traits). The partitioning of polygenic traits and oligogenic traits cannot be achieved using the interval mapping approach. The LOD scores of individual markers for the two polygenic traits were all smaller than the individual LOD scores for the six oligogenic traits.

We used the leave-one-out cross-validation to evaluate the accuracy of the empirical Bayesian method. When all 495 markers were included in the model, the *R*-square values ranged from 0·45 (for Extract) to 0·82 (for Height). Clearly, GS is effective for all eight traits. We also performed a variable selection approach using the full model (including all 495 markers) to rank the markers from the highest LOD score to the lowest LOD score. The number of top markers included in the model ranged from 1 to 495. For example, when the top five markers were used to evaluate the accuracy of GS, we only used

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

Table 2. *Accuracies of genome prediction* (*R-squares*) *using the empirical Bayesian method for eight quantitative traits in the barley experiment*

| Trait | *R*-square(1) (1 marker) | *R*-square(495) (495 markers) | *R*-square(opt.) (optimal number) | Optimal number of markers |
|---|---|---|---|---|
| (1) Yield | 0·0042 | 0·4685 | 0·5865 | 123 |
| (2) Lodging | 0·0415 | 0·5394 | 0·6141 | 13 |
| (3) Height | 0·4357 | 0·8182 | 0·8389 | 33 |
| (4) Head | 0·6625 | 0·8095 | 0·8381 | 5 |
| (5) Protein | 0·0026 | 0·6206 | 0·7303 | 165 |
| (6) Extract | 0·1366 | 0·4495 | 0·5650 | 12 |
| (7) Amylase | 0·0943 | 0·4842 | 0·6274 | 11 |
| (8) Power | 0·1651 | 0·6674 | 0·7941 | 16 |

*R*-square (1): The *R*-square value for model including the top marker.
*R*-square (495): The *R*-square value for model including all markers.
*R*-square (opt.): The *R*-square value for model including the optimal number of markers.
Optimal number: The number of markers that produces the maximum *R*-square value.

Table 3. *Comparison of the accuracies of prediction* (*R-squares*) *of MAS and GS*

| Trait | MAS | GS | Gain (GS-MAS) |
|---|---|---|---|
| (1) Yield | 0·5673 | 0·5865 | 0·0192 |
| (2) Lodging | 0·5593 | 0·6141 | 0·0548 |
| (3) Height | 0·6489 | 0·8389 | 0·19 |
| (4) Head | 0·7213 | 0·8381 | 0·1168 |
| (5) Protein | 0·5869 | 0·7303 | 0·1434 |
| (6) Extract | 0·4847 | 0·565 | 0·0803 |
| (7) Amylase | 0·5629 | 0·6274 | 0·0645 |
| (8) Power | 0·5782 | 0·7941 | 0·2159 |
| Average | 0·5886 | 0·6993 | 0·1106 |

these five markers to predict the genomic values. The *R*-square value for each trait formed an *R*-square profile for each trait. Different traits have different patterns for the *R*-square profiles. However, they all show a common feature: each curve starts with a low *R*-square, quickly increases to a maximum value and then progressively decreases (Fig. 6). The maximum *R*-square varied across different traits, but it was higher than the one when all markers were used for prediction. Table 2 provides a summary of the *R*-square profile for each trait. If the top marker was included in the model for prediction, only one trait (Head) had a high prediction value ($R^2 = 0{\cdot}66$), while Extract and Power each had a reasonable predictability ($R^2 = 0{\cdot}14$ and $R^2 = 0{\cdot}17$). The maximum *R*-square ranged from 0·56 (Extract) to 0·84 (Height). The number of markers that generated the maximum *R*-square values also varied across different traits. The two polygenic traits, Yield and Protein, required 123 and 165 markers, respectively, to reach the highest accuracies for prediction. Heading date is an oligogenic trait because the top five markers collectively contribute 85% of the phenotypic variance. Table 3 shows a comparison of the GS using the empirical Bayesian method and MAS using the multiple regression method (ordinary LS method). The GS had a higher *R*-square value than the MAS for every trait. The average *R*-square values for MAS and GS were 0·59 and 0·70, respectively, with an average gain of 0·11.

By definition, GS uses all markers to predict genomic values of candidate plants. However, some marker selection remains beneficial. Once the optimal number of markers is reached, including more markers appears to be slightly detrimental to GS. This conclusion is consistent with that of the Che & Xu (2010) study of flowering time in Arabidopsis. However, the decline of the accuracy by adding more markers afterwards is not dramatic, provided the marker effects are estimated using the empirical Bayesian method.

## 5. Discussion

Genome-wide epistasis may play an important role in agronomic traits. The GS tools reviewed above also apply to epistatic models. The epistatic model simply has a higher dimensionality than the additive model, and requires a fast computational algorithm. Whether or not epistatic effects are important depends on the properties of the traits and plant species. The analysis of Xu & Jia (2007) using data from a different barley crossing experiment showed that epistatic effects are not as important as additive effects. The cross in the Xu & Jia (2007) study involved two different parental lines and seven traits: (1) Heading date, (2) Height, (3) Yield, (4) Lodging, (5) Kernel weight, (6) Maturity and (7) Test weight. Four of the seven traits in the Xu & Jia (2007) study were the same as four of the eight traits in the current study. Because of the similarity of the traits and plant species in the two data sets, we do not expect to see more important roles of epistatic effects than additive effects in this data set.

https://doi.org/10.1017/S0016672310000583 Published online by Cambridge University Press

Unfortunately, the current version of the empirical Bayes program cannot handle all pair-wise interactions for 495 markers and thus we cannot test this hypothesis. Part of the reasons for the unimportant role of epistasis may be due the difficulty in detecting epistatic effects (Hill *et al*., 2008); or the importance of epistasis may vary among traits. Dudley & Johnson (2009) used the epistatic model to predict the genomic values for several quantitative traits in corn and showed significant increases in predictability over the additive model. Methods and software packages for epistatic models are available (Yi & Xu, 2002; Yi *et al*., 2003; Zhang & Xu, 2005) and have been reviewed in detail by Yi (2010).

All effective methods for GS are related to mixed model methodology. The QTL effects are always treated as random effects, either explicitly or implicitly. Therefore, understanding the mixed model methodology is fundamentally important in GS. The biggest hurdle in the mixed model approach to GS is computational speed. Efficient algorithms are always required for increased marker density.

The corresponding technology of GS for discrete traits or any traits deviating from normality is the generalized linear mixed model (GLMM; McCulloch & Neuhaus, 2005). However, many discrete traits may be analysed as if they were quantitative (Rebai, 1997; Kadarmideen *et al*., 2000), and yield similar results from the analyses using the correct GLMM. If investigators decide not to implement GLMM for discrete trait analysis, data transformation is recommended prior to the analysis. For example, the binomial trait defined as a ratio can be transformed using the Box-Cox transformation (Yang *et al*., 2006) or other simple transformations (Freeman & Tukey, 1950) prior to the analysis. GLMM is more appropriate for binary traits than other discrete traits because there is no appropriate transformation to make binary traits normal.

With the current pace of technology development, DNA sequence data will be available very soon for all agricultural crops. Sequencing the genome for all individuals in a target population is no longer a dream. With complete sequence data, pedigree analysis is no longer necessary. Pedigree analysis is one of the most difficult problems in GS. Once pedigree information becomes irrelevant, a polygenic effect is no longer required in the model for genome prediction, as it will be absorbed by the saturated markers. Therefore, GS will be easier with complete sequence data than the one with partial genomic information due to the irrelevance of pedigree information and the disappearance of the polygenic effect.

This project was supported by the Agriculture and Food Research Initiative (AFRI) of the USDA National Institute of Food and Agriculture under the Plant Genome, Genetics and Breeding Program 2007-35300-18285 to SX.

## References

Abdi, H. (2003). Partial least squares (PLS) regression. In *Encyclopedia of Social Sciences Research Methods* (ed. M. Lewis-Beck, A. Bryman & T. F. Liao). Thousand Oaks: Sage.

Beavis, W. D. (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*, p. 250–266. Washington, DC: American Seed Trade Association.

Boulesteix, A.-L. & Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics* **8**, 32–44.

Calvin, J. A. (1993). REML estimation in unbalanced multivariate variance components models using an EM algorithm. *Biometrics* **49**, 691–701.

Camus-Kulandaivelu, L., Veyrieras, J.-B., Madur, D., Combes, V., Fourmann, M., Barraud, S., Dubrevil, P., Gouesnand, B., Manicacci, D. & Charcosset, A. (2006). Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* **172**, 2449–2463.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician* **39**, 83–87.

Chahal, G. S. & Gosal, S. S. (2002). *Principles and Procedures of Plant breeding: Biotechnological and Conventional Approaches*. Boca Raton, FL: CRC Press.

Che, X. & Xu, S. (2010). Significance test and genome selection in Bayesian shrinkage analysis. *International Journal of Plant Genomics 2010*, 11 pages, doi: 10.1155/2010/893206.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251–263.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics* **22**, 67–90.

Draper, N. R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd edn. New York: John Wiley and Sons.

Dudley, J. W. (1993). Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Science* **33**, 660–668.

Dudley, J. W. & Johnson, G. R. (2009). Epistatic models improve prediction of performance in corn. *Crop Science* **49**, 763–770.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Harlow, Essex, UK: Addison Wesley Longman.

Fisher, R. A. (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh* **52**, 399–433.

Freeman, M. & Tukey, J. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics* **21**, 607–611.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems.

*Journal of the American Statistical Association* **72**, 320–338.

Hayes, P. M. & Jyambo, O. (1994). Summary of QTL effects in the Steptoe × Morex population. *Barley Genetics Newsletter* **23**, 98–143.

Hayes, P. M., Liu, B. H., Knapp, S. J., Chen, F., Jones, B., Blake, T., Fronckowiak, J., Rasmusson, D., Sorrells, M., Ullrich, S. E., Wesenberg, D. & Kleinhofs, A. (1993). Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theoretical and Applied Genetics* **87**, 392–401.

Hazel, L. (1943). The genetic basis for constructing selection indexes. *Genetics* **28**, 476–490.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

Henderson, C. R. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics* **21**, 309–310.

Hill, W. G., Goddard, M. E. & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics* **4**, e1000008.

Hoerl, A. & Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **42**, 80–86.

Hu, Z. & Xu, S. (2009). PROC QTL – A SAS procedure for mapping quantitative trait loci. *International Journal of Plant Genomics 2009*, 3 pages, doi:10.1155/2009/141234.

Jeffreys, H. (1939). *Theory of Probability*, 1st edn. Oxford: The Clarendon Press.

Kadarmideen, H., Janss, L. & Dekkers, J. (2000). Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. *Genetics Research* **76**, 305–317.

Lamkeya, K. R. & Lee, M. (1993). Focused plant improvement: towards responsible and sustainable agriculture. In *Proceedings of the 10th Australian Plant Breeding Conference*, Gold Coast, p. 18–23.

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

Legates, D. R. & McCabe, G. J. Jr (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**, 233–241.

Lindgren, F., Geladi, P. & Wold, S. (1993). The kernel algorithm for PLS. *Journal of Chemometrics* **7**, 45–59.

Lynch, M. & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753–1766.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates, Inc.

McCulloch, C. E. & Neuhaus, J. M. (2005). Generalized Linear Mixed Models. *Encyclopedia of Biostatistics*. New York: John Wiley & Sons Ltd.

Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Moose, S. P. & Mumm, R. H. (2008). Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiology* **147**, 969–977.

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S. & Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution* **41**, 56.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when the block sizes are unequal. *Biometrika* **58**, 545–554.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000*a*). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. (2000*b*). Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

Queller, D. C. & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 258–275.

Rebai, A. (1997). Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genetics Research* **69**, 69–74.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–32.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Smith, H. (1936). A discriminant function for plant selection. *Annals of Eugenics* **7**, 240–250.

Solberg, T., Sonesson, A., Woolliams, J. & Meuwissen, T. (2009). Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* **41**, 29.

ter Braak, C. J. F., Boer, M. P. & Bink, M. C. A. M. (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**, 1435–1438.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D. & Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* **28**, 286–289.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **58**, 267–288.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2**, e41.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modeling: some current developments. In *Multivariate Analysis* (ed. P. R. Krishnaiah), New York: Academic Press.

Xie, C. & Xu, S. (1997). Restricted multistage selection indices. *Genetics Selection Evolution* **29**, 193–203.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**, 513–21.

Xu, S. (2010). An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity*, doi: 10.1038/hdy.2009.180.

Xu, S. & Jia, Z. (2007). Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics* **175**, 1955–63.

Xu, S. & Muir, W. (1992). Selection index updating. *Theoretical and Applied Genetics* **83**, 451–458.

Yandell, B. S., Mehta, T., Banerjee, S., Shriner, D., Venkataraman, R., Moon, J. Y., Neely, W. W., Wu, H., von Smith, R. & Yi, N. (2007). R/qtlbim: QTL with

Bayesian interval mapping in experimental crosses. *Bioinformatics* **23**, 641–643.
Yang, R., Yi, N. & Xu, S. (2006). Box-Cox transformation for QTL mapping. *Genetica* **128**, 133–143.
Yi, N. (2010). Statistical analysis of genetic interactions. *Genetics Research* **92**, 443–459.
Yi, N. & Xu, S. (2002). Mapping quantitative trait loci with epistatic effects. *Genetical Research* **79**, 185–198.
Yi, N. & Xu, S. (2008). Bayesian Lasso for quantitative trait loci mapping. *Genetics* **179**, 1045–1055.
Yi, N., Xu, S. & Allison, D. B. (2003). Bayesian model choice and search strategies for mapping interacting quantitative trait Loci. *Genetics* **165**, 867–883.
Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208.
Zhang, Y. M. & Xu, S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **95**, 96–104.