Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 7959508, 12 pages
https://doi.org/10.1155/2020/7959508

*Research Article*

# A Two-Stage Method for Improving the Prediction Accuracy of Complex Traits by Incorporating Genotype by Environment Interactions in *Brassica napus*

**Sican Xiong**,[1,2] **Meng Wang**,[3] **Jun Zou**,[3] **Jinling Meng**,[3] **and Yanyan Liu**[2]

[1]*School of Science, East China University of Technology, Nanchang, Jiangxi 330013, China*
[2]*School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China*
[3]*National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei 430070, China*

Correspondence should be addressed to Sican Xiong; xsc060@126.com

Received 4 July 2020; Revised 23 July 2020; Accepted 24 July 2020; Published 14 August 2020

Academic Editor: Chin-Chia Wu



Improving the prediction accuracy of a complex trait of interest is key to performing genomic selection (GS) for crop breeding. For the complex trait measured in multiple environments, this paper proposes a two-stage method to solve a linear model that jointly models the genetic effects and the genotype × environment interaction ($G \times E$) effects. In the first stage, the least absolute shrinkage and selection operator (LASSO) penalized method was utilized to identify quantitative trait loci (QTL). Then, the ordinary least squares (OLS) approach was used in the second stage to reestimate the QTL effects. As a case study, this approach was used to improve the prediction accuracies of flowering time (FT), oil content (OC), and seed yield per plant (SY) in *Brassica napus* (*B. napus*). The results showed that the $G \times E$ effects reduced the mean squared error (MSE) significantly. Numerous QTL were environment-specific and presented minor effects. On average, the two-stage method, named OLS post-LASSO, offers the highest prediction accuracies (correlations are 0.8789, 0.9045, and 0.5507 for FT, OC, and SY, respectively). It was followed by the marker × environment interaction ($M \times E$) genomic best linear unbiased prediction (GBLUP) model (correlations are 0.8347, 0.8205, and 0.4005 for FT, OC, and SY, respectively), the LASSO method (correlations are 0.7583, 0.7755, and 0.2718 for FT, OC, and SY, respectively), and the stratified GBLUP model (correlations are 0.6789, 0.6361, and 0.2860 for FT, OC, and SY, respectively). The two-stage method showed an obvious improvement in the prediction accuracy, and this study will provide methods and reference to improve GS of breeding.

## 1. Introduction

In the last three decades, the development of molecular marker technology has provided numerous molecular markers for the most important species [1]. Regarding the use of molecular markers in the selection of a genetic trait, marker-assisted selection (MAS) [2] became a valuable tool in animal and plant breeding in the 1990s and works well for traits with a simple genetic architecture. However, MAS is not suitable for complex traits controlled by multiple genes, many of which have minor effects. Therefore, genomic selection (GS) as an advanced form of MAS was first propounded by Meuwissen et al. [3]. Instead of using a subset of significant markers for selection in MAS, GS uses whole genome-wide markers to predict the genome-estimated breeding values (GEBVs) of selected individuals and thus avoids biased marker effects estimates due to the detection process in MAS. However, with high-density molecular markers, the number of markers ($p$) can vastly exceed the sample size ($n$), which is referred to as a "large $p$ small $n$" problem. Thus, it is impossible to obtain the estimates of markers effects via a linear model by OLS [4].

To deal with the "large $p$ small $n$" problem, one can impose some constraints on the linear model, resulting in penalized estimation methods, such as ridge regression (RR) [5] and LASSO [6]. RR performs parameter shrinkage only, while LASSO offers both parameter shrinkage and variable selection simultaneously. Both RR and LASSO can generate

parsimonious models in the presence of a large number of predictors. The predictor size selected by LASSO is generally less than the sample size ($n$) [7], so applying OLS to the model selected by LASSO is feasible. The post-model-selection estimator was called "OLS post-LASSO" in the study of Belloni and Chernozhukov [8] and it possesses the advantage of a smaller bias than LASSO. The debiasing in OLS post-LASSO often improves the prediction error of the model [9], and this two-stage process is also known as the relaxed LASSO [10].

Bayesian methods are also applied to fit this "large $p$ small $n$" problem in GS [11]. In Bayesian inference, the marker effects are considered as random instead of fixed, and the mixed-effects model is often adopted to describe the phenotypic variation. By specifying different priors for the random marker effects, many different models, including BLUP (best linear unbiased prediction) [12], BayesA, BayesB [3], BayesC [13], and Bayesian LASSO [14], have been proposed in GS (see de los Campos et al. [11] for an overview). BLUP is a statistical procedure used to estimate the random effects and is easily obtained by solving the well-known Henderson's mixed model equations (MME) [15]. Thus, BLUP and its extensions, including rrBLUP (ridge regression BLUP) [16,17] and GBLUP [18], have become the most widely used methods in GS. Many software packages for those methods, for example, rrBLUP [17] and BGLR (Bayesian Generalized Linear Regression) [19], are freely available online. Recently, the GBLUP model has been extended to multienvironment data. For example, Lopez-Cruz et al. [20] proposed a $M \times E$ GBLUP model to accommodate the $G \times E$, and they also compared the $M \times E$ GBLUP model with the stratified (within the environment) GBLUP model. The results showed that the prediction accuracy of the $M \times E$ GBLUP model was substantially greater than the stratified GBLUP model. The significant increase in the prediction accuracy of using multienvironment models compared with single-environment analysis has been confirmed in many crops, such as maize [21] and rice [22].

*B. napus* is one of the most important oil crops worldwide. To better understand the genetic control of important agronomic traits in *B. napus*, a doubled haploid (DH) population, named TNDH population, which was derived from the F1 cross between European cultivar Tapidor and Chinese cultivar Ningyou7, was developed [23]. After several years and trial locations, the phenotypic data were collected from a multienvironment, and the TNDH population has been adopted as reference resources by the OREGIN (Oilseed Rape Genetic Improvement Network) management team. Based on the TNDH population, many QTL for the complex traits have been detected (see Shi et al. [24], etc.). Recently, a high-density genetic map of the TNDH population with a total of 2041 molecular markers was constructed [25]. Using this high-density genetic map, the genomic prediction accuracy of the FT trait in *B. napus* was evaluated via eight existing models by Li et al. [26]. However, the authors did not incorporate the $G \times E$ effects into their study. As previously noted, the $G \times E$ effects play a very important role in explaining the variation of the complex traits. Accumulating studies showed that incorporating $G \times E$ effects into the GS model could substantially increase the prediction accuracy of the complex trait. Therefore, in this study, based on the representative TNDH population, we will evaluate the performance of a two-stage approach via a linear model that jointly models the genetic effects and $G \times E$ effects. The objective of the present study is to improve the prediction accuracy of complex traits for *B. napus*. In contrast to the most commonly used $G \times E$ GS models, such as the $M \times E$ GBLUP model of Lopez-Cruz et al. [20], we treat the marker effects as fixed instead of random. We assume that the LASSO method can be used to identify the main effect and environment-specific effect QTL. Based on the identified QTL, a parsimonious linear model can be established and the OLS method is used to reestimate the QTL's effects. The performances of this two-stage approach named OLS post-LASSO and other comparison methods, including LASSO, $M \times E$ GBLUP model, and the stratified GBLUP model of Lopez-Cruz et al. [20], are evaluated in terms of prediction accuracy for FT, OC, and SY.

## 2. Materials and Methods

*2.1. Genotypic and Phenotypic Data.* A published dataset for the TNDH population was used in this study (see Luo et al. [27] for details). The TNDH population was derived from an $F_1$ progeny of a cross between a European winter cultivar "Tapidor" and a Chinese semiwinter cultivar "Ningyou7" [23]. The population comprises 182 DH lines grown at five different sites (Wuhan, Jiangling, Daye, Hangzhou, and Dali) in China for over five years (2002–2007). Combining the harvest year and the location, a total number of ten environments (year-location combinations) are available and are coded as "S3," "S4," "S5," "S6," "S7," "E7," "N3," "N4," "N6," and "N7," separately. There are a total of 2041 molecular markers for each DH line genotyped ("A" and "B" were denoted for "Tapidor" and "Ningyou7", respectively) by the *Brassica* 60K Illumina Infinium SNP array, and a total of 22 traits, including SY, OC, and FT, are collected from all the ten environments. Details of phenotypic and genotypic data and how the TNDH population was developed can be found in Luo et al. [27]. These 182 TNDH lines, the 2041 markers, and the phenotypic data for three complex traits (SY, OC, and FT) across all the ten environments were used in the present study.

*2.2. Two-Stage Approach.* Suppose that there are several $n$ population lines (individuals) cultivated in a total of $m$ environments, $y_{ki}$ is individual $i$'s trait value collected from environment $k$ ($i = 1, \ldots, n$, $k = 1, \ldots, m$). Since those $n$ individuals cultivated in different environments share the same genotypes, we use $x_{ij}$ to denote the genotype of individual $i$ at locus $j$ ($j = 1, \ldots, p$). $x_{ij} = 1$ and 0, respectively, represent A and B genotypes, where $p$ is the number of markers. For jointly modeling the genetic effect and the $G \times E$ effect, a $G \times E$ linear model by regressing phenotypes on markers across all the multiple environments can be described as follows:

$$y_{ki} = \mu + \mu_k + \sum_{j=1}^{p} x_{ij}\left(\beta_j + \alpha_{kj}\right) + \epsilon_{ki} \quad (i = 1, \ldots, n; k = 1, \ldots, m), \tag{1}$$

where $\mu$ is the overall mean (the intercept term), which is stable across the environments, and $\mu_k$ is the environment effect ($E$) that may vary by environment, $\beta_j$is the main effect across all the environments ($Q$), $\alpha_{kj}$is the environment-specific effect or equivalently the interaction effect between the $j^{th}$ locus and the $k^{th}$ environment ($Q \times E$), and $\epsilon_{ki} \sim N(0, \sigma^2)$ is the residual error. If some $\beta_j s$ or $\alpha_{kj} s$ are not equal to zero, we considered that there exist the main effects or the $Q \times E$ effects.

Model (1) is very similar to the "$M \times E$ GBLUP model" of Lopez-Cruz et al. [20]. In $M \times E$ GBLUP model, both the main effects and the environment-specific effects are treated as random effects. Also, the $M \times E$ GBLUP model does not include the overall mean and can be expressed as follows:

$$y_{ki} = \mu_k + \sum_{j=1}^{p} x_{ij}\left(\beta_j + \alpha_{kj}\right) + \epsilon_{ki} \quad (i = 1, \ldots, n, k = 1, \ldots, m), \tag{2}$$

where $\alpha_{kj} s$ are called $M \times E$ effects by Lopez-Cruz et al. [20]. Furthermore, ignoring the $M \times E$ effect and analyzing data separately in each environment, model (1) was reduced to as the "stratified GBLUP model" by Lopez-Cruz et al. [20]. The stratified GBLUP model can be expressed as follows:

$$y_{ki} = \mu_k + \sum_{j=1}^{p} x_{ij}\beta_{kj} + \epsilon_{ki} \quad (i = 1, \ldots, n, k = 1, \ldots, m), \tag{3}$$

where $\beta_{kj}$is the effect of the $j^{\text{th}}$ marker on the $k^{\text{th}}$ environment.

In the study, marker effects, including the main effect and the environment-specific effect, are considered as fixed instead of random. Considering that the number of parameters in model (1) is often larger than the sample size in GS, a two-stage approach, that is, OLS post-LASSO, was used to solve this problem, and the approach can be described as follows.

In the first stage, the LASSO method was used to select markers that have nonzero effects or equivalently detect QTL from the markers as pointed out by Zhang et al. [28]. In matrix form, model (1) can be expressed as

$$y = \mu 1_{mn} + 1_n \otimes I_m \mu + x \otimes 1_m \beta + x \otimes I_m \alpha + \varepsilon, \tag{4}$$

where $\otimes$ is the Kronecker product of two matrices and $1_{mn}$, $1_n$, and $1_m$ are the vectors of ones of order $mn$, $n$, and $m$, respectively. $I_m$ is the identity matrix of order $m$, $y = (y_{11}, \ldots, y_{1n}, \ldots, y_{m1}, \ldots, y_{mn})^T$ are the phenotypic values across environments, $x = (x_{ij})_{n \times p}$ is the genotype matrix, $\mu = (\mu_1, \ldots, \mu_m)$ are the environmental effects, $\beta = (\beta_1, \cdots, \beta_p)^T$ are the main marker effects, $\alpha = (\alpha_{11}, \ldots, \alpha_{1p}, \ldots, \alpha_{m1}, \ldots, \alpha_{mp})^T$ are the environment-specific marker effects, and $\epsilon = (\epsilon_{11}, \ldots, \epsilon_{1n}, \ldots, \epsilon_{m1}, \ldots, \epsilon_{mn})^T$is the residual error.

Let $Z = 1_n \otimes I_m$, $X = (x \otimes 1_m, x \otimes I_m)$, $\theta = (\beta^T, \alpha^T)^T$, and $K = (m+1)\, p$ is the number of columns in matrix $X$; then $Z$ is an $mn \times m$ matrix, $X$ is an $mn \times K$ matrix, and $\theta$ is a $K$-dimensional column vector. Matrices $Z$ and $X$ can be partitioned by columns, such that $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$ and $X = (X_1, \ldots, X_K)$, and vector $\theta$ can be written as $\theta = (\theta_1, \cdots, \theta_K)^T$. Then, model (4) can be expressed as

$$y = \mu 1_{mn} + \sum_{k=1}^{m} Z_k \mu_k + \sum_{l=1}^{K} X_l \theta_l + \varepsilon. \tag{5}$$

Equation (5) is the standard form of a linear model. Considering that the number of predictors ($K + m$) in model (5) is often much larger than the sample size ($mn$) in the context of $G \times E$, we use the LASSO method to solve the model first. The LASSO estimator of model (5) can be obtained by minimizing

$$\frac{1}{2}\left\| y - \left( \mu 1_{mn} + \sum_{k=1}^{m} Z_k \mu_k + \sum_{l=1}^{K} X_l \theta_l \right) \right\|^2 + \lambda \sum_{l=1}^{K} |\theta_l|, \tag{6}$$

where $\| \cdot \|$ denotes the $l_2$-norm and $\lambda \geq 0$is a tuning parameter. The tuning parameter can be selected by $k$-fold cross-validation, for example, using 10-fold cross-validation. As noted by many studies, such as Zou and Hastie [7], the number of nonzero effects selected by LASSO is generally less than the sample size, that is, $mn$ for model (6). Therefore, after LASSO, OLS regression using the selected QTL is feasible and possesses some advantages, especially reducing the shrinkage bias of LASSO [8].

Thus, in the second stage, we adopt the OLS to reestimate the QTL effects. This two-stage method was referred to as OLS post-LASSO by Belloni and Chernozhukov [8].

*2.3. Full Data Analysis.* To estimate the marker effects and show $G \times E$ relevance and structure in the dataset, model (5) was first fitted to the full dataset using OLS post-LASSO. As noted previously, the markers with nonzero main effects and environment-specific effects by LASSO in the first stage were reported as QTL. Based on the selected QTL, the OLS method was used to reestimate its values. The reestimated values from OLS are reported as the final estimated effects of QTL. We use the $t$-test to test whether the corresponding reestimated effect of each QTL is equal to zero or not. If the $p$ value of the $t$-test is less than 0.05, the corresponding QTL is reported as significant QTL. Otherwise, it is reported as a nonsignificant QTL. For the nonsignificant QTL, the corresponding effects are not significantly different from zero or, equivalently, the corresponding effects are ignorable. Meanwhile, the OLS approach can produce the corresponding standard errors (S.E.) of the estimate for the parameters. Based on the standard errors of the estimate for the QTL's effect, we can construct the 95% confidence intervals for the estimate of QTL's effect, including the main effects and the environment-specific

effects. The 95% confidence intervals are calculated by the estimated effects plus or minus 1.96 times the standard errors.

For the linear regression, R-squared is often reported as the measure that represents the proportion of the variation in the dependent variable explained by the model. However, R-squared always does not decrease as more predictors are added to the model. Thus, R-squared cannot be used to measure the contributions of each predictor. The adjusted R-squared does not increase as more predictors are added; thus it is chosen as the measure to evaluate the contribution of each component of the model. However, we cannot interpret the adjusted R-squared the same as R-squared. Note that the adjusted R-squared is equal to the percentage of the decrement of the MSE from the null model that only contains the intercept term to the alternative model that contains both the intercept term and other components of the model. Therefore, we calculate the MSE of both the null model and the alternative model. Meanwhile, we calculate the decrement and the percentage of decrement between them. To better understand the contributions of the three components of the model, that is, $E$, Q, and $Q \times E$, five alternative models incorporating the three components of the model and its combinations, $E + Q$ and $E + Q + Q \times E$, are evaluated in the article. Those alternative models with higher values of the adjusted R-squared, that is, higher percentages of the decrement of the MSE from the null model to the corresponding alternative model, are the better models. The corresponding components included in the better models would play an important role in prediction.

*2.4. Randomly Splitting the Data for Assessing Prediction Accuracy.* For comparison, the existing "$M \times E$ GBLUP model" and "stratified GBLUP model" mentioned above are chosen as comparison methods. Meanwhile, the $n = 182$ TNDH lines across all the ten environments (i.e., $m = 10$) are chosen as a working example for assessing the prediction accuracy of the two-stage method and the comparison methods. Based on the 182 TNDH lines, for each complex trait, that is, SY, OC, and FT, we merge the phenotypic datasets as a long vector just as described in the methodology from all the ten environments into one dataset. After merging, the sample size is enlarged to 1820 ($= nm = 182 \times 10$). Then, for each merged dataset, we randomly partition it into training and testing datasets at a proportion of 2 : 1. This random partition is repeated 100 times, resulting in a total of 100 random training datasets and the corresponding 100 random testing datasets. The marker effects are estimated on each training dataset across environments using LASSO, OLS post-LASSO, and $M \times E$ GBLUP model and within each environment using stratified GBLUP model. The GEBVs are computed in the corresponding testing dataset across environments using estimated LASSO, OLS post-LASSO, and $M \times E$ GBLUP model and within each environment using the estimated stratified GBLUP model. Then, we calculate the correlation between the GEBVs and the observed phenotypes for each trait within each environment. Taking the average across 100 replicated partitions, we obtain the average correlation and report it as prediction accuracy within each environment. Meanwhile, the standard deviation (SD) of the sampling distribution of the prediction accuracy among 100 replicated partitions is also reported to indicate the deviation of the accuracy.

*2.5. Software.* The minimizing problem of equation (6) can be efficiently solved by Least Angle Regression [29] in *R* software [30] using "lars" package or Alternating Direction Method of Multipliers (ADMM) [31] algorithms in MATLAB software using "lasso" function, which is used in the present study. The $M \times E$ GBLUP model (2) and stratified GBLUP model (3) are implemented using the *R* package BGLR [19].

## 3. Results

*3.1. Marker Effects.* The number of detected QTL and the frequency analysis of significant or nonsignificant QTL are reported in Table 1. From Table 1, we can see that the total number of QTL with main effects varied across traits. There are a total of 46, 77, and 26 QTL with main effects for FT, OC, and SY, respectively, and there are also a total of 231, 237, and 146 QTL with environment-specific effects for FT, OC, and SY, respectively. For main marker effects, fewer of them have significantly nonzero effects, and the percentages of significant QTL are 39.13%, 32.47%, and 42.31% for FT, OC, and SY, respectively. Equivalently, most identified main effect QTL by LASSO have small or ignorable effects, and the percentages of nonsignificant QTL are 60.87%, 67.53%, and 57.69% for FT, OC, and SY, respectively. For environment-specific marker effects, 15.58%, 16.03%, and 6.85% for FT, OC, and SY, respectively, have effects significantly different from zero. Thus, most identified environment-specific effects QTL by LASSO have small or ignorable effects.

Figures 1–3 show the point estimates and 95% confidence interval (95% CI) of marker main and environment-specific effects along the chromosomes. The vertical green confidence intervals that overlap the horizontal line of zero contain the value of zero; thus, the corresponding marker effects are nonsignificant. The vertical blue confidence intervals that represent the corresponding marker effects are statistically significant under the significance level of 0.05. As noted from these figures, most of the main and environment-specific marker effects are small and not significantly different from zero. Figures 1(c)–3(c) show the standard error (SE) of environment-specific marker effects for the same detected environment-specific QTL. The positive values (blue lines) of S.E. indicate that the corresponding QTL have environment-specific effects in multiple environments. Figures 1–3 show that few environment-specific effects QTL interact with multiple environments, and most of them display their effects in only one environment.

Table 1: The total number of QTL detected by LASSO and the frequency of significant and nonsignificant QTL by OLS post-LASSO (significance level $\alpha$ =0.05).

| Trait | Effect | Total | Significant | Nonsignificant |
|---|---|---|---|---|
| FT | Main | 46 | 18 | 28 |
| | Environment-specific | 231 | 36 | 195 |
| OC | Main | 77 | 25 | 52 |
| | Environment-specific | 237 | 38 | 199 |
| SY | Main | 26 | 11 | 15 |
| | Environment-specific | 146 | 10 | 136 |

FT, flowering time; OC, oil content; SY, seed yield per plant; QTL: quantitative trait loci; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least squares.

*3.2. Decrement of MSE.* The decrements of the MSE from the null model that only includes the intercept term to the alternative model that includes both the intercept term and one of the components of $E$, Q, $Q \times E$, $E + Q$, and $E + Q + Q \times E$ are reported in Table 2.

From Table 2, we can see that the MSE of the null model is 198.4734 for FT, 6.2661 for OC, and 0.3503 for SY, respectively. After adding the component of the $Q \times E$ into the model, the decrement of the MSE is 189.3257 for FT, 4.6635 for OC, and 0.2143 for SY, respectively, and the corresponding percentage of the decrement, that is, the adjusted R-squared, is 95.3910% for FT, 74.4235% for OC, and 61.1769% for SY, respectively. That means $Q \times E$ plays a key role in the model. If adding the combination of $E$ and $Q$ into the null model, the corresponding percentage of the decrement of the MSE is 98.2912% for FT, 74.6738% for OC, and 62.0056% for SY, respectively, which is slightly larger than that of the alternative model that contains both intercept and $Q \times E$. The percentage of the decrement of MSE for the full model that contains all of the three components, $E + Q + Q \times E$, has the highest values of 98.6731%, 88.0751%, and 66.6784% for FT, OC, and SY, respectively. Thus, the full model is most appropriate to use to predict the complex traits.

Another interesting finding is that the percentage of the decrement of MSE for the model with combined components, such as $E + Q$ and $E + Q + Q \times E$, is not equal to the sum of the percentages of decrement of the MSE for all separated models that contain only one of them. This is because the main effects QTL and the environment-specific effects QTL are highly correlated. Correlations among them can change the percentages of the decrement of the MSE dramatically from what they would be in a separate model.

*3.3. Prediction Accuracy.* Respectively, the prediction accuracies of FT, OC, and SY are evaluated (Tables 3–5). From Tables 3–5, we can see that the highest average prediction accuracy among ten environments was obtained using the OLS post-LASSO method (average correlations are 0.8789, 0.9045, and 0.5507 for FT, OC, and SY, respectively). This two-stage method is followed by the $M \times E$ GBLUP model (average correlations are 0.8347, 0.8205, and 0.4005 for FT, OC, and SY, respectively). For FT and OC, the third performing method is the LASSO approach (average correlations are 0.7583 and 0.7755 for FT and OC, respectively). However, for SY, the stratified GBLUP method is the third performing method (average correlation is 0.2860). Thus, on average, the two-stage method always performs the best in prediction accuracy for all the three complex traits.

Also, although the performances of various methods vary in different environments, the OLS post-LASSO method is advantageous in all the 10 environments except for "N6" for FT and "S5" for SY. The OLS post-LASSO achieves its best accuracy in "S4" for FT (correlation is 0.9333), in "S7" for OC (correlation is 0.9188), and "N6" for SY (correlation is 0.7039). For FT, in the environment "N6," the $M \times E$ GBLUP model has higher prediction accuracy than the OLS post-LASSO method (correlations of 0.8512 and 0.8270 for the $M \times E$ GBLUP model and OLS post-LASSO method, respectively). For SY, in the environment "S5," the $M \times E$ GBLUP model also has higher prediction accuracy than the OLS post-LASSO method (correlations of 0.3924 and 0.2285 for the $M \times E$ GBLUP model and OLS post-LASSO method, respectively).

Generally, the LASSO method yielded lower prediction accuracies compared with the $M \times E$ GBLUP method for FT, OC, and SY, whereas the OLS post-LASSO method, which refits the model again based on the QTL identified by the LASSO method, outperforms the $M \times E$ GBLUP model. The improvement of prediction accuracy from LASSO to OLS post-LASSO is significant. For example, the average prediction accuracy of the $M \times E$ GBLUP model (correlation is 0.8347) for FT locates outside the 95% confidence interval of OLS post-LASSO ($0.8789 \pm 1.96 \times 0.0038 = [0.8715, 0.8864]$) (Table 3). In other words, the probability of the difference between the average prediction accuracies of the OLS post-LASSO and $M \times E$ GBLUP model is less than 0.05. Thus, the improvement is significant for FT, and this finding is also true for OC and SY as shown in Tables 4 and 5, respectively.

## 4. Discussion

Since GS was proposed by Meuwissen et al. [3] in 2001, numerous studies have been performed to increase the prediction accuracy of the trait of interest, and numerous approaches have been proposed for GS in different situations, especially BLUP type methods. As one of the derivations of the BLUP method, the GBLUP method has become a commonly used GS method and has shown success in many situations, such as in the presence of the $G \times E$. In this study, we established a general $G \times E$ linear model to simultaneously model the genetic effects and the $G \times E$ effects. By treating the marker effects, including the main marker effects across all environments and the environment-specific marker effects, as fixed instead of random, a two-stage method named OLS post-LASSO was used to solve the model and obtain a genomic prediction.

The OLS method was also used by Meuwissen et al. [3] but not in the context of $G \times E$. For using the OLS method in GS, the biggest effects selected by some procedure, such as single segment regression analysis performed by Meuwissen et al. [3], were included. However, this stepwise procedure

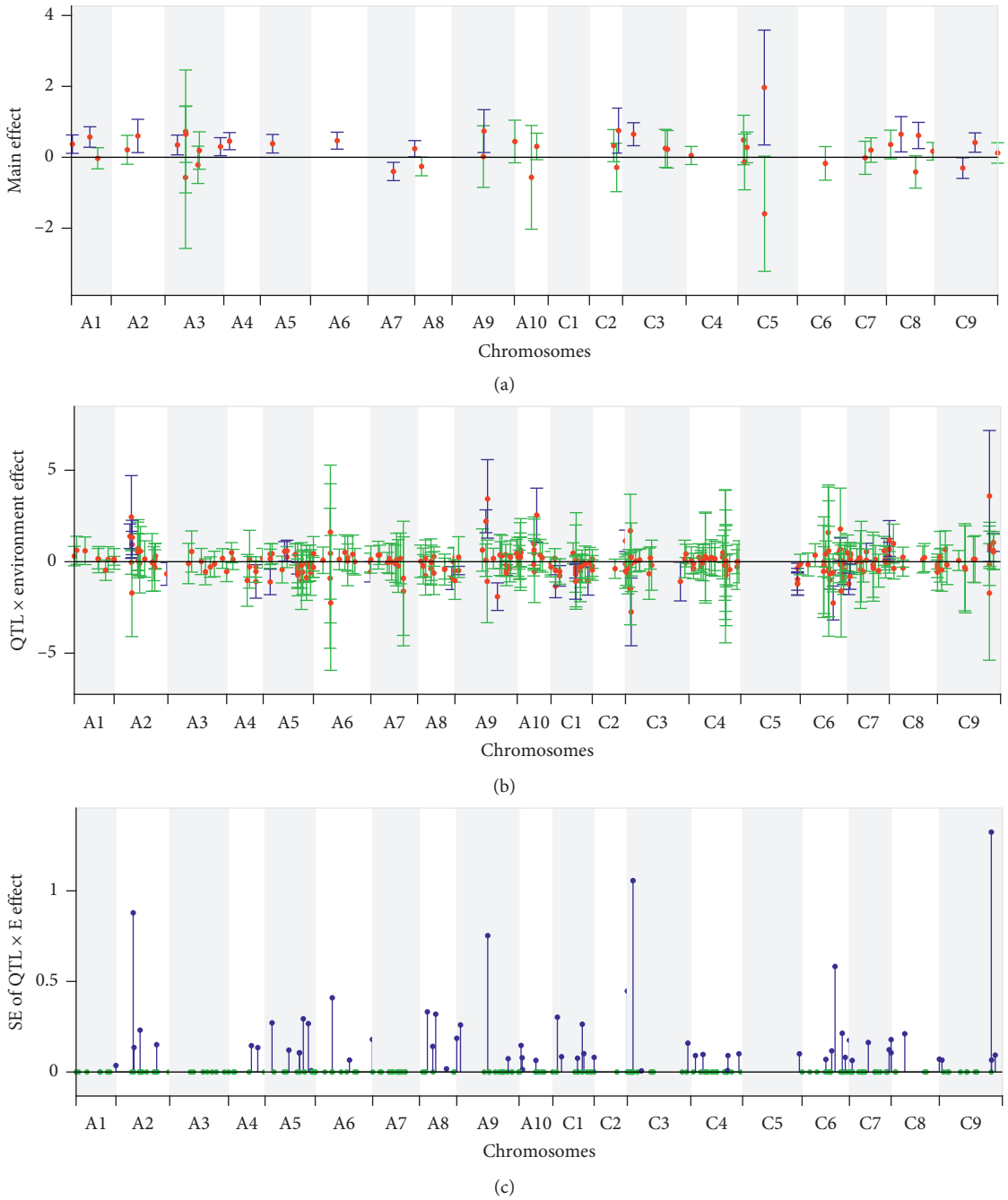



Figure 1: The point estimates and 95% confidence intervals (95% CI) for marker main and environment-specific effects and the standard error (SE) of environment-specific effects for FT. (a) Point estimates (red dots) and 95% CI for main effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (b) Point estimates (red dots) and 95% CI for environment-specific effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (c) SE for environment-specific effects (the blue and green stems represent environment-specific effects found in multiple environments or only one environment, respectively).

tends to overestimate the marker effects and cause lower prediction accuracy. This issue also exists in the context of QTL mapping using linkage disequilibrium (LD) or linkage analysis, especially in the context of $G \times E$. In the study, we adopt the LASSO method to estimate the effects of all the markers simultaneously in the first stage. As we know, the LASSO method can shrink the marker effects estimates precisely and select the biggest effects. When we have done QTL selection by LASSO beforehand, the OLS estimates are no longer unbiased in the second stage. Therefore, the two-stage method can mitigate this issue in GS, especially in the context of $G \times E$.

Table 2: The decrement and its percentages of the MSE between the null model and the alternative model.

| Trait | Type | Null model (intercept) | Alternative model (adding components as below) | | | | |
|---|---|---|---|---|---|---|---|
| | | | $E$ | $Q$ | $Q\times E$ | $E+Q$ | $E+Q+Q\times E$ |
| FT | MSE[a] | 198.4734 | 17.0725 | 196.5320 | 9.1477 | 3.3914 | 2.6336 |
| | Decrement[b] | — | 181.4009 | 1.9414 | 189.3257 | 195.0819 | 195.8397 |
| | % of decrement[c] | — | 91.3981 | 0.9781 | 95.3910 | 98.2912 | 98.6731 |
| OC | MSE | 6.2661 | 4.4863 | 4.1189 | 1.6027 | 1.5870 | 0.7472 |
| | Decrement | — | 1.7798 | 2.1473 | 4.6635 | 4.6792 | 5.5189 |
| | % of decrement | — | 28.4040 | 34.2680 | 74.4235 | 74.6738 | 88.0751 |
| SY | MSE | 0.3503 | 0.2259 | 0.3284 | 0.1360 | 0.1331 | 0.1167 |
| | Decrement | — | 0.1245 | 0.0219 | 0.2143 | 0.2172 | 0.2336 |
| | % of decrement | — | 35.5276 | 6.2476 | 61.1769 | 62.0056 | 66.6784 |

[a]Mean square error; [b]the decrement of MSE between the null model that only includes the intercept term and the alternative model that includes both intercept and the corresponding components; [c]the percentage of the decrement of the MSE between the null model and the alternative model or, equivalently, the adjusted R-squared; FT, flowering time; OC, oil content; SY, seed yield per plant; $E$, the environment effect; $Q$, the main effect of locus; Q×E, the interaction effect between the locus and the environment.

Table 3: The prediction accuracy (standard deviation, SD) of four methods for FT.

| Environment | LASSO | OLS post-LASSO | Stratified GBLUP | $M\times E$ GBLUP |
|---|---|---|---|---|
| E7 | 0.6030 (0.0077) | 0.9008 (0.0040) | 0.6906 (0.0056) | 0.7563 (0.0049) |
| N3 | 0.7871 (0.0045) | 0.8774 (0.0036) | 0.6724 (0.0064) | 0.8656 (0.0028) |
| N4 | 0.8091 (0.0040) | 0.8924 (0.0037) | 0.7032 (0.0057) | 0.8882 (0.0026) |
| N6 | 0.7864 (0.0056) | 0.8270 (0.0048) | 0.6791 (0.0057) | 0.8512 (0.0040) |
| N7 | 0.7634 (0.0063) | 0.9048 (0.0049) | 0.6429 (0.0062) | 0.8284 (0.0053) |
| S3 | 0.8124 (0.0032) | 0.8706 (0.0037) | 0.6907 (0.0052) | 0.8536 (0.0023) |
| S4 | 0.7935 (0.0039) | 0.9333 (0.0023) | 0.7497 (0.0045) | 0.8948 (0.0025) |
| S5 | 0.7563 (0.0053) | 0.7889 (0.0047) | 0.6038 (0.0076) | 0.7659 (0.0051) |
| S6 | 0.8048 (0.0042) | 0.8858 (0.0033) | 0.6916 (0.0055) | 0.8630 (0.0031) |
| S7 | 0.6675 (0.0071) | 0.9083 (0.0031) | 0.6646 (0.0062) | 0.7796 (0.0044) |
| Average | 0.7583 (0.0052) | 0.8789 (0.0038) | 0.6789 (0.0058) | 0.8347 (0.0037) |

FT, flowering time; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least squares; GBLUP, genomic best linear unbiased prediction; $M\times E$, marker × environment interaction.

Table 4: The prediction accuracy (standard deviation, SD) of four methods for OC.

| Environment | LASSO | OLS post-LASSO | Stratified GBLUP | $M\times E$ GBLUP |
|---|---|---|---|---|
| E7 | 0.7552 (0.0060) | 0.8665 (0.0041) | 0.5956 (0.0075) | 0.7784 (0.0057) |
| N3 | 0.8239 (0.0043) | 0.9097 (0.0030) | 0.6721 (0.0065) | 0.8514 (0.0033) |
| N4 | 0.7594 (0.0049) | 0.9153 (0.0027) | 0.6261 (0.0078) | 0.8048 (0.0045) |
| N6 | 0.8255 (0.0038) | 0.9130 (0.0025) | 0.7300 (0.0055) | 0.8729 (0.0030) |
| N7 | 0.7625 (0.0073) | 0.8866 (0.0044) | 0.5837 (0.0096) | 0.7859 (0.0065) |
| S3 | 0.8127 (0.0037) | 0.9175 (0.0021) | 0.6767 (0.0063) | 0.8336 (0.0034) |
| S4 | 0.7809 (0.0048) | 0.9114 (0.0028) | 0.6600 (0.0068) | 0.8372 (0.0042) |
| S5 | 0.7189 (0.0065) | 0.9030 (0.0029) | 0.5806 (0.0075) | 0.7990 (0.0040) |
| S6 | 0.7174 (0.0045) | 0.9037 (0.0028) | 0.5625 (0.0070) | 0.7846 (0.0035) |
| S7 | 0.7986 (0.0047) | 0.9188 (0.0026) | 0.6741 (0.0059) | 0.8572 (0.0034) |
| Average | 0.7755 (0.0051) | 0.9045 (0.0030) | 0.6361 (0.0070) | 0.8205 (0.0042) |

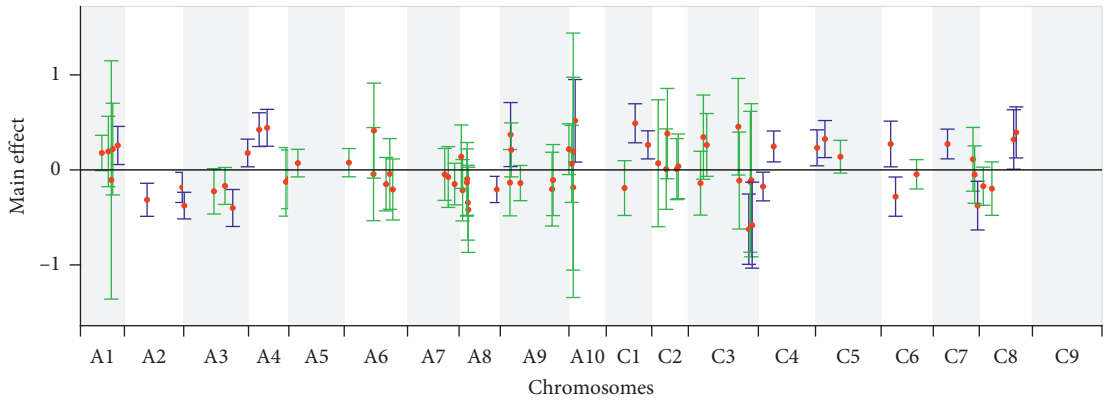
OC, oil content; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least squares; GBLUP, genomic best linear unbiased prediction; $M\times E$, marker × environment interaction.

Using the TNDH population as a working example, prediction accuracy was compared for three complex traits (FT, OC, and SY) using four methods, namely, OLS post-LASSO, LASSO, $M\times E$ GBLUP model, and stratified GBLUP model. Generally, the two-stage method, that is, OLS post-LASSO, achieved the highest prediction accuracies on average across environments and also achieved the highest prediction accuracies within most of the ten environments. The $M\times E$ GBLUP model performed worse than the two-stage method but better than the other two approaches, namely, the LASSO and the stratified GBLUP model. Although LASSO performed worse than the $M\times E$ GBLUP model, the OLS after LASSO, that is, OLS post-LASSO, outperformed the $M\times E$ GBLUP model, and the improvement in prediction accuracy is significant. The results show that the OLS post-LASSO could always outperform the

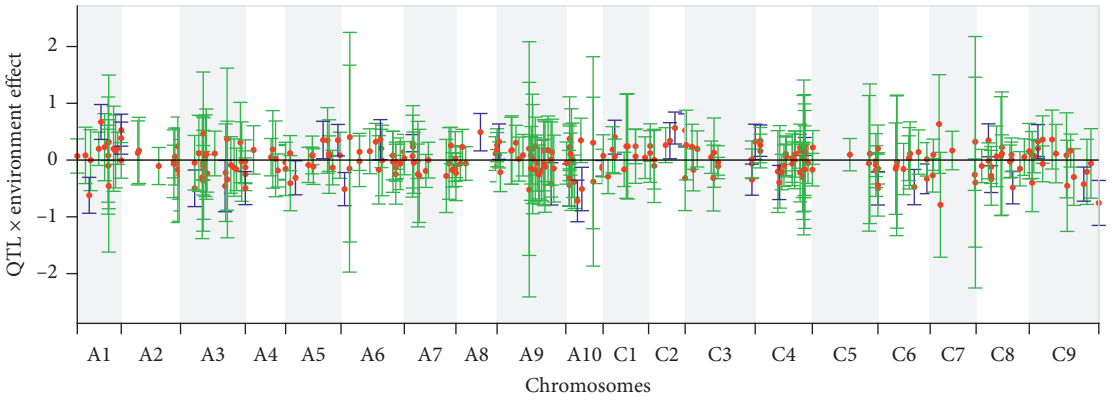
Table 5: The prediction accuracy (standard deviation, SD) of four methods for SY.

| Environment | LASSO | OLS post-LASSO | Stratified GBLUP | $M \times E$ GBLUP |
|---|---|---|---|---|
| E7 | 0.4952 (0.0082) | 0.5735 (0.0075) | 0.4418 (0.0076) | 0.4767 (0.0071) |
| N3 | 0.1600 (0.0104) | 0.6958 (0.0092) | 0.1336 (0.0097) | 0.2322 (0.0104) |
| N4 | 0.2635 (0.0115) | 0.3816 (0.0089) | 0.1509 (0.0110) | 0.3049 (0.0098) |
| N6 | 0.2811 (0.0097) | 0.7039 (0.0092) | 0.4038 (0.0089) | 0.4653 (0.0097) |
| N7 | 0.2224 (0.0106) | 0.6312 (0.0089) | 0.3251 (0.0079) | 0.4464 (0.0087) |
| S3 | 0.1351 (0.0111) | 0.4698 (0.0108) | 0.0907 (0.0110) | 0.2088 (0.0114) |
| S4 | 0.2887 (0.0109) | 0.6395 (0.0082) | 0.3056 (0.0099) | 0.4269 (0.0094) |
| S5 | 0.1232 (0.0093) | 0.2285 (0.0113) | 0.3595 (0.0106) | 0.3924 (0.0079) |
| S6 | 0.4142 (0.0100) | 0.6203 (0.0074) | 0.3973 (0.0099) | 0.5919 (0.0070) |
| S7 | 0.3351 (0.0093) | 0.5627 (0.0092) | 0.2517 (0.0099) | 0.4597 (0.0087) |
| Average | 0.2718 (0.0101) | 0.5507 (0.0091) | 0.2860 (0.0096) | 0.4005 (0.0090) |

SY, seed yield per plant; LASSO, least absolute shrinkage and selection operator; OLS, ordinary least squares; GBLUP, genomic best linear unbiased prediction; $M \times E$, marker × environment interaction.

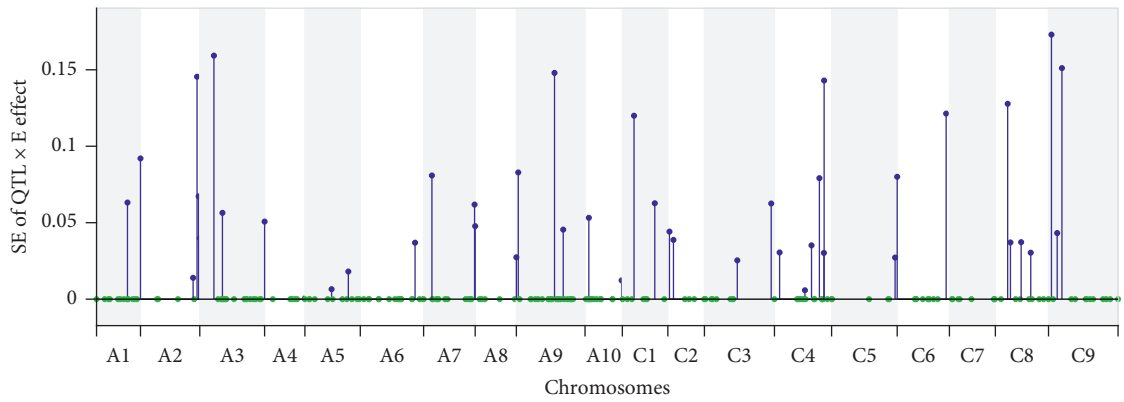



(a)




(b)

Figure 2: Continued.




(c)

Figure 2: The point estimates and 95% confidence intervals (95% CI) for marker main and environment-specific effects and the standard error (SE) of environment-specific effects for OC. (a) Point estimates (red dots) and 95% CI for main effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (b) Point estimates (red dots) and 95% CI for environment-specific effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (c) SE for environment-specific effects (the blue and green stems represent environment-specific effects found in multiple environments or only one environment, respectively).

LASSO, whether for FT, OC, or SY. Although the advantage of the two-stage method has been reported by Belloni and Chernozhukov [8], the application in GS and its benefit in increasing predictive ability are first studied in this study.

From the computational aspects of the model, the two-stage approach took around 45 minutes to compute (Windows10 Pro with a 1.6 GHz Intel Core i5-8250U processor and 8 GB of memory) in the first stage and around 0.1 seconds to compute in the second stage. The computation time is higher than that of the stratified and the $G \times E$ GBLUP models. It took around 30 seconds and 10 minutes for the stratified and the $G \times E$ GBLUP models, respectively. However, in the case of $G \times E$, the two-stage method tends to fit easily compared to traditional methods, such as the factor-analytic method. The factor-analytic method tries to simplify a complex covariance structure and, in some cases, for example, in the case of $G \times E$, difficulty in reaching convergence [32].

As a penalized regression method, LASSO was first implemented in GS by Usai et al. [33], and its prediction performance was evaluated by many studies, such as Ogutu et al. [34] and Xu et al. [35]. LASSO as well as GBLUP, the most commonly used method in GS, always outperformed other methods, such as rr-BLUP [34] and support vector machine (SVM) [35]. Based on the FT trait dataset of TNDH population, the study of Li et al. [26] indicated that the average prediction accuracies across the ten environments varied from 0.593 to 0.651 using the existing eight models: rr-BLUP, reproducing kernel Hilbert spaces (RKHS), Bayesian LASSO, BayesA, BayesB, random forest (RF), and SVM (linear kernel and Gaussian kernel). The average prediction accuracies obtained by Li et al. [26] for the FT trait were lower than those in the four methods evaluated in the present study (average correlations are 0.8789, 0.8347, 0.7583, and 0.6789 for OLS post-LASSO, $M \times E$ GBLUP model, LASSO, and the stratified GBLUP model, respectively). The stratified GBLUP model performed similarly bad to the eight models evaluated by Li et al. [26] because those methods ignore the $G \times E$ effects in the analysis. The results of our study confirmed that incorporating $G \times E$ effects into the GS model increased prediction accuracy, which has been noted by many studies, such as Lopez-Cruz et al. [20]. In particular, the two-stage method performed the best for complex traits FT, OC, and SY.

The percentage of decrement of MSE by our model (1) (corresponding to the alternative model with $E + Q + Q \times E$) is very close to 100% for the FT trait (98.6731%). This finding indicates that our proposed model fits the FT trait dataset perfectly. However, the performance is reduced when applying the same model (1) to other traits, for example, to OC (the percentage of decrement of MSE is 88.0751%) and SY (the percentage of decrement of MSE is 66.6784%). As noted by Luo et al. [27], the FT shows very high heritability; however, the SY shows low heritability. Thus, it is reasonable that our proposed model accounts for more variation of FT but less variation of SY. However, even in the case of more complex traits, that is, the OC and SY, the prediction performance of our proposed method remains superior compared with previous approaches, like the $M \times E$ GBLUP model and other models.

Although the FT is not complex as SY, the identified number of QTL for FT is larger than that for SY (Table 1). If we only focus on the number of identified QTL, there seems to be some irrationality. We can explain the issue from at least the following two aspects. First, the identified QTL by LASSO is suggestive, and further experimental identification is required. That means the detected QTL may not be the true QTL. Second, from the first subplot (a) of Figure 1, we can see that there exists a major QTL on chromosome C5 for FT, which has the largest main marker effect, and the other significant QTL (the blue lines) have smaller main marker effects than the major QTL. However, we cannot find the

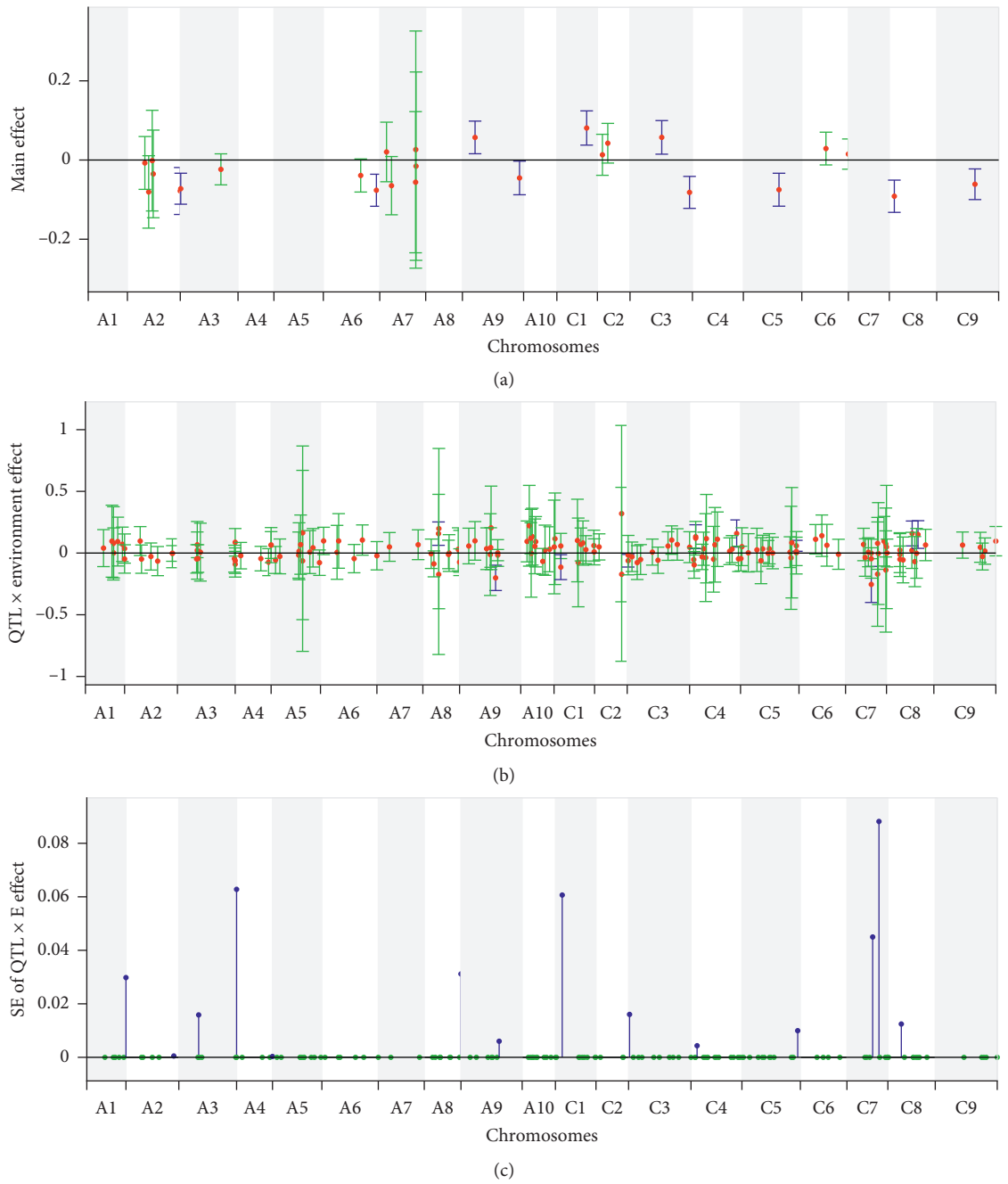



Figure 3: The point estimates and 95% confidence intervals (95% CI) for marker main and environment-specific effects and the standard error (SE) of environment-specific effects for SY. (a) Point estimates (red dots) and 95% CI for main effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (b) Point estimates (red dots) and 95% CI for environment-specific effects (the vertical blue and green CIs represent statistically significant or nonsignificant values, respectively). (c) SE for environment-specific effects (the blue and green stems represent environment-specific effects found in multiple environments or only one environment, respectively).

major QTL for OC and SY (please see subplots (a) in Figures 2 and 3). Meanwhile, from the magnitude of the absolute value of the main marker effects, we can see that it decreases from around 4 for FT to around 1.5 for OC and around 0.3 for SY. Similar patterns can be found for the environment-specific marker effects (please see the subplots (b) of Figures 1–3 for details). Thus, the finding of our study also supports that the FT trait is not as complex as OC and SY as we expected.

## Data Availability

Phenotypic and marker data used in the article can be found in Supplemental file S1.

## Conflicts of Interest

The authors declare that there are no conflicts of interest.

## Acknowledgments

The authors acknowledge support by the National Natural Science Foundation of China (NSFC) (Project nos. 31970564, 11971362, 11661003, and 11661004) and funding for the Doctoral Research of ECUT under Grant no. DHBK2018052.

## Supplementary Materials

The compressed supplementary file named "S1_TN182 Phenotypic and marker data.zip" includes an Excel file named "TN182 Phenotypic and marker data.xlsx." The Excel file includes four sheets. The "Environment" sheet shows the information about the ten environments, including the name of macro environments, the code of the experiment, and so on. The "Trait name" sheet shows the names, the abbreviation, and the measurement of the three traits which are studied in the paper. The "Phenotype" sheet shows all the phenotypic values of the three traits collected from all the ten environments. The "Genotype" sheet shows the genotype matrix of all the individuals. (*Supplementary Materials*)

## References

[1] F. Khan, "Molecular markers: an excellent tool for genetic analysis," *Journal of Molecular Biomarkers & Diagnosis*, vol. 06, no. 03, p. 233, 2015.

[2] R. Lande and R. Thompson, "Efficiency of marker-assisted selection in the improvement of quantitative traits," *Genetics*, vol. 124, no. 3, pp. 743–756, 1990.

[3] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[4] P. Pérez, G. de Los Campos, J. Crossa, and D. Gianola, "Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R," *The Plant Genome*, vol. 3, no. 2, pp. 106–116, 2010.

[5] A. E. Hoerl and R. W. Kennard, "ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[8] A. Belloni and V. Chernozhukov, "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, vol. 19, no. 2, pp. 521–547, 2013.

[9] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity*, Chapman and Hall/CRC, New York, NY, USA, 2015.

[10] N. Meinshausen, "Relaxed lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 374–393, 2007.

[11] G. de los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, "Whole-genome regression and prediction methods applied to plant and animal breeding," *Genetics*, vol. 193, no. 2, pp. 327–345, 2013.

[12] C. R. Henderson and R. L. Quaas, "Multiple trait evaluation using relatives' records," *Journal of Animal Science*, vol. 43, no. 6, pp. 1188–1197, 1976.

[13] D. Habier, R. L. Fernando, K. Kizilkaya, and D. J. Garrick, "Extension of the Bayesian alphabet for genomic selection," *BMC Bioinformatics*, vol. 12, no. 1, p. 186, 2011.

[14] T. Park and G. Casella, "The bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[15] G. K. Robinson, "That BLUP is a good thing: the estimation of random effects," *Statistical Science*, vol. 6, no. 1, pp. 15–32, 1991.

[16] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*, Cambridge University Press, Cambridge; NY, USA, 2003.

[17] J. B. Endelman, "ridge regression and other kernels for genomic selection with R package rrBLUP," *The Plant Genome*, vol. 4, no. 3, pp. 250–255, 2011.

[18] P. M. VanRaden, "Efficient methods to compute genomic predictions," *Journal of Dairy Science*, vol. 91, no. 11, pp. 4414–4423, 2008.

[19] P. Pérez and G. de los Campos, "Genome-wide regression and prediction with the BGLR statistical package," *Genetics*, vol. 198, no. 2, pp. 483–495, 2014.

[20] M. Lopez-Cruz, J. Crossa, D. Bonnett et al., "Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model," *Genes|Genomes|Genetics*, vol. 5, no. 4, pp. 569–582, 2015.

[21] M. Bandeira E Sousa, J. Cuevas, E. G. de Oliveira Couto et al., "Genomic-enabled prediction in maize using kernel models with genotype × environment interaction," *Genes|Genomes|Genetics*, vol. 7, no. 6, pp. 1995–2014, 2017.

[22] E. Monteverde, J. E. Rosas, P. Blanco et al., "Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice," *Crop Science*, vol. 58, no. 4, pp. 1519–1530, 2018.

[23] D. Qiu, C. Morgan, J. Shi et al., "A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content," *Theoretical and Applied Genetics*, vol. 114, no. 1, pp. 67–80, 2006.

[24] J. Shi, R. Li, D. Qiu et al., "Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*," *Genetics*, vol. 182, no. 3, pp. 851–861, 2009.

[25] Y. Zhang, C. L. Thomas, J. Xiang et al., "QTL meta-analysis of root traits in *Brassica napus* under contrasting phosphorus supply in two growth systems," *Scientific Reports*, vol. 6, no. 1, 2016.

[26] L. Li, Y. Long, L. Zhang et al., "Genome wide analysis of flowering time trait in multiple environments via high-throughput genotyping technique in *Brassica napus* L," *PLoS One*, vol. 10, no. 3, p. e0119425, Article ID e0119425, 2015.

[27] Z. Luo, M. Wang, Y. Long et al., "Incorporating pleiotropic quantitative trait loci in dissection of complex traits: seed yield in rapeseed as an example," *Theoretical and Applied Genetics*, vol. 130, no. 8, pp. 1569–1585, 2017.

[28] M. Zhang, K. L. Montooth, M. T. Wells, A. G. Clark, and D. Zhang, "Mapping multiple quantitative trait loci by bayesian classification," *Genetics*, vol. 169, no. 4, pp. 2305–2318, 2005.

[29] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[30] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, Austria, 2017.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[32] L. S. Peixouto, J. A. R. Nunes, and D. F. Furtado, "Factor analysis applied to the G+GE matrix via REML/BLUP for multi-environment data," *Crop Breeding and Applied Biotechnology*, vol. 16, no. 1, pp. 1–6, 2016.

[33] M. G. Usai, M. E. Goddard, and B. J. Hayes, "LASSO with cross-validation for genomic selection," *Genetics Research*, vol. 91, no. 6, pp. 427–436, 2009.

[34] J. O. Ogutu, T. Schulz-Streeck, and H. P. Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions," *BMC*, vol. 6, no. Suppl 2, 2012.

[35] Y. Xu, X. Wang, X. Ding et al., "Genomic selection of agronomic traits in hybrid rice using an NCII population," *Rice*, vol. 11, no. 1, p. 32, 2018.