entropy

MDPI

Article

# A Two-Stage Mutual Information Based Bayesian Lasso Algorithm for Multi-Locus Genome-Wide Association Studies

**Hongping Guo [1,2], Zuguo Yu [1,3,*], Jiyuan An [4], Guosheng Han [1], Yuanlin Ma [1] and Runbin Tang [1]**

1 Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan 411105, China; guohongping0501@163.com (H.G.); hangs@xtu.edu.cn (G.H.); 201590110068@smail.xtu.edu.cn (Y.M.); 201831510085@smail.xtu.edu.cn (R.T.)
2 School of Mathematics and Computer Science, Hanjiang Normal University, Shiyan 442000, China
3 School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4001, Australia
4 Centre for Tropical Crops and Biocommodities, Queensland University of Technology, Brisbane, QLD 4001, Australia; j.an@qut.edu.au
* Correspondence: yuzg@xtu.edu.cn

Received: 01 February 2020; Accepted: 10 March 2020; Published: 13 March 2020


**Abstract:** Genome-wide association study (GWAS) has turned out to be an essential technology for exploring the genetic mechanism of complex traits. To reduce the complexity of computation, it is well accepted to remove unrelated single nucleotide polymorphisms (SNPs) before GWAS, e.g., by using iterative sure independence screening expectation-maximization Bayesian Lasso (ISIS EM-BLASSO) method. In this work, a modified version of ISIS EM-BLASSO is proposed, which reduces the number of SNPs by a screening methodology based on Pearson correlation and mutual information, then estimates the effects via EM-Bayesian Lasso (EM-BLASSO), and finally detects the true quantitative trait nucleotides (QTNs) through likelihood ratio test. We call our method a two-stage mutual information based Bayesian Lasso (MBLASSO). Under three simulation scenarios, MBLASSO improves the statistical power and retains the higher effect estimation accuracy when comparing with three other algorithms. Moreover, MBLASSO performs best on model fitting, the accuracy of detected associations is the highest, and 21 genes can only be detected by MBLASSO in *Arabidopsis thaliana* datasets.

**Keywords:** GWAS; Pearson correlation; mutual information; feature screening; Bayesian Lasso

## 1. Introduction

Genome-wide association study (GWAS) has evolved to be an essential technology for exploring the genetic mechanism of complex traits [1]. It concentrates on identifying the significant single nucleotide polymorphisms (SNPs) associated with the given traits. In past years, several single-locus GWAS methods have been developed [1–5], and have detected a few variants among various traits successfully. However, they still have some drawbacks, such as the combined effects of multiple loci are ignored and the threshold in multiple test correction is hard to be determined [6].

To address these drawbacks, some classical high-dimensional statistical methods were well used in GWAS when the number of SNPs is not far more than that of individuals, such as the least absolute shrinkage and selector operator (Lasso) [7], the elastic net [8], and Bayesian Lasso [9,10]. However, the current situation is the opposite, because the number of SNPs is much larger than

that of individuals. In the case of ultrahigh-dimensional data, the aforementioned methods will fail due to the internal computational complexity. Fortunately, Fan and Lv [11] proposed a two-stage feature screening (or variable selection) method. The main idea of this method is: The dimension of features are firstly cut down by sure independence screening (SIS), and then a certain popular high-dimensional feature screening method (such as Lasso, the smoothly clipped absolute deviation (SCAD) [12], or the adaptive Lasso [13]) is used to select significant features and estimate regression coefficients simultaneously. The extension of SIS is iterative sure independence screening (ISIS), which can revive those non-negligible features that are single uncorrelated while indirectly correlated to the respond variables [11]. Instead of the Pearson correlation based SIS, statisticians have exploited some other SIS methods from different measurements, such as rank correlation [14], the distance correlation [15], the partial correlation [16] and so on. Among these methods, Pearson correlation and distance correlation based screening have been applied in GWAS successfully [6,17], and some genes associated with crop quantitative traits such as rice salt-tolerance and poplar growth have been identified [18,19]. ISIS expectation-maximization Bayesian LASSO (ISIS EM-BLASSO) [6] selects potentially associated SNPs in single-objective screening methodology based on the Pearson correlation between the SNPs and phenotype. In reality, the intrinsic heterogeneity is likely to be present in big data [20], thus the characterization of correlations via multi-objective method can bring higher power [21]. Although two-side high-dimensional genome-wide association studies (2HiGWAS) [17] efficiently selects the associated SNPs by combining Pearson correlation and distance correlation, the computational burden of constructing distance correlation is very high.

Since mutual information can detect broader classes of relationships [22], and the computational complexity is relatively low [23]. We propose to modify the screening method in the first stage of ISIS EM-BLASSO to a multi-objective one, which is based on the combination of Pearson correlation and mutual information. Then EM-Bayesian Lasso (EM-BLASSO) [10] is applied to further select SNPs and estimate the effects by shrinking the weak effects to zero, and likelihood ratio test is used to identify the true quantitative trait nucleotides (QTNs), these procedures are the same as those in the second stage of ISIS EM-BLASSO (also denoteds EM-BLASSO). We call our method a two-stage mutual information based Bayesian Lasso (MBLASSO). In order to validate the effectiveness of our method, we compare it with three GWAS methods, EM-BLASSO [10], ISIS EM-BLASSO [6] and genome-wide efficient mixed model association (GEMMA) [5]. EM-BLASSO represents the single-stage GWAS method without pre-screening, ISIS EM-BLASSO is a typical two-stage GWAS method using only Pearson correlation screening, and GEMMA is a golden standard GWAS method widely used for comparison.

## 2. Materials and Methods

### *2.1. Statistical Framework*

In this study, we consider the linear mixed genetic model [6] as follows:

$$\mathbf{y} = \mathbf{1}\mu + Q\alpha + X\beta + \varepsilon \tag{1}$$

where $\mathbf{y}$ is a $n \times 1$ phenotypic vector of quantitative trait, and $n$ is the number of individuals; $\mathbf{1}$ is a $n \times 1$ vector in which every element is equal to 1, and $\mu$ is the overall mean; $Q = (Q_1, Q_2, \ldots, Q_q)$ is a $n \times q$ matrix of fixed effects, such as the population structure, and $q$ is the number of fix effects; $\alpha$ is a $q \times 1$ vector of fixed effects; and $X$ is a $n \times p$ matrix of SNP genotype values. For each SNP, homozygous genotype are coded as 1, and $-1$, respectively, and the heterozygous ones are indicated by 0. $p$ is the number of presumed QTNs, $\beta$ is the QTN effects, and $\varepsilon \sim MVN_n(0, \sigma_e^2\mathbf{I})$ is a $n \times 1$ vector of residual error.

### 2.2. Simulation Experiments

To assess the performance of methods, we considered simulation scenarios based on the *Arabidopsis thaliana* datasets consisting of 216,130 SNPs, 199 accessions, and 107 phenotype traits [24]. For genotype simulation, we randomly selected 10,000 SNPs, 2000 for each of the five chromosomes, i.e., 11,226,256–12,038,776 bp on Chr.1, 5,045,828–6,6412,875 bp on Chr.2, 1,916,588–3,196,442 bp on Chr.3, 2,232,796–3,3143,893 bp on Chr.4, and 19,999,868–21,039,406 bp on Chr.5. Additionally, we generated the phenotype simulation data with sample size 199 from three different scenarios, and undertook 1000 times for each simulation. Six QTNs were assumed to be genuine; their heritabilities were set as 0.10, 0.05, 0.05, 0.15, 0.05, and 0.05, respectively; and their allelic frequencies we are all nearly 0.30. Both the overall mean and residual variance we are set as 10.0, and the positions and effects of the six QTNs are shown in Tables S1–S3. The genotype and phenotype simulations were the same as those used by Wang et al. [25].

The first model (only six QTNs' additive effects) is: $\mathbf{y} = \mu + \sum_{i=1}^{6} x_i b_i + \varepsilon$, $\varepsilon \sim MVN_n(0, \sigma_e^2 \mathbf{I})$. The second model (six QTNs' additive effects plus polygenic effect) is: $\mathbf{y} = \mu + \sum_{i=1}^{6} x_i b_i + \mathbf{u} + \varepsilon$, $\mathbf{u} \sim MVN_n(0, \sigma_{pg}^2 \mathbf{K})$, $\varepsilon \sim MVN_n(0, \sigma_e^2 \mathbf{I})$, and $\mathbf{K}$ is the kinship matrix. Set $\sigma_{pg}^2 = 2$, thus $h_{pg}^2 = 0.092$. The third model (six QTNs' additive effects plus three other pairs of QTNs' epistatic effects) is: $\mathbf{y} = \mu + \sum_{i=1}^{6} x_i b_i + \sum_{j=1}^{3} (A_j \# B_j) b_{jj} + \varepsilon$, $\varepsilon \sim MVN_n(0, \sigma_e^2 \mathbf{I})$, # denotes Hadamard product (element-wise multiplication), three other pairs of epistatic QTNs (unrelated to the six true QTNs) are placed on 3063784bp (Chr.4) and 5227063bp (Chr.2), 5986135bp (Chr.2) and 2031781bp (Chr.3), and 2668059bp (Chr.3) and 11824678bp (Chr.1), respectively. Each pair of QTNs was set with $\sigma_{epi}^2 = 1.25$, thus $h_{epi}^2 = 0.05$.

### 2.3. Real Data and Preprocessing

We used four flowering-time related traits of *Arabidopsis thalina* datasets [24,26] for analysis. The four traits are days to flowering time under long days with vernalization (LDV), days to flowering time under short days with vernalization (SDV), days to flowering time under long days with two weeks vernalization (2W), and days to flowering time under long days with four weeks vernalization (4W), respectively. We removed the SNPs with minor allele frequency (MAF) less than 0.01, and 178376 SNPs remained ultimately. For phenotypes, we deleted the individuals with missing phenotype value, thus 168, 159, 152 and 119 individuals were reserved for each of the four traits LDV, SDV, 2W and 4W, respectively, and then a logarithmic transformation was performed to each phenotype value. Due to the strong population structure in *Arabidopsis thaliana*, we were obliged to eliminate the impact of population structure. We reorganized the SNP genotype data via the software PLINK (Version 1.09) [27] at first, then chose a suitable value for population number $q$ from 1 to 5 with the minimum cross-validation error, and calculated the population structure matrix $Q$ synchronously by using the software ADMIXTURE (Version 1.3) [28], and finally corrected the primary phenotype vector $\mathbf{y}$ by $Q_j, j = 1, 2, \ldots, q$, whose effects $\hat{\alpha}_j$ were estimated by least-square method. Therefore, the corrected phenotype vector is:

$$\mathbf{y}' = \mathbf{y} - \sum_{j=1}^{q} Q_j \hat{\alpha}_j = \mathbf{1}\mu + \sum_{i=1}^{p} X_i \beta_i + \varepsilon \tag{2}$$

### 2.4. Mutual Information

Mutual information proposed by Shannon [29] is based on the concept of entropy and has been widely used in feature selection [23]. Given two discrete random variables $X$ and $Y$, the mutual information of $X$ and $Y$ is defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

where $H(X)$ is the entropy of $X$ and $H(X,Y)$ is the joint entropy of $X$ and $Y$. They can be specified as:

$$H(X) = -\sum_{x} p(x) \cdot \log p(x) \tag{4}$$

$$H(X,Y) = -\sum_{x,y} p(x,y) \cdot \log p(x,y) \tag{5}$$

where $p(x) = P(X = x)$ is the marginal probability density function, and $p(x,y) = P(X = x, Y = y)$ is the joint probability density function. Mutual information can also be defined as:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

where $H(X|Y)$ is the conditional entropy of $X$ given $Y$. We calculated the mutual information by using the matlab package "MutualInfo" (Version 0.9) written by Peng et al. [23].

In fact, mutual information can be illustrated as the amount of information one random variable contained in another random variable. The larger the mutual information is, the stronger correlation between the two random variables is. In GWAS, we consider the phenotypic vector as one random variable, and the genotype vector of a SNP as another random variable. In this way, we can calculate the mutual information between each of the SNPs and phenotype.

### *2.5. SCAD*

SCAD is a penalized likelihood approach that enables to selecting variables and estimating coefficients simultaneously due to its Oracle properties [12]. The objective function $\xi$ is:

$$\xi_{\lambda,\gamma}(\beta) = argmin_{\beta} \left( \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} (X_{ij}\beta_j))^2 + \sum_{j=1}^{p} \rho_{\lambda,\gamma}(|\beta_j|) \right) \tag{7}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the regression coefficient vector to be estimated and, $\lambda$ and $\gamma$ are penalty and shrinkage parameter, respectively, both of them are greater than 0. The former term of Equation (7) is the loss function, and the latter term is the penalty function defined by:

$$\rho_{\lambda,\gamma}(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if} \quad |\beta_j| < \lambda, \\ \frac{-(|\beta_j|^2 - 2\gamma\lambda|\beta_j| + \lambda^2)}{2(\gamma-1)}, & \text{if} \quad \lambda \leq |\beta_j| < \gamma\lambda \quad \text{and} \quad \gamma > 2, \\ \frac{(\gamma+1)\lambda^2}{2}, & \text{if} \quad |\beta_j| \geq \gamma\lambda. \end{cases} \tag{8}$$

$\gamma = 3.7$ as suggested in the original study [12]. We performed SCAD by using the R package "ncvreg" from https://CRAN.R-project.org/package=ncvreg.

### *2.6. Likelihood Ratio Test*

Likelihood ratio test is to compare the maximum of likelihood function in null hypothesis $H_0$ and alternative hypothesis $H_1$, and further determine whether the hypothesis is effective. LOD (log of odds) score is a statistic criterion used in likelihood ratio test. The definition is:

$$LOD = log_{10}(\frac{l_0}{l_1}) = \frac{-2(L_0 - L_1)}{4.6052} \tag{9}$$

$l_0 = e^{L_0}$, $l_1 = e^{L_1}$, $L_0 = L(\theta_{-k})$ and $L_1 = L(\theta)$ are the natural logarithms of the likelihood functions for null hypothesis $H_0 : \beta_k = 0$ and alternative hypothesis $H_1$, respectively, $\theta_{-k} = \{\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_o\}$ and $\theta = \{\beta_1, \ldots, \beta_o\}$, and $o$ is the number of markers potentially associated with the trait. $LOD \geq 3$ was proposed to be the significant criterion in multi-locus GWAS [25], which is slightly stringent and equivalent to $P = Pr(\chi_1^2 > 3 \times 4.6052) \approx 0.0002$. Under $H_0$,

$LOD \times 4.6052$ follows a $\chi^2$ distribution with one degree of freedom. We set the significant criterion of MBLASSO, ISIS EM-BLASSO, and EM-BLASSO as $LOD \geq 3$, which is Bonferroni correction for GEMMA by referring the published study [30].

### *2.7. A Two-Stage Mutual Information Based Bayesian Lasso (MBLASSO) Method*

On the whole, this procedure is a two-stage strategy for multi-locus GWAS. In the first phase, we used a modified ISIS approach based on Pearson correlation and mutual information to obtain a subset of SNPs, the elements of which can be divided into two types, separately. As to Pearson correlation screening, Type I includes those SNPs with strong correlated to phenotype, and Type II consists of those SNPs weak correlated while indirectly correlated to phenotype with some SNPs from Type I. For mutual information screening, Types I and II are similar as those in Pearson correlation screening. The first phase of our method can be considered to select SNPs from two different measurements. In the second phase, we adopted EM-BLASSO [10] to estimate the effects and select the SNPs with nonzero effect ($\geq 10^{-5}$) to further likelihood ratio test procedure. We call this method MBLASSO. The flow chart is shown in Figure 1.

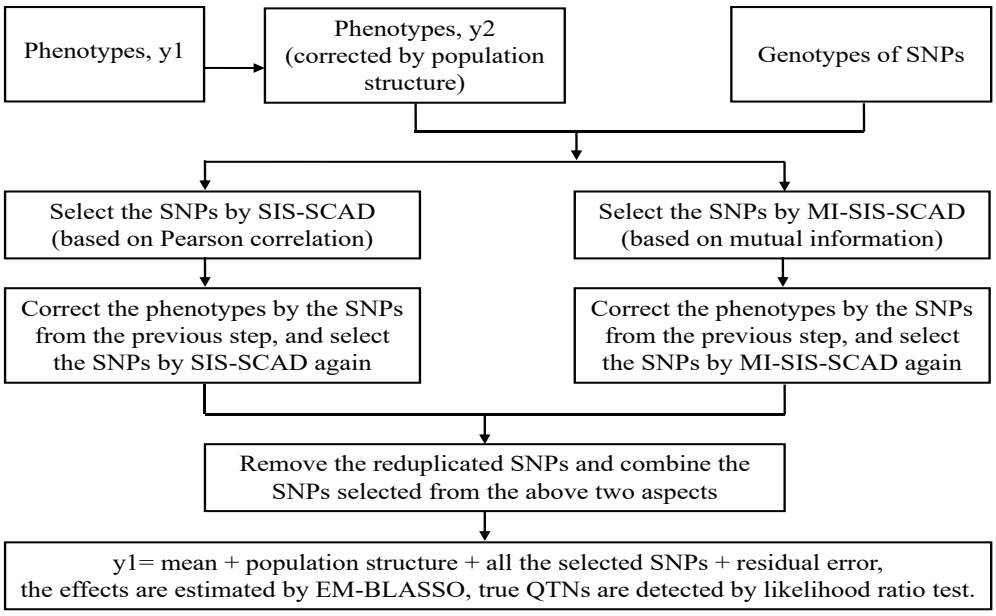


**Figure 1.** A flow chart of MBLASSO method.

More specifically, MBLASSO works as follows:

- Step 1: Correct the initial phenotype vector ($\mathbf{y}$) by the fixed effects, which indicate the population structure in our model.
- Step 2: Calculate the Pearson correlation of the $i$th SNP with the corrected phenotype ($\mathbf{y}^{'}$), that is,

$$\omega_i = \rho_{X_i,\mathbf{y}'} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x_i})(y_j^{'} - \bar{\mathbf{y}^{'}})}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x_i})^2} \cdot \sqrt{\sum_{j=1}^{n}(y_j^{'} - \bar{\mathbf{y}^{'}})^2}} \tag{10}$$

where $x_{ji}$ is the $i$th SNP genotype value of the $j$th individual, $y_j^{'}$ is the corrected phenotype value of the $j$th individual, $\bar{x_i}$ is the average of the genotype value of the $i$th SNP, $\bar{\mathbf{y}^{'}}$ is the mean of the

corrected phenotype value of all individuals, and $\omega = (\omega_1, \omega_2, \ldots, \omega_p)^T$ is a vector of Pearson correlation coefficients.

- Step 3: Sort the components of vector $\omega$ in descending order and define a subset:

$$\Omega = \{1 \leq i \leq p : |\omega_i| \text{ is among the } (n-1) \text{ largest of all}\} \tag{11}$$

where $n-1$ is one of the two sizes recommended by Fan and Lv [11], and it is more appropriate in our work. Suppose that there are $k_1$ SNPs corresponding to $\Omega$, $k_1 \geq n-1$, for the reason that more than one SNP may share a common Pearson correlation coefficient; the subset consisting of these SNPs is denoted as $\mathcal{A}_1 = \{X_{jm_1}, X_{jm_2}, \ldots, X_{jm_{k_1}}\}$, $m_1, m_2, \ldots, m_{k_1}$ are the orders of the $k_1$ selected SNPs in all the $p$ SNPs. Then implement SCAD to estimate the effects. Select the SNPs with nonzero effect to form another subset $\mathcal{A}_2 = \{X_{jl_1}, X_{jl_2}, \ldots, X_{jl_{k_2}}\} \subseteq \mathcal{A}_1$, $k_2 \leq k_1$, and $\{l_1, l_2, \ldots, l_{k_2}\} \subseteq \{m_1, m_2, \ldots, m_{k_1}\}$. The SNPs in $\mathcal{A}_2$ correspond to Type I in Pearson correlation screening. This Pearson correlation based SIS followed by SCAD is called SIS-SCAD [11].

- Step 4: Undertake ISIS-SCAD [11] to revive those non-negligible SNPs that are single uncorrelated but jointly correlated with phenotype, only one iteration is implemented here. Firstly correct the phenotype in Step 1 ($\mathbf{y}'$) by the $k_2$ SNPs selected by SIS-SCAD in Step 3, that is,

$$\mathbf{y}'' = \mathbf{y}' - \sum_{t=1}^{k_2} X_{l_t} \beta_{l_t} \tag{12}$$

where $\beta_{l_t}$ is estimated by SCAD, and then repeat SIS-SCAD to the rest of the $p - k_2$ SNPs, which results in another subset of $k_3$ SNPs, $\mathcal{A}_3 = \{X_{js_1}, X_{js_2}, \ldots, X_{js_{k_3}}\}$. The SNPs in $\mathcal{A}_3$ correspond to Type II in Pearson correlation screening. The union of the two disjoint subsets $\mathcal{A}_2$ and $\mathcal{A}_3$ is denoted as $\mathcal{A}$, $\mathcal{A} = \mathcal{A}_2 \bigcup \mathcal{A}_3$, whose size is $k$, $k = k_2 + k_3$.

- Step 5: Under the same conditions as in Step 2, calculate the mutual information of the $i$th SNP and the corrected phenotype ($\mathbf{y}'$) by

$$\psi_i = I(X_i, \mathbf{y}') = \sum_{j=1}^{n} p(x_{ji}, y_j') \cdot \log \frac{p(x_{ji}, y_j')}{p(x_{ji}) p(y_j')} \tag{13}$$

and $\psi = (\psi_1, \psi_2, \ldots, \psi_p)^T$ is a vector of mutual information for all of the $p$ SNPs with the corrected phenotype. $p(x_{ji}, y_j')$ is the joint probability, $p(x_{ji})$ and $p(y_j')$ are the marginal probabilities of $x_{ji}$ and $y_j'$, separately.

- Step 6: Similar to Step 3, sort the components of vector $\psi$ in descending order and define another subset:

$$\Psi = \{1 \leq i \leq p : \psi_i \text{ is among the } (n-1) \text{ largest of all}\} \tag{14}$$

Assume that $\tau_1$ SNPs corresponding to $\Psi$, $\tau_1 \geq n-1$, because more than one SNP may share a public mutual information with phenotype. The subset is $\mathcal{B}_1 = \{X_{jh_1}, X_{jh_2}, \ldots, X_{jh_{\tau_1}}\}$. Then use SCAD to estimate the effects of SNPs in $\mathcal{B}_1$ and select the SNPs with nonzero effect to constitute a new subset $\mathcal{B}_2 = \{X_{jr_1}, X_{jr_2}, \ldots, X_{jr_{\tau_2}}\} \subseteq \mathcal{B}_1$, $\tau_2 \leq \tau_1$, and $\{r_1, r_2, \ldots, r_{\tau_2}\} \subseteq \{h_1, h_2, \ldots, h_{\tau_1}\}$. The SNPs in $\mathcal{B}_2$ correspond to Type I in mutual information screening. We call this mutual information based SIS followed by SCAD as MI-SIS-SCAD.

- Step 7: Refering to Step 4, correct the phenotype in Step 1 ($\mathbf{y}'$) by $\tau_2$ SNPs selected by MI-SIS-SCAD, and repeat MI-SIS-SCAD once for to the remaining of the $p - \tau_2$ SNPs, which generates a subset of $\tau_3$ SNPs, $\mathcal{B}_3 = \{X_{jt_1}, X_{jt_2}, \ldots, X_{jt_{\tau_3}}\}$. The SNPs in $\mathcal{B}_3$ correspond to Type II in mutual information screening. The union of the disjoint subsets $\mathcal{B}_2$ and $\mathcal{B}_3$ is denoted as $\mathcal{B}$, $\mathcal{B} = \mathcal{B}_2 \bigcup \mathcal{B}_3$, the size of which is $\tau$, $\tau = \tau_2 + \tau_3$. We call this process as MI-ISIS-SCAD.

- Step 8: Gather the SNPs selected from Steps 4 and 7 and remove the reduplicated ones. Then obtain a new subset of SNPs, that is, $\mathcal{C} = \mathcal{A} \bigcup \mathcal{B}$, the size of which is $\nu = k + \tau$.
- Step 9: Use EM-BLASSO to estimate the effect of the $\nu$ SNPs from $\mathcal{C}$ and further eliminate the SNPs with zero effect, the source code for EM-BLASSO can be found at https://CRAN.R-project.org/package=mrMLM, where we can also download the program of ISIS EM-BLASSO. Note that the phenotype vector in this step refers to the original one ($\mathbf{y}$).
- Step 10: Apply the likelihood ratio test to identify the true QTNs, and set the significant criterion as $LOD \geq 3$.

## 3. Results

### 3.1. The Overlap Ratio between Pearson Correlation and Mutual Information Based Screening in MBLASSO

To illustrate the necessity of considering the correlation measured in mutual information between the SNPs and phenotype, we calculated the overlap ratio and average number of SNPs selected by Pearson correlation and mutual information in the first variable selection stage. The SNPs selected by Pearson correlation and mutual information can be divided into two types (Types I and Type II), respectively. We found that each type of screening obtains phenotype-associated SNPs without large overlapping (Table 1), which suggests that the SNPs from our MBLASSO method may have more associations with phenotype than ISIS EM-BLASSO.

**Table 1.** Screening results based on Pearson correlation and mutual information in MBLASSO under three simulation scenarios (each cell includes the overlap ratio and average number of SNPs after screening in the parentheses).

| Simulations | Pearson Correlation Screening | | | Mutual Information Screening | | |
|---|---|---|---|---|---|---|
| | Type I | Type II | Total | Type I | Type II | Total |
| 1 | 0.470 (15.8) | 0.086 (50.4) | 0.184 (66.2) | 0.417 (18.2) | 0.298 (15.5) | 0.356 (33.7) |
| 2 | 0.452 (16.6) | 0.091 (50.3) | 0.181 (66.9) | 0.398 (19.0) | 0.285 (17.5) | 0.334 (36.5) |
| 3 | 0.457 (14.6) | 0.090 (50.8) | 0.173 (65.4) | 0.383 (18.4) | 0.278 (17.4) | 0.323 (35.8) |

### 3.2. Statistical Power for QTN Detection

The power for the $i$th QTN is: $power_i = \ell_i/1000, i = 1, 2, 3, 4, 5, 6$, where $\ell_i$ is the frequency that $i$th hypothetical QTN is successfully detected among all 1000 repetitions. A detected SNP within 1kb of the candidate QTN is regarded as true QTN [6,25,30]. In three simulations, powers of the six QTNs in MBLASSO are highest, except the second QTN powers are lower than those of EM-BLASSO (Figure 2a–c and Tables S1–S3). The average powers of MBLASSO are 72.4, 71.4, and 65.2 (%) in three simulations, respectively. They are improved by 26.4, 28.9, and 26.1 (%) compared to GEMMA; 5.6, 4.3, and 3.8 (%) compared to EM-BLASSO; and 2.2, 2.0, and 3.0 (%) compared to ISIS EM-BLASSO. We supposed four QTNs (QTN2, QTN3, QTN5, and QTN6) with the same 5% heritability, but the detection powers of QTN5 are much lower than three other values for MBLASSO, ISIS EM-BLASSO, and EM-BLASSO (Figure 2a–c and Tables S1–S3). To measure the robustness of methods, we used the standard deviation of powers across the four QTNs, which was proposed by Ren et al. [30]. In Simulation 1, the standard deviations for MBLASSO, ISIS EM-BLASSO and EM-BLASSO are 8.14, 8.16 and 13.99, respectively, indicating the best stability of MBLASSO. The stability comparisons in Simulations 2 and 3 are the same as that in Simulation 1. Therefore, MBLASSO improves the power and has best stability in different scenarios. A Violin plot of average statistical powers for MBLASSO in three simulation scenarios is shown in Figure S1a.

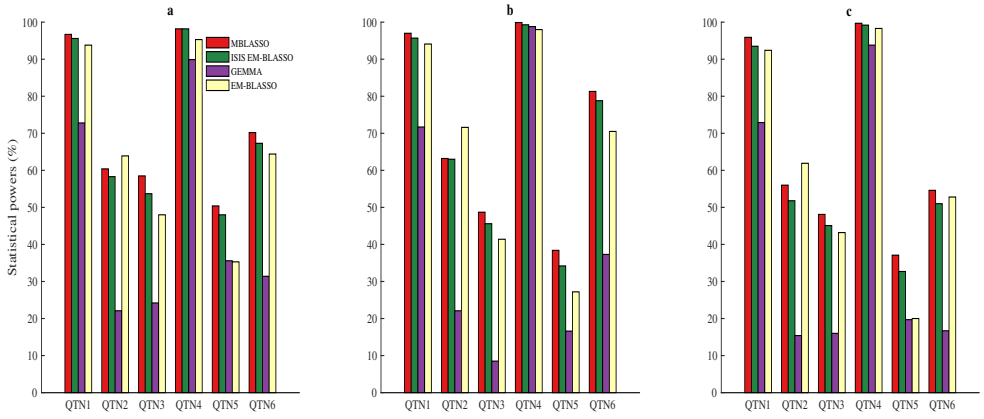


**Figure 2.** Statistical powers for the six simulated QTNs in three simulation scenarios. (**a**) only six QTNs' additive effects; (**b**) six QTNs' additive effects and polygenic background effect; and (**c**) six QTNs' additive effects and three other pairs of QTNs' epistatic effects.

### *3.3. Average Accuracy for QTN Effects*

Mean squared error (MSE) was used to quantify the bias of effect estimation. The MSE of the $i$th QTN is: $MSE_i = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\beta}_{ij} - \beta_i)^2$, $i = 1, 2, 3, 4, 5, 6$, where $\hat{\beta}_{ij}$ is the effect of the $i$th QTN in the $j$th repetition, and $\beta_i$ is the theoretical effect of the $i$th QTN. The smaller the MSE is, the better the accuracy of the algorithm is. We applied the average MSE of the six QTNs to totally measure the accuracy of different algorithms. They are 0.0610, 0.0812, 0.5467 and 0.0561 for MBLASSO, ISIS EM-BLASSO, GEMMA and EM-BLASSO, respectively, in Simulation 1, the similar case is shown in Simulation 2, and the average MSE for MBLASSO is the lowest in Simulation 3 (Figure 3a–c and Tables S1–S3), indicating the better estimation accuracy of MBLASSO on the whole. A violin plot of average MSEs for MBLASSO in three simulation scenarios is shown in Figure S1b.

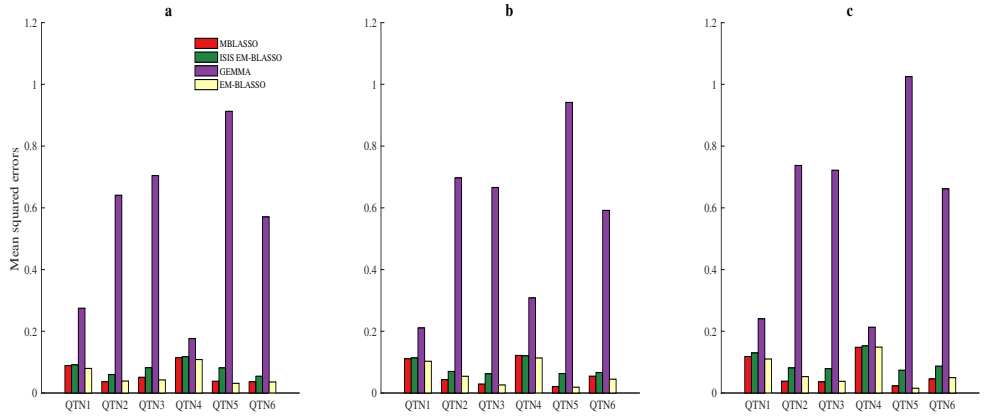


**Figure 3.** Average mean squared errors (MSEs) for the six simulated QTNs in three simulation scenarios. The description of (**a**–**c**) is the same as that in Figure 2.

### *3.4. Type 1 Error Ratio*

Type 1 error ratio, also known as false positive ratio, is an important problem that needs to be overcome in GWAS. In Simulation 1, they are 0.0302%, 0.0325%, 0.0325% and 0.0259% for MBLASSO, ISIS EM-BLASSO, GEMMA, and EM-BLASSO, respectively, and GEMMA has the lowest Type 1 error ratio in Simulations 2 and 3 (Figure 4). Note that all Type 1 error ratios are less than 0.05% (Figure 4 and Tables S1–S3), which indicates that all the four algorithms ensure the Type 1 error is at a very low level. A violin plot of Type 1 error ratios for MBLASSO in three simulation scenarios is shown in Figure S1c.

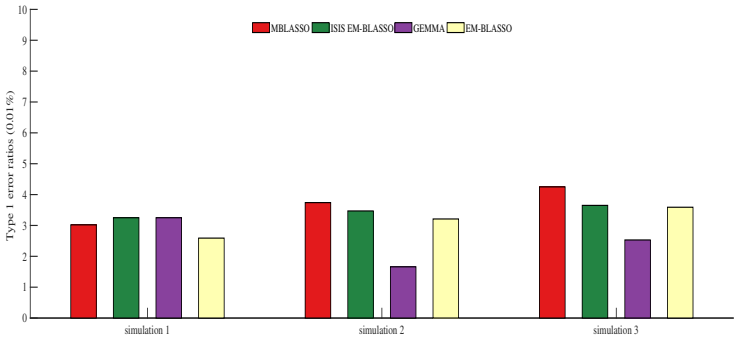


**Figure 4.** Type 1 error ratios (0.01%) in three simulation scenarios. The descriptions of Simulations 1–3 corresponding to (**a**–**c**) in Figure 2.

### *3.5. Computational Efficiency*

The computing time of MBLASSO is longer than that of ISIS EM-BLASSO, because it needs additional computation of mutual information between all the SNPs and phenotype, but it takes less time than EM-BLASSO. For example, in Simulation 1, MBLASSO finishes the analysis of 199 individuals with 10,000 SNPs for 1000 repetitions in 4.12 h, ISIS EM-BLASSO takes 2.90 h, GEMMA spends 2.20 h, and while EM-BLASSO needs 28.86 h for the same dataset (Table S1). The specific hours spent on the other two simulations are largely identical with only minor differences to those in Simulation 1 (Tables S2 and S3), and the operations of computation are on a computer of Intel Xeon E5-2640 CPU 2.40 GHz.

### *3.6. Arabidopsis Thaliana Dataset Analysis*

We analyzed four flowering-time related traits (LDV, SDV, 2W, and 4W) using by MBLASSO, ISIS EM-BLASSO, GEMMA and EM-BLASSO. Suppose that the candidate genes for the traits are in the proximity of 20 kb with the associated SNPs [6,25]; MBLASSO identifies 17, 18, 17 and 18 SNPs significant associated with each of the four traits LDV, SDV, 2W, and 4W, respectively. ISIS EM-BLASSO detects 14, 18, 19, and 16 remarkable associated SNPs; GEMMA identifies 3, 5, 1, and 2 significant SNPs; and EM-BLASSO tests 3, 0, 4, and 6 SNPs, respectively. A Venn diagram showings the overlap numbers of SNPs detected by the four algorithms in the four traits is presented in Figure S2.

To measure the model fitting degree of the detected SNPs, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were computed for each trait in four various methods, where a lower value indicates a better model fitting. We can explicitly see that MBLASSO shows the lowest AIC and BIC for the four traits (Table 2), thus it is the best algorithm in model fitting, followed by ISIS EM-BLASSO, EM-BLASSO, and GEMMA.

**Table 2.** Degree of model fitting (AIC, BIC) for SNPs identified in four flowering-time related traits for *Arabidopsis thaliana*.

| Traits | MBLASSO | | ISIS EM-BLASSO | | GEMMA | | EM-BLASSO | |
|---|---|---|---|---|---|---|---|---|
| | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| LDV | −360.543 | −307.436 | −318.966 | −275.230 | 1312.693 | 1322.065 | −113.638 | −104.266 |
| SDV | −169.269 | −114.028 | −140.485 | −85.245 | 1356.907 | 1372.251 | 149.095 | 149.095 |
| 2W | −103.363 | −51.957 | −65.172 | −7.718 | 584.000 | 587.024 | 148.247 | 160.342 |
| 4W | −124.109 | −74.084 | −98.993 | −54.527 | 1253.281 | 1258.839 | 22.893 | 39.568 |

Meanwhile, by referring to the latest GO annotation [31] for *Arabidopsis thalina* genes at www.arabidopsis.org, we extracted the known genes related to flowering-time traits and found 5, 4, 2, and 3 known genes closed to the detected SNPs with MBLASSO; 3, 2, 1, and 2 known genes with ISIS

EM-BLASSO; 0, 1, 0, and 1 known genes with GEMMA; and none of known genes could be identified by using EM-BLASSO for LDV, SDV, 2W and 4W, respectively (Table 3). These results suggest that the accuracy of associations retrieved by MBLASSO are the highest.

**Table 3.** The accuracy of detected associations in four flowering-time related traits for *Arabidopsis thaliana* (the number behind slash in each cell is the count of detected SNPs, and the number in front of slash is the count of known genes in GO annotation),

| Traits | MBLASSO | ISIS EM-BLASSO | GEMMA | EM-BLASSO |
|---|---|---|---|---|
| LDV | 5/17 | 3/14 | 0/3 | 0/3 |
| SDV | 4/18 | 2/18 | 1/5 | 0/0 |
| 2W | 2/17 | 1/19 | 0/1 | 0/4 |
| 4W | 3/18 | 2/16 | 1/2 | 0/6 |

In addition, totally 21 genes were only detected by MBLASSO, among which five genes (AT5G45830, AT5G45840, AT3G57230, AT5G15850, and AT5G04240) are in the 98 candidate genes [24], and AT5G45830 (alias: DOG1) is the gene with the highest frequency significant associated with flowering-time related phenotypes. Nearly all of the 23 flowering-time related phenotypes are associated with this gene [24]. Meanwhile, AT5G45840 (alias: MDIS1) is one of the Top 5 flowering-time related genes studied by researchers in *Arabidopsis thaliana* (www.arabidopsis.org). The detailed GWAS results are listed in Table S4.

About the computation speed, despite MBLASSO being is slower than ISIS EM-BLASSO and GEMMA, it is much faster than EM-BLASSO, for example, for the trait LDV, the time for MBLASSO is 2.31 min, ISIS EM-BLASSO requires 1.92 min, GEMMA takes 0.85 min and EM-BLASSO consumes to 183.6 min. We notice that the time costs of all the four flowering-time traits in MBLASSO, ISIS EM-BLASSO, and GEMMA are all less than 3 min (Table S5).

## 4. Discussion

MBLASSO is a GWAS method modified from ISIS EM-BLASSO, that is, iterative sure independence screening (ISIS) in the first stage of ISIS EM-BLASSO is replaced by a combination ISIS based on Pearson correlation and mutual information. We assume a subset of loci jointly affects the phenotype. In the first stage, we focus on selecting those SNPs that are likely to be highly associated. Considering some SNPs may have different correlations under various phenotypes, which are hard to measure only by Pearson correlation, so we adopt the mutual information to obtain the SNPs with potential correlation to phenotype. Meanwhile, since those SNPs individually irrelevant but jointly relevant to phenotype can be revived, this multi-objective screening process is a crucial component of our methodology to improve the statistical power. In the second stage, we apply the existing EM-BLASSO method [10], which is actually a single stage multi-locus GWAS strategy, to estimate the effects of selected SNPs and further filter out the SNPs with very small effect ($<10^{-5}$). Finally, we use likelihood ratio test to identify the true QTNs.

In fact, the method and criterion of hypothesis testing in different approaches may be different, e.g., the Wald test is applied in RMLM [25] and original EM-BLASSO [10], the significant level is $P = 0.01$ or 0.05, and a looser likelihood ratio test criterion $LOD \geq 2$ is employed in pLARmEB [32]. Since different significant criteria will lead to changes in results, for above three simulations, we listed the performances (average power, average MSE and Type 1 error ratio) of MBLASSO in three different significant criteria ($LOD = 3$, $LOD = 2$ and $P = 0.01$) in Table S6. We can see the average power increased with the decrease of $LOD$ value, but the Type 1 error ratio and average MSE also increased. This means that with the relaxation of significant criteria, high statistical power will be achieved, while false positives will be increased and estimation accuracy will be reduced. In addition, the performances at the significant criterion $P = 0.01$ in Wald test are between $LOD = 2$ and 3 in likelihood ratio test. GEMMA is a single-locus GWAS approach, and the

significant threshold for each test is determined by Bonferroni correction ($0.05/p$, $p$ is the number of SNPs). MBLASSO, ISIS EM-BLASSO, and EM-BLASSO are multi-locus approaches and do not require multiple test correction.

We conducted paired t-test (also used in [6,25,30]) for statistical power and MSE between MBLASSO and three other methods in three simulation scenarios (Table S7). We can see it has significant improvements compared with ISIS EM-BLASSO and GEMMA. For the traits SDV and 2 W in real *Arabidopsis thaliana* datasets, the numbers of significant SNPs identified by MBLASSO are not more than ISIS EM-BLASSO, but the degrees of model fitting are better (Table 2); and the number of known candidate genes adjacent to the detected SNPs is still larger (Table 3), this phenomenon indicates MBLASSO may be more effective to capture the inherent relationship between SNPs and phenotype. The traditional EM-BLASSO [10] and GEMMA perform well in terms of Type 1 error ratio in the three simulations, but their performances in *Arabidopsis thaliana* dataset are worse than expected, not only achieving the worse model fitting performance but also fewer of genes are detected. On the whole, our algorithm MBLASSO is slightly slower than ISIS EM-BLASSO and GEMMA, but it is more effective and accurate for both simulation and real datasets.

## 5. Conclusions

Our algorithm MBLASSO is a modified version of ISIS EM-BLASSO; it integrates Pearson correlation and mutual information to the feature screening stage, and it considers different types of correlation between the SNPs and phenotype. In three different simulation scenarios, MBLASSO improves the statistical power and retains the higher effect estimation accuracy when comparing with three other methods. Meanwhile, the GWAS results in four flowering-time related traits are superior in model fitting; the accuracy of detected associations are the highest; and 21 genes can only be detected by MBLASSO.

**Supplementary Materials:** The following are available online at http://www.mdpi.com/1099-4300/22/3/329/s1.

**Author Contributions:** Conceptualization, H.G. and Z.Y.; methodology, H.G.; Writing—Original draft preparation, H.G.; Writing—Review and editing, H.G., Z.Y., J.A., G.H., Y.M. and R.T.; Supervision, Z.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (Grant No. 11871061); Collaborative Research project for Overseas Scholars (including Hong Kong and Macau) of National Natural Science Foundation of China (Grant No. 61828203); Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT)(Grant No. IRT_15R58); Hunan Provincial Innovation Foundation for Postgraduate (Grant No. CX2018B375); and The project for Excellent Young and Middle-aged Science and Technology Innovation Team of Hubei Province (Grant No. T201731).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Yu, J.; Pressoir, G.; Briggs, W.H.; Bi, I.V.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **2006**, *38*, 203–208. [CrossRef]
2. Kang, H.M.; Zaitlen, N.A.; Wade, C.M.; Kirby, A.; Heckerman, D.; Daly, M.J.; Eskin, E. Efficient control of population structure in model organism association mapping. *Genetics* **2008**, *178*, 1709–1723. [CrossRef] [PubMed]
3. Zhang, Z.; Ersoz, E.; Lai, C.Q.; Todhunter, R.J.; Tiwari, H.K.; Gore, M.A.; Bradbury, P.J.; Yu, J.; Arnett, D.K.; Ordovas, J.M.; et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **2010**, *42*, 355–360. [CrossRef] [PubMed]
4. Lippert, C.; Listgarten, J.; Liu, Y.; Kadie, C.M.; Davidson, R.I.; Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **2011**, *8*, 833–835. [CrossRef]
5. Zhou, X.; Stephens, M. Genome-wide efficient mixed model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [CrossRef]

6. Tamba, C.L.; Ni, Y.L.; Zhang, Y.M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* **2017**, *13*, e1005357. [CrossRef]
7. Wu, T.T.; Chen, Y.F.; Hastie, T.; Sobel, E.; Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **2009**, *25*, 714–721. [CrossRef]
8. Cho, S.; Kim, H.; Oh, S.; Kim, K.; Taesung, P. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc.* **2009**, *3*, S25. [CrossRef]
9. Li, J.; Das, K.; Fu, G.; Li, R.; Wu, R. The Bayesian lasso for genome-wide association studies. *Bioinformatics* **2011**, *27*, 516–523. [CrossRef]
10. Xu, S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **2010**, *105*, 483–494. [CrossRef]
11. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 849–911. [CrossRef] [PubMed]
12. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
13. Zou, H. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
14. Li, G.; Peng, H.; Zhang, J.; Zhu, L. Robust rank correlation based screening. *Ann. Stat.* **2012**, *40*, 1846–1877. [CrossRef]
15. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef]
16. Li, R.; Liu, J.; Lou, L. Variable selection via partial correlation. *Statistica Sinica* **2017**, *27*, 983–996. [CrossRef]
17. Jiang, L.; Liu, J.; Zhu, X.; Ye, M.; Sun, L.; Lacaze, X.; Wu, R. 2HiGWAS: A unifying high-dimensional platform to infer the global genetic architecture of trait development. *Brief. Bioinform.* **2015**, *16*, 905–911. [CrossRef]
18. Cui, Y.; Zhang, F.; Zhou, Y. The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. *Front. Plant Sci.* **2018**, *9*, 1464. [CrossRef]
19. Liu, J.; Ye, M.; Zhu, S.; Jiang, L.; Sang, M.; Gan, J.; Wang, Q.; Huang, M.; Wu, R. Two-stage identification of SNP effects on dynamic poplar growth. *Plant J.* **2018**, *93*, 286–296. [CrossRef]
20. Fan, J.; Han, F.; Liu, H. Challenges of big data analysis. *Nat. Sci. Rev.* **2014**, *1*, 293–314. [CrossRef]
21. Jing, P.J.; Shen, H.B. MACOED: A multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics* **2015**, *31*, 634–641. [CrossRef] [PubMed]
22. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; Mcvean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
23. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
24. Atwell, S.; Huang, Y.S.; Vilhjalmsson, B.J.; Willems, G.; Horton, M.; Li, Y.; Meng, D. Genome-wide association study of 107 phenotypes in *Arabidopsis* thaliana inbred lines. *Nature* **2010**, *465*, 627–631. [CrossRef] [PubMed]
25. Wang, S.B.; Feng, J.Y.; Ren, W.L.; Huang, B.; Zhou, L.; Wen, Y.J.; Zhang, J.; Dunwell, J.M.; Xu, S.; Zhang, Y.M. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* **2016**, *6*, 19444. [CrossRef] [PubMed]
26. Togninalli, M.; Seren, Ü.; Freudenthal, J.A.; Monroe, J.G.; Meng, D.; Nordborg, M.; Weigel, D.; Borgwardt, K.; Korte, A.; Grimm, D.G. AraPheno and the AraGWAS Catalog 2020: A major database update including RNA-Seq and knockout mutation data for *Arabidopsis thaliana*. *Nucleic Acids Res.* **2019**, *48*, D1063–D1068. [CrossRef]
27. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; Bakker, P.I.W.D.; Daly, M.J. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
28. Alexander, D.H.; Novembre, J.; Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **2009**, *19*, 1655–1664. [CrossRef]
29. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

30. Ren, W.L.; Wen, Y.J.; Dunwell, J.M.; Zhang, Y.M. pKWmEB: Integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* **2018**, *120*, 208–218. [CrossRef]
31. Berardini, T.Z.; Mundodi, S.; Reiser, L.; Huala, E.; Garcia-Hernandez, M.; Zhang, P.; Mueller, L.A.; Yoon, J.; Doyle, A.; Lander, G.; et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol.* **2004**, *135*, 745–755. [CrossRef] [PubMed]
32. Zhang, J.; Feng, J.Y.; Ni, Y.L.; Wen, Y.J.; Niu, Y.; Tamba, C.L.; Yue, C.; Song, Q.; Zhang, Y.M. pLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* **2017**, *118*, 517–524. [CrossRef] [PubMed]