genes

MDPI

Article

Hybrid of Restricted and Penalized Maximum Likelihood Method for Efficient Genome-Wide Association Study

Wenlong Ren [1], Zhikai Liang [2], Shu He [1] and Jing Xiao [1,*]

[1] Department of Epidemiology and Medical Statistics, School of Public Health, Nantong University, Nantong 226019, China; wenlongren@ntu.edu.cn (W.R.); he_shu@ntu.edu.cn (S.H.)

[2] Plant and Microbial Biology Department, University of Minnesota, Saint Paul, MN 55108, USA; liang795@umn.edu

* Correspondence: jxiaont@ntu.edu.cn

Received: 8 October 2020; Accepted: 27 October 2020; Published: 29 October 2020

Abstract: In genome-wide association studies, linear mixed models (LMMs) have been widely used to explore the molecular mechanism of complex traits. However, typical association approaches suffer from several important drawbacks: estimation of variance components in LMMs with large scale individuals is computationally slow; single-locus model is unsatisfactory to handle complex confounding and causes loss of statistical power. To address these issues, we propose an efficient two-stage method based on hybrid of restricted and penalized maximum likelihood, named HRePML. Firstly, we performed restricted maximum likelihood (REML) on single-locus LMM to remove unrelated markers, where spectral decomposition on covariance matrix was used to fast estimate variance components. Secondly, we carried out penalized maximum likelihood (PML) on multi-locus LMM for markers with reasonably large effects. To validate the effectiveness of HRePML, we conducted a series of simulation studies and real data analyses. As a result, our method always had the highest average statistical power compared with multi-locus mixed-model (MLMM), fixed and random model circulating probability unification (FarmCPU), and genome-wide efficient mixed model association (GEMMA). More importantly, HRePML can provide higher accuracy estimation of marker effects. HRePML also identifies 41 previous reported genes associated with development traits in *Arabidopsis*, which is more than was detected by the other methods.

Keywords: restricted maximum likelihood; penalized; computational efficiency; linear mixed model; GWAS

1. Introduction

Genome-wide association studies (GWAS) can advance our understanding of molecular mechanism of complex traits [1–4]. Testing each SNP (single nucleotide polymorphism) one time is the most popular method, which is flexible to perform on all kinds of models. However, each SNP requires multiple testing adjustment, which will result in strict *p*-values. One strategy to solve this problem is to use more information beyond the *p*-value. For example, Xu, et al. [5] proposed a model-based clustering method that borrowed information across SNPs and increased the signal strength by properly clustering SNPs. Lee and Lee [6] presented a web application for the network-based *Arabidopsis* genome-wide association boosting, which can identify weak association signals by integrating co-functional gene network information. Apart from this, the linear mixed model (LMM) has become a widely used methodology due to its capability in controlling for population stratification and the inclusion of related individuals [7]. However, the implementation of LMM requires estimating the variance component of each random effect, leading to increased computational burden. The restricted maximum likelihood

(REML) is the widely used method for the estimation of variance components. Conventional REML algorithms are impractical to handle large-scale genomic datasets with thousands of individuals and millions of SNPs. There are two main reasons limiting REML application: on one hand, it is hard to obtain closed-form solutions of REML or posterior estimations of variance components. On the other hand, the inversion of the covariance matrix is required to perform on each computation of likelihood, an operation that is proportional to the cube of individual number. As a result, improving the computational efficiency of REML for estimating variance components has become one of the research hotspots [8–12].

With regards to this, several approaches based on sparse matrix operations have been developed to improve the calculating speed [13,14]. Lippert, et al. [8] performed spectral decomposition on the covariance matrix, converting matrix inversion to diagonal reciprocal operation. This strategy not only greatly improves the computational efficiency but, also, takes advantage of genetic relatedness matrix to adjust the correlation. Similarly, Zhou and Stephens [9] implemented their method in a genome-wide efficient mixed model association (GEMMA), which only required a small amount of matrix vector multiplications to obtain variance components. A different idea was proposed by Loh et al. [10], which used preconditioned conjugate gradients and stochastic trace estimators to avoid all cubic operations. This is an asymptotic method via linear system transformations, particularly suitable for Bio Bank large-scale individuals. In addition to the above two popular ideas, some specialized methods have been developed to solve variance component estimations efficiently. Lourenco et al. [15] proposed a robust derivative-free restricted-maximum likelihood framework (DF-REML), which can tackle normality violations, as well as other model misspecifications. Cesarani et al. [16] investigated bias in variance components under different genotyping strategies, showing that single-step genomic restricted maximum likelihood (ssGREML) is more robust compared to GREML. Ganjgahi et al. [4] proposed a weighted least squares REML (WLS-REML) using a noniterative one-step random effect estimator to decrease the computational cost. Border and Becker [12] developed stochastic *Lanczos* derivative-free REML and *Lanczos* first-order Monte Carlo REML to further improve the computing speed. However, these existing methods for REML variance components estimation are mainly aimed at single-locus LMM, which is not effective enough to handle a complex genetic structure.

Several classical multivariate selection methods have good performances in association analyses when the number of SNPs is not far more than that of individuals, including Lasso and its derivatives [17–19], penalized maximum likelihood (PML) [20–23], and Bayesian methods [24,25]. However, most of these methods are not available to analyze large-scale genomic data due to ultra-high dimensional variables. To address this issue effectively, some improved multi-locus GWAS methods were proposed. For example, multi-locus mixed-model (MLMM) [26] adopts stepwise mixed-model regression with forward inclusion and backward elimination using a Bayesian approach and performs well when the structure is complex, fixed and random model circulating probability unification (FarmCPU) [27] incorporates multiple markers simultaneously as covariates in a stepwise LMM to partially remove the confounding between testing markers and kinship, iterative nonlocal prior-based selection (GWASinlps) [28] considers an iterative structured screen-and-select strategy and nonlocal priors within it and provides an efficient and parsimonious variable selection for continuous phenotypes, a machine-learning method combines a random forest-based technique with the modeling of linkage disequilibrium through latent variables [29] and accelerates the computing speed for multi-locus GWAS, a gene set analysis with Generalized Berk-Jones (GBJ) statistic [30] introduces a permutation-free parametric framework, which can increase the power by incorporating information from multiple signals in the same gene, the SNP set GWAS approach RAINBOW [31] achieves faster computation by using linear kernel for constructing the Gram matrix of the SNP set of interest, and the multi-locus random SNP effect mixed linear model (mrMLM) [32] uses the Wald test based on a random SNP effect linear mixed model to reduce dimensionality; then, all the selected markers are placed into an empirical Bayes [33] multi-locus model, showing the advantage in controlling a complex population structure. A limitation of Bayesian method is that Markov Chain Monte Carlo (MCMC) sampling comes at

the cost of intensive computation, or the posterior distribution of fitness is not easy to calculate [34]. Penalized maximum likelihood is similar to the Bayesian method involving the posterior distribution of parameters; the difference is that PML adopts a fast approach to obtain the maximum posteriori estimates of fitness via numerical optimization. Therefore, PML provides an efficient way to perform multivariate selection.

In this study, we developed an efficiently hybrid method HRePML to perform GWAS, which takes full advantage of REML and PML. Under the linear mixed model framework, we firstly conducted single-locus marker scanning using REML and then carried out multi-locus marker selection based on the reduced subset, taking genetic relatedness and population stratification into account. We used pure C++ language to implement HRePML and overcome one key issue in the programming limited memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [35]. In order to validate the effectiveness of HRePML, we conducted a series of simulation studies and real data analyses and compared it with three methods: MLMM [26], FarmCPU [27], and GEMMA [9].

2. Materials and Methods

2.1. The Arabidopsis thaliana Dataset

A publicly available dataset of *Arabidopsis thaliana* [36] is used to conduct a simulation study and real data analysis. This dataset contains 216,130 markers and 199 individuals. There are four development related traits to be analyzed, which are the number of days between the appearance of the first flower and the senescence of the last flower (FT duration GH), number of days between germination and plant complete senescence (LC duration GH), number of days between germination and senescence of the last flower (LFS GH), and number of days between last flower senescence and complete plant senescence (MT GH), respectively.

2.2. Restricted Maximum Likelihood (REML) Method in Single-Locus Screening Stage

2.2.1. Single-Locus Linear Mixed Model

A standard linear mixed model for association mapping can be expressed as

$$\boldsymbol{y} = \boldsymbol{Fb} + \boldsymbol{X}\beta + \boldsymbol{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ denotes the $n \times 1$ observed phenotypic vector of n individuals, $\boldsymbol{F}$ is an $n \times c$ fixed effect design matrix, including $\mathbf{1}$'s column vector, $\boldsymbol{b}$ is a $c \times 1$ vector of their corresponding effect sizes, $\boldsymbol{X}$ denotes the $n \times 1$ marker genotype vector of focal variant, β is the random effect of one focal marker with normal distribution $\beta \sim N(0, \sigma_g^2)$, the variance σ_g^2 is changed with different markers, and $\boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{\Sigma_u})$ is an $n \times 1$ random vector and is typically used to account for polygenic effects or confounding factors; here, σ_u^2 is the variance, and $\boldsymbol{\Sigma_u}$ is an $n \times n$ covariance structure defined as $\boldsymbol{\Sigma_u} = \boldsymbol{G} \cdot \boldsymbol{G}^T/m$; $\boldsymbol{G}$ is an $n \times m$ genotype matrix, and m is the number of markers; $\boldsymbol{\varepsilon} \sim N(0, \sigma_n^2 \boldsymbol{I_n})$ denotes an $n \times 1$ independently and identically distributed (i.i.d.) residual vector, and σ_n^2 is the residual variance. The covariance of $\boldsymbol{y}$ can be denoted as

$$\begin{aligned} Cov(\boldsymbol{y}) =& \boldsymbol{X}\boldsymbol{X}^T\sigma_g^2 + \boldsymbol{\Sigma_u}\sigma_u^2 + \boldsymbol{I_n}\sigma_n^2 \\ =& \left[\boldsymbol{X}\boldsymbol{X}^T\omega_g + (\boldsymbol{\Sigma_u}\omega_u + \boldsymbol{I_n})\right]\sigma_n^2 \\ =& \left(\boldsymbol{X}\boldsymbol{X}^T\omega_g + \boldsymbol{P}\right)\sigma_n^2 \end{aligned} \tag{2}$$

where $\omega_g = \sigma_g^2/\sigma_n^2$, $\omega_u = \sigma_u^2/\sigma_n^2$, and $\boldsymbol{P} = \boldsymbol{\Sigma_u}\omega_u + \boldsymbol{I_n}$. The estimate of ω_u can be obtained under the null model, defined as $\hat{\omega}_u$, which only needs to be computed once. Using spectral decomposition, $\boldsymbol{\Sigma_u}$ can be expressed as $\boldsymbol{\Sigma_u} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, where $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$ is a diagonal matrix of eigenvalues, and $\boldsymbol{Q}$ is an $n \times n$ eigenvector matrix corresponding to these eigenvalues [32,37,38].

2.2.2. Equation Transformation and Update Covariance

Transform Equation (1) by left-multiplying $\boldsymbol{Q}^T$ [32] and generate the following model

$$\boldsymbol{y}_Q = \boldsymbol{F}_Q\boldsymbol{b} + \boldsymbol{X}_Q\beta + \boldsymbol{Q}^T\boldsymbol{u} + \boldsymbol{Q}^T\boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{y}_Q = \boldsymbol{Q}^T\boldsymbol{y}$, $\boldsymbol{F}_Q = \boldsymbol{Q}^T\boldsymbol{F}$, and $\boldsymbol{X}_Q = \boldsymbol{Q}^T\boldsymbol{X}$. After transformation, the covariance of y becomes

$$\begin{aligned} Cov(\boldsymbol{y}_Q) =& \boldsymbol{X}_Q\boldsymbol{X}_Q^T\sigma_g^2 + \boldsymbol{Q}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\sigma_u^2\boldsymbol{Q} + \boldsymbol{Q}^T\boldsymbol{I}_{\boldsymbol{n}}\sigma_n^2\boldsymbol{Q} \\ =& \boldsymbol{X}_Q\boldsymbol{X}_Q^T\sigma_g^2 + \boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{Q}\sigma_u^2 + \boldsymbol{Q}^T\boldsymbol{I}_{\boldsymbol{n}}\boldsymbol{Q}\sigma_n^2 \\ =& \boldsymbol{X}_Q\boldsymbol{X}_Q^T\sigma_g^2 + \boldsymbol{\Lambda}\sigma_u^2 + \boldsymbol{I}_{\boldsymbol{n}}\sigma_n^2 \\ =& \left(\boldsymbol{X}_Q\boldsymbol{X}_Q^T\omega_g + \boldsymbol{\Lambda}\hat{\omega}_u + \boldsymbol{I}_{\boldsymbol{n}}\right)\sigma_n^2 \end{aligned} \tag{4}$$

Let $\boldsymbol{V}_0 = \boldsymbol{\Lambda}\hat{\omega}_u + \boldsymbol{I}_{\boldsymbol{n}}$ and $\boldsymbol{V} = \boldsymbol{X}_Q\boldsymbol{X}_Q^T\omega_g + \boldsymbol{V}_0$; clearly, diagonal matrix $\boldsymbol{V}_0$ is estimated. To determine whether a marker has an effect, hypothesis testing needs to be conducted. To estimate the marker effect β, its variance ratio ω_g needs to be obtained firstly, so that the estimation of each ω_g is the most interesting issue for each corresponding marker.

2.2.3. Optimal Solution via Efficient REML

In order to get the estimation $\hat{\omega}_g$, optimize the following restricted log-likelihood function with respect to ω_g,

$$\begin{aligned} L_r(\omega_g) =& -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log\left|\boldsymbol{F}_Q^T\boldsymbol{V}^{-1}\boldsymbol{F}_Q\right| - \frac{n-c}{2}\left(\boldsymbol{y}_Q - \boldsymbol{F}_Q\hat{\boldsymbol{b}}\right)^T\boldsymbol{V}^{-1}\left(\boldsymbol{y}_Q - \boldsymbol{F}_Q\hat{\boldsymbol{b}}\right) \\ =& -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log\left|\boldsymbol{F}_Q^T\boldsymbol{V}^{-1}\boldsymbol{F}_Q\right| - \frac{n-c}{2}\boldsymbol{y}_Q^T\boldsymbol{S}\boldsymbol{y}_Q \end{aligned} \tag{5}$$

where

$$\hat{\boldsymbol{b}} = \left(\boldsymbol{F}_Q^T\boldsymbol{V}^{-1}\boldsymbol{F}_Q\right)^{-1}\boldsymbol{F}_Q^T\boldsymbol{V}^{-1}\boldsymbol{y}_Q \tag{6}$$

$$\boldsymbol{S} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{F}_Q\left(\boldsymbol{F}_Q^T\boldsymbol{V}^{-1}\boldsymbol{F}_Q\right)^{-1}\boldsymbol{F}_Q^T\boldsymbol{V}^{-1} \tag{7}$$

Limited-memory BFGS (L-BFGS) [35,39] is an optimized algorithm of quasi-Newton methods for efficient solution. The libLBFGS is a user-friendly C library implementation of the L-BFGS method written by Nocedal [40]. This library requires that the objective function $L_r(\omega_g)$ and its gradient $\partial L_r(\omega_g)/\partial\omega_g$ are computable. However, the gradient function cannot be obtained directly by closed form. Fortunately, the well-known C++ Boost library [41] provides a finite difference method for solving the gradient, which is located in a boost/math/differentiation/finite_difference.hpp file. After $\hat{\omega}_g$ is estimated, $\hat{\beta}$ and $\hat{\sigma}_n^2$ can be easily obtained for restricted log-likelihood functions, which are as follows, respectively,

$$\hat{\beta} = \omega_g\boldsymbol{X}_Q^T\boldsymbol{S}\boldsymbol{y}_Q \tag{8}$$

$$\hat{\sigma}_n^2 = \frac{1}{n-c}\boldsymbol{y}_Q\boldsymbol{S}\boldsymbol{y}_Q \tag{9}$$

so that the variance of $\hat{\beta}$ denotes

$$\mathrm{var}(\hat{\beta}) = \omega_g\hat{\sigma}_n^2 - \omega_g\boldsymbol{X}_Q^T\boldsymbol{V}^{-1}\boldsymbol{X}_Q\omega_g\hat{\sigma}_n^2 \tag{10}$$

The Wald test is used to conduct a hypothesis test on each marker effect β—that is, $H_0 : \beta = 0$, $H_1 : \beta \neq 0$. The Wald test is

$$W = \frac{\hat{\beta}^2}{\mathrm{var}(\hat{\beta})} \tag{11}$$

W follows the chi-square distribution with 1 degree of freedom, and the p-value corresponding to W can be computed by the C++ Boost library, which is located in a boost/math/distributions/chi_squared.hpp file. In the single-marker screening stage, a relatively loose and flexible P cutoff is adopted.

2.3. Penalized Maximum Likelihood (PML) Method in Multi-Locus Screening Stage

2.3.1. Multi-Locus Linear Mixed Model and Penalized Likelihood Function

All the markers passing the REML step are placed into a multi-locus model [20] as

$$\boldsymbol{y} = \boldsymbol{Fb} + \sum_{i=1}^{t} \boldsymbol{x_i}\beta_i + \boldsymbol{\varepsilon} \tag{12}$$

where $\boldsymbol{y}$, $\boldsymbol{F}$, $\boldsymbol{b}$, and $\boldsymbol{\varepsilon}$ are the same definitions as Equation (1). $\boldsymbol{x_i}$ is the ith $n \times 1$ genotypic vector, and β_i is the fixed effect of corresponding marker. t denotes the total number of selective markers from the REML step.

Penalized maximum likelihood (PML) is a fast approach to use numerical optimization to obtain the maximum posteriori estimates when the penalty function is a probability density on the parameters [20,22,42]. Let $\boldsymbol{\theta} = \left\{\boldsymbol{b}, \beta_1, \cdots, \beta_t, \sigma_n^2\right\}$ is the interesting vector of the parameters. The log-likelihood function denotes

$$L(\boldsymbol{\theta}) = \log \varphi\left(\boldsymbol{y}; \boldsymbol{Fb} + \sum_{i=1}^{t} \boldsymbol{x_i}\beta_i, \boldsymbol{I}\sigma_n^2\right) \tag{13}$$

where $\varphi\left(\boldsymbol{y}; \boldsymbol{Fb} + \sum_{i=1}^{t} \boldsymbol{x_i}\beta_i, \boldsymbol{I}\sigma_n^2\right)$ is the normal density with the mean $\boldsymbol{Fb} + \sum_{i=1}^{t} \boldsymbol{x_i}\beta_i$ and covariance $\boldsymbol{I}\sigma_n^2$. A factor is introduced to penalize on the marker effects β_i

$$p(\beta_i) = \varphi\left(\beta_i; \mu_i, \sigma_i^2\right) \qquad i = 1, \cdots, t \tag{14}$$

where $\varphi\left(\beta_i; \mu_i, \sigma_i^2\right)$ is a normal prior with a mean μ_i and variance σ_i^2. Then, μ_i is also assigned a normal prior distribution

$$p(\mu_i) = \varphi\left(\mu_i; 0, \sigma_i^2/\tau\right) \qquad i = 1, \cdots, t; \tau > 0 \tag{15}$$

Let $\boldsymbol{\delta} = \left\{\mu_1, \cdots, \mu_t, \sigma_1^2, \cdots, \sigma_t^2\right\}$ be the hyperparameters that can be estimated along with the interested parameters at the same time. The logarithm of the penalized prior distribution is

$$P(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{t}\left[\log \varphi\left(\beta_i; \mu_i, \sigma_i^2\right) + \log \varphi\left(\mu_i; 0, \sigma_i^2/\tau\right)\right] \tag{16}$$

Then, the logarithm of penalized likelihood function can be defined as

$$L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = L(\boldsymbol{\theta}) + P(\boldsymbol{\theta}, \boldsymbol{\delta}) \tag{17}$$

2.3.2. Iterative Method for Parameter Estimation

The parameter vector $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are estimated by the penalized maximum likelihood simultaneously, and the solution of PML needs an iterative method to implement. For the interested parameter vector $\boldsymbol{\theta}$, find the first-order partial derivatives of the elements in $\boldsymbol{\theta}$ and then make them equal to 0.

Via $\frac{\partial}{\partial \boldsymbol{b}} L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$, $\frac{\partial}{\partial \beta_i} L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$, and $\frac{\partial}{\partial \sigma_n^2} L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$ and $i = 1, \cdots, t$, it can obtain a closed-form expression on $\boldsymbol{b}$, β_i, and σ_i^2, respectively.

$$\boldsymbol{b} = \left(\boldsymbol{F}^T \boldsymbol{F}\right)^{-1} \boldsymbol{F}^T \left(\boldsymbol{y} - \sum_{i=1}^{t} \boldsymbol{x_i} \beta_i\right) \tag{18}$$

$$\beta_i = \left[\boldsymbol{x_i}^T \boldsymbol{x_i} + \sigma_n^2 / \sigma_i^2\right]^{-1} \left[\boldsymbol{x_i}^T \left(\boldsymbol{y} - \boldsymbol{Fb} - \sum_{\substack{k=1 \\ k \neq i}}^{t} \boldsymbol{x_k} \beta_k\right) + \mu_i \sigma_n^2 / \sigma_i^2\right] \tag{19}$$

$$\sigma_n^2 = \frac{1}{n} \left(\boldsymbol{y} - \boldsymbol{Fb} - \sum_{i=1}^{t} \boldsymbol{x_i} \beta_i\right)^T \left(\boldsymbol{y} - \boldsymbol{Fb} - \sum_{i=1}^{t} \boldsymbol{x_i} \beta_i\right) \tag{20}$$

For the nuisance parameter vectors, $\boldsymbol{\delta}$, μ_i, and σ_i^2 are acquired in the same way. Via $\frac{\partial}{\partial \mu_i} L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$ and $\frac{\partial}{\partial \sigma_i^2} L_p(\boldsymbol{\theta}, \boldsymbol{\delta}) = 0$, $\mu_i = \beta_i / (\tau + 1)$ and $\sigma_i^2 = \frac{1}{2}\left[(\beta_i - \mu_i)^2 + \tau \mu_i^2\right]$ can be obtained.

Set the initial values for $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, and τ and update the values of $\boldsymbol{b}$, β_i, σ_n^2, μ_i, and σ_i^2 until convergence. In order to reach convergence quickly, a criterion according to the experience needs to be considered. After the parameter estimation, $\left|\beta_i\right| > 10^{-4}$ are selected to further conduct a likelihood-ratio test. The logarithm of the odds (LOD) score is used to determine the final identification.

2.4. Design of Simulation Experiments

Four simulation experiments were designed to validate our new method HRePML. In the first simulation study, our goal was to explore the new method's performance on the statistical power, mean square error (MSE), and running time. We generated a set of genotype data consisting of 500 individuals and 10,000 markers, which was based on the *Arabidopsis thaliana* dataset [36]. Eight quantitative trait nucleotides (QTNs) were simulated with heritability of 0.01, 0.03, 0.03, 0.05, 0.08, 0.01, 0.05, and 0.05, respectively. Their positions and true effects are described in Table 1. The total genetic heritability is $h_T^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2) = 0.01 \times 2 + 0.03 \times 2 + 0.05 \times 3 + 0.08 = 0.31$, and the residual variance is $\sigma_e^2 = 10.0$. Then, the total genetic variance σ_G^2 can be obtained, as well as the genetic variance of each QTN $\sigma_{gi}^2 (i = 1, \cdots, 8)$. The population mean is set to 10.0. The phenotype is generated by the model $y = \mu + \sum_{i=1}^{8} \boldsymbol{x_i} \beta_i + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(0, 10.0 \times \boldsymbol{I_n})$. The experiment is repeated 1000 times. The statistical power for each QTN is defined as the percentage of the number of detected QTN to the number of repetitions. We used the logarithm of the odds (LOD = 3.0) as the criterion for detecting QTN [32,43,44]. Mean squared error was calculated as $MSE_k = \frac{1}{S} \sum_{i=1}^{S} \left(\hat{\beta}_{ki} - \beta_k\right)^2$, where $k = 1, \cdots, 8$, S was the number of detected QTN k among 1000 repetitions, $\hat{\beta}_{ki}$ was the effect estimation of QTN k in the ith repeat, and β_k was the kth QTN's true effect.

In the second simulation study, we aimed at exploring the influence of polygenic background on HRePML. The polygenic effects were introduced with multivariate normal distribution $MVN(0, \sigma_{pg}^2 \boldsymbol{K})$, where the polygenic variance σ_{pg}^2 was set to 2.0, genetic relatedness matrix $\boldsymbol{K}$ was calculated as $\boldsymbol{G}^T \boldsymbol{G} / M$, $\boldsymbol{G}$ was the genotype matrix, and M was the number of markers. Other parameters were set to the same as the first simulation study. The position and true effect of each QTN are listed in Table S1. Based on the model $y = \mu + \sum_{i=1}^{8} \boldsymbol{x_i} \beta_i + \boldsymbol{u} + \boldsymbol{\varepsilon}$, where $\boldsymbol{u} \sim MVN(0, 2.0 \times \boldsymbol{K})$, the phenotypes are simulated.

Table 1. Comparison of the statistical power and mean squared errors (MSE) for each quantitative trait nucleotide (QTN) among the hybrid of restricted and penalized maximum likelihood (HRePML), multi-locus mixed-model (MLMM), fixed and random model circulating probability unification (FarmCPU), and genome-wide efficient mixed model association (GEMMA) methods in the first simulation study *.

QTN	Chr.	Position(bp)	R^2	Effect	Power (%)				Mean Squared Errors (MSE)			
					HRePML	MLMM	FarmCPU	GEMMA	HRePML	MLMM	FarmCPU	GEMMA
1	1	404108	0.01	0.4328	9.9	2.4	1.6	0.0	0.0509	0.1334	0.1224	na #
2	1	636788	0.03	0.7497	45.1	39.9	53.7	0.2	0.0193	0.0440	0.0241	0.3112
3	3	507976	0.03	0.7497	66.7	40.1	13.9	8.1	0.1443	0.0992	0.2756	0.1597
4	3	931437	0.05	0.9679	89.5	69.5	58.8	55.3	0.0321	0.0276	0.0434	0.0770
5	4	75898	0.08	1.2243	100.0	99.8	100.0	97.5	0.0407	0.0375	0.0527	0.0283
6	4	461978	0.01	0.4328	12.7	5.0	8.9	0.7	0.2488	0.3808	0.3429	0.5502
7	4	607026	0.05	0.9679	69.6	80.9	98.5	73.8	0.0421	0.0988	0.0367	0.1544
8	5	282008	0.05	0.9679	89.6	87.6	90.3	55.1	0.0397	0.0334	0.0345	0.0725

* In the first simulation study, the dataset consists of 500 individuals and 10,000 single nucleotide polymorphism (SNP) markers with 1000 replicates. Eight true QTNs are set in each replicate. Then, this dataset can be regarded as having 10,000,000 SNPs and 8000 true QTNs in total. # "na" represents not available.

In the third simulation study, our goal was to investigate the influence of the sample size on the running time and statistical power. The sample size was set to 500, 1000, 2000, and 4000, respectively. Meanwhile, the number of markers was fixed at 10,000. In the fourth simulation study, our aim was to investigate the impact on running time as the number of markers increased. The number of markers was set to 10,000, 50,000, 100,000, and 200,000, respectively. At the same time, the sample size was fixed at 500. In these two simulation studies, the repeat times were set to 100. The position and heritability of each QTN were set to the same as those of first simulation study. Their parameters are listed in Table S3.

3. Results

3.1. Statistical Properties

We compared the statistical properties of the new HRePML method with those of the multi-locus mixed-model (MLMM) [26], fixed and random model circulating probability unification (FarmCPU) [27], and genome-wide efficient mixed model association (GEMMA) [9] methods. Here, the statistical properties mainly included statistical power and mean squared error (MSE). In the first simulation study, the dataset consisted of 500 individuals and 10,000 SNP markers with 1000 replicates. Eight true QTNs were set in each replicate. Then, this dataset was regarded as having 10,000,000 SNPs and 8000 true QTNs in total. The average power of the four methods were 60.39%, 53.15%, 53.21%, and 36.34%, respectively. HRePML obtained the highest statistical power, which was at least about 7% higher than the other three methods. In particular, HRePML performed well on QTNs with lower heritability, such as QTN 1, 3, 4, and 6 (Tables 1 and 2 and Figure 1A). The mean squared error was used to measure the accuracy of the QTN effect estimates, and smaller MSE represented better accuracy. The average MSE of the above four methods were 0.0772, 0.1068, 0.1165, and 0.1933, respectively, demonstrating that the average MSE of HRePML was the minimum (Tables 1 and 2 and Figure 2A).

Table 2. Comparison of average statistical power, average mean squared errors (MSE), and running time among the HRePML, MLMM, FarmCPU, and GEMMA methods in the first simulation study *.

Statistical Properties	HRePML	MLMM	FarmCPU	GEMMA
Average power (%)	60.39	53.15	53.21	36.34
Average MSE	0.0772	0.1068	0.1165	0.1933
Running time (Hour)	3.1419	22.7274	4.6653	2.4186

* The dataset used in Table 2 is the same as that used in Table 1.

To further validate the performance of HRePML, an additive polygenic effect was involved in the second simulation study. The dataset was the same with that used in the first simulation study, except that polygenic effect was added to the phenotype. The same trend in statistical power was observed, and the average powers of HRePML, MLMM, FarmCPU, and GEMMA were 62.45%, 57.08%, 55.18%, and 40.46%, respectively, which showed that HRePML was still powerful and robust under polygenic interference (Tables S1 and S2 and Figure 1B). As far as the mean squared error was concerned, the average MSE of the above four methods were 0.0926, 0.1343, 0.1184, and 0.2125, respectively. HRePML had the most accuracy of the QTN effect estimates, followed by FarmCPU, MLMM, and GEMMA (Tables S1 and S2 and Figure 2B).

In the third simulation study, we investigated the effect of the sample size on the statistical power of HRePML. There were four datasets consisting of 500, 1000, 2000, and 4000 individuals, respectively, and 10,000 SNP markers, with 100 replicates. Eight true QTNs were set in each replicate. Then, each dataset could be regarded as having 1,000,000 SNPs and 800 true QTNs in total. The average powers of sample sizes 500, 1000, 2000, and 4000 were 59.88%, 75.75%, 82.00%, and 91.13%, respectively. Clearly, the statistical power improved as the sample size increased. The results demonstrated that the statistical power could be more than 80% for QTN, with heritability equal or greater than 0.03 when the

sample size reached 1000. However, for QTN with very small heritability (0.01), the required sample size was at least 4000, and then, the statistical power could exceed 60% (Table 3 and Figure 3).

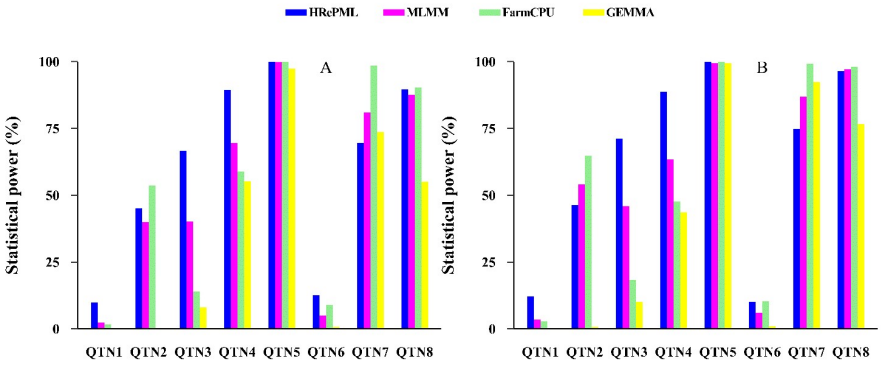

Figure 1. Comparison of statistical powers of eight simulated quantitative trait nucleotides (QTNs) using four genome-wide association study (GWAS) methods (hybrid of restricted and penalized maximum likelihood (HRePML), multi-locus mixed-model (MLMM), fixed and random model circulating probability unification (FarmCPU), and genome-wide efficient mixed model association (GEMMA)). (**A**) The first simulation study: no polygenic background. (**B**) The second simulation study: an additive polygenic variance involved.

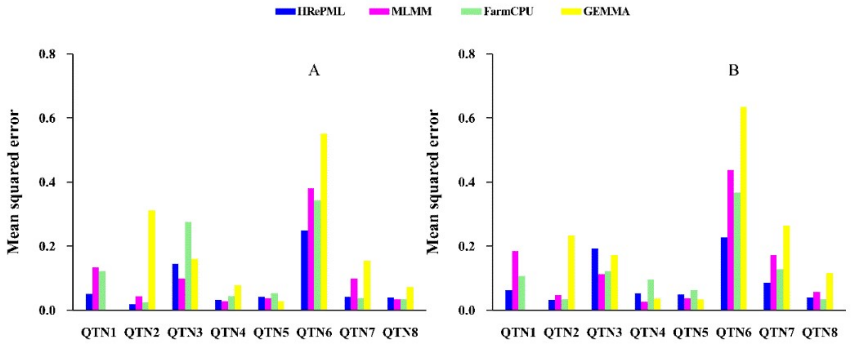

Figure 2. Comparison of mean squared errors of each simulated QTN effect using four GWAS methods (HRePML, MLMM, FarmCPU, and GEMMA). The descriptions in (**A**,**B**) are the same as those in Figure 1.

Table 3. Effect of the sample size on the statistical power and running time using the HRePML method in the third simulation study *.

QTN	R^2	Sample Size: Power (%)			
		500	1000	2000	4000
1	0.01	12	13	30	61
2	0.03	48	84	94	99
3	0.03	61	91	93	96
4	0.05	91	99	97	99
5	0.08	100	100	100	100
6	0.01	9	23	46	77
7	0.05	71	98	98	99
8	0.05	87	98	98	98
Average power (%)		59.88	75.75	82.00	91.13
Running time (Hour)		0.3142	1.1244	3.9969	39.5439

* In the third simulation study, there are four datasets consisting of 500, 1000, 2000, and 4000 individuals, respectively, and 10,000 SNP markers, with 100 replicates. Eight true QTNs are set in each replicate. Then, each dataset can be regarded as having 1,000,000 SNPs and 800 true QTNs in total.

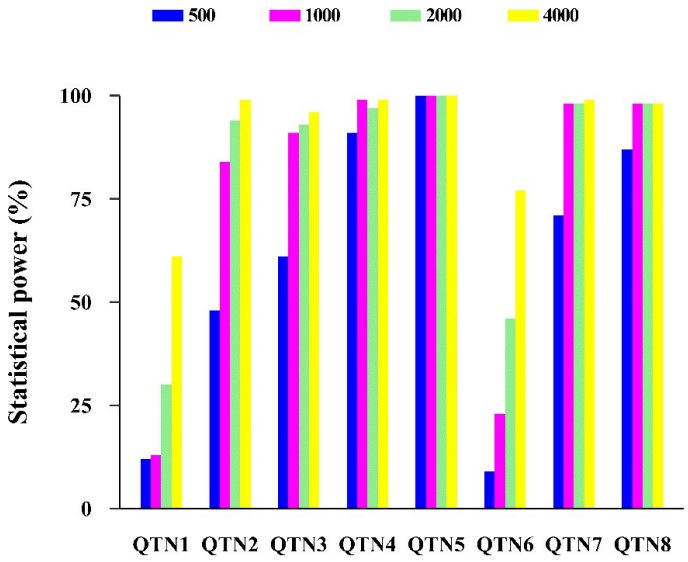

Figure 3. Effect of the sample size on the statistical power using HRePML in the third simulation study.

3.2. Running Time

All above four methods were carried out on the same machine (Intel® Core™ i5-7300HQ CPU 2.50 GHz, Memory 8.00 GB, Houston, TX, USA). In the first simulation consisting of 1000 repetitions, the total running time of HRePML, MLMM, FarmCPU, and GEMMA were 3.1419, 22.7274, 4.6653, and 2.4186 h, respectively. Compared with the other two multi-locus MLMM and FarmCPU methods, HRePML was the most computationally efficient, which was only slightly slower than the single-locus GEMMA method. In particular, HRePML achieved about seven times faster than the popular multi-locus method MLMM (Table 2 and Figure 4A). The second simulation also conducted with 1000 repetitions, and the same trend in total running time was observed, which was 3.2273, 27.4473, 4.9198, and 2.3855 h for the four methods, respectively. GEMMA was the fastest method, followed by HRePML, FarmCPU, and MLMM (Table S2 and Figure 4B). In the third and fourth simulations of 100 repeated experiments on the HRePML, the sample sizes and number of markers were investigated on the influence of running time. With sample sizes 500, 1000, 2000, and 4000, the running times were 0.3142, 1.1244, 3.9969, and 39.5439 h, respectively. The results showed that, as the sample size increased, the running time increased nonlinearly (Table 3 and Figure 5A). In the fourth simulation study, there were four datasets consisting of 10,000, 50,000, 100,000, and 200,000 SNP markers, respectively, and 500 individuals. With the number of markers 10,000, 50,000, 100,000, and 200,000, the running time was 0.3142, 1.5460, 3.2735, and 6.3574 h, respectively. Clearly, the running time increased almost linearly with the markers increasing (Figure 5B).

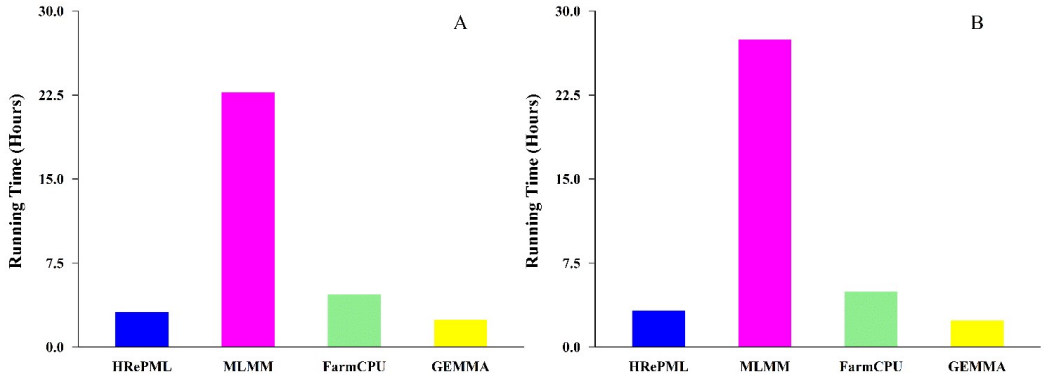

Figure 4. Comparison of total running time using four GWAS methods (HRePML, MLMM, FarmCPU, and GEMMA). The descriptions in (**A**,**B**) are the same as those in Figure 1.

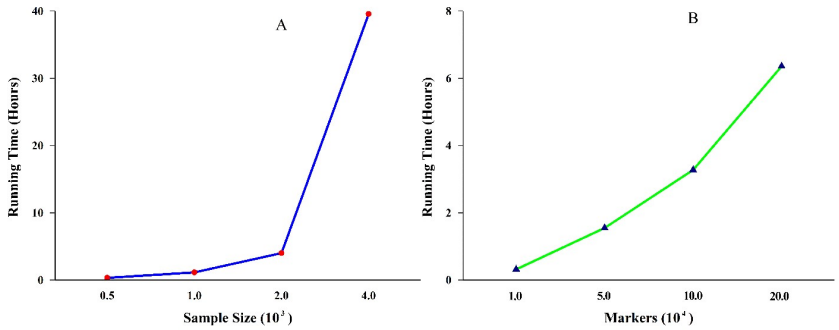

Figure 5. (**A**). Effect of the sample size on running time in the third simulation. (**B**) Effect of markers on the running time in the fourth simulation.

3.3. Association Analysis of Real Data in Arabidopsis

We performed GWAS on four development traits of *Arabidopsis* using HRePML, MLMM, FarmCPU, and GEMMA. The four methods identified 77, 43, 32, and 17 SNPs significantly associated with four traits. HRePML had the highest number of detected SNPs, which was more than four times than that of GEMMA detected (Table S4). Then, we performed gene ontology (GO) functional annotations on detected SNPs within their physical position 10 KB. As a result, the number of candidate genes detected by four methods were 41, 19, 25, and 5, which demonstrated that HRePML had the strongest ability to mine candidate genes, followed by FarmCPU, MLMM, and GEMMA (Tables S4 and S5). A total of eight genes were detected simultaneously by at least two methods. Interestingly, most of these eight genes were located on chromosome 5, while there was none located on chromosome 1. We found good agreement between the new methods HRePML and FarmCPU. It was worth noting that *AT5G45900* and *AT5G45940* could be identified by at least two methods on traits LC duration GH and LFS GH (Table 4). *AT5G45900* is a component of the autophagy conjugation pathway and contributes to plant basal immunity towards fungal infection. *AT5G45940* encodes a CoA pyrophosphatase and also has ppGpp pyrophosphohydrolase and exhibits minor activity of NADH pyrophosphatase and was most strongly expressed in embryo cotyledon and the hypocotyl, flower, and phloem of vascular tissues [45]. In summary, HRePML identified the most numbers of significantly associated SNPs and genes in the real data analysis (Table S5).

3.4. An Example for the Use of HRePML

To run HRePML requires four input files. The first input file is a genotype file, where each row represents the SNP marker, and each column represents an individual. The first two columns of the genotype file provide the chromosome and physical position information about the SNP marker. The genotype is coded as 0, 1, and 2, representing aa, Aa, and AA, where "A" is a dominant allele and "a" is a recessive allele. The second input file is the phenotype file, which is a column vector. The third input file is the genetic relatedness matrix or kinship file, which is a $N \times N$ matrix, and N is the number of individuals. The fourth input file is the covariates file, where the first column is the unit column vector, followed by the population structure or principal component matrix. The example data can be found at https://github.com/wenlongren/HRePML/tree/master/Example%20Data.

In a Linux system (Ubuntu), the compiling command is: g++ -I/path/liblbfgs-1.10/include HRePML-Linux.cpp -llapack -lblas -llbfgs -o output, where it needs to include the C++ library of limited-memory BFGS. If Math Kernel Library (MKL) has been installed for Intel CPU users, the following compiling command is recommended: g++ -I/path/liblbfgs-1.10/include HRePML-Linux.cpp -lmkl_gf_lp64 -lmkl_sequential -lmkl_core -llbfgs -o output. In order to save the results, the HRePML program requires two output files, which are the results file and the computational time file. After compilation, the execution command is: ./output Genotype.csv Phenotype.csv Kinship.csv Fixed.csv Results.csv Time.csv. Then, the results and running time are output into Results.csv and Time.csv files, respectively.

Table 4. Previously reported genes that were identified at least by two methods simultaneously with HRePML, MLMM, FarmCPU, and GEMMA.

Detected Genes	Associated Trait	Chr.	Position	Effect Estimate	LOD/*p*-Value	Methods
AT2G16440	LFS GH	2	7140030	−7.461, −9.107, −5.16	3.90×10^{-11}, 1.28×10^{-17}, 9.56×10^{-8}	FarmCPU, MLMM, GEMMA
AT3G07160	LFS GH	3	2280271	−5.934, −8.845	1.16×10^{-7}, 9.37×10^{-15}	FarmCPU, MLMM
AT3G54280	MT GH	3	20090780	1.002, 1.762	9.90×10^{-13}, 5.65×10^{-8}	FarmCPU, MLMM
AT4G09960	FT Duration GH	4	6228754	0.822, 1.136	3.74, 3.69×10^{-8}	HRePML, FarmCPU
AT4G33620	LC Duration GH	4	16140068	2.996, 2.540	4.78, 4.29×10^{-29}	HRePML, MLMM
AT5G45900, *AT5G45940*	LC Duration GH	5	18625634, 18625726	−3.707, −6.051	4.78, 2.51×10^{-28}	HRePML, FarmCPU
AT5G45900, *AT5G45940*	LFS GH	5	18625634, 18625726, 18625726	−4.318, −5.147, −5.616	5.23, 1.83×10^{-8}, 1.05×10^{-7}	HRePML, FarmCPU, GEMMA
AT5G53360	MT GH	5	21646741	0.236, 0.267	3.05×10^{-14}, 1.55×10^{-7}	FarmCPU, GEMMA

4. Discussion

In this paper, we proposed a new fast multi-locus method HRePML for GWAS, which is based on a restricted and penalized maximum likelihood function. HRePML can take genetic relatedness and population stratification into account under the linear mixed model. In addition, we implemented the algorithm in pure C++ language and provide Windows and Linux platform versions for the researcher's choice.

The new method adopts a two-stage approach to conduct multi-locus GWAS, which is widely used to improve the computational efficiency when hundreds of thousands of SNPs appear. The core idea is to conduct an initial screening of the marginal effects of all SNPs and select the ones with reasonably large effects for the next phase of the multi-locus test. Recently, multi-locus GWAS methods have become more and more popular, such as, MLMM [26], FarmCPU [27], mrMLM [32], pKWmEB (integration of Kruskal-Wallis test with empirical Bayes with polygenic background control) [43], and ISIS EM-BLASSO (iterative modified-sure independence screening and expectation-maximization bayesian least absolute shrinkage and selection operator) [44]. We drew on their successful experience and used the LOD value instead of the p value to determine the final identified SNPs. Using LOD equal to 3.0 as the threshold, many real data analyses show that it is feasible to improve the statistical power [32,43,44]. However, these multi-locus methods mentioned above are all programmed in R language, which are limited in analyzing large samples and massive SNP data. We implemented a new method HRePML in pure C++ language with the aid of the lapack, blas, libfgs, and boost C++ library. More importantly, the HRePML program can be further sped up with math kernel library (MKL) for Intel CPU users. Our first and second simulation experiments indicated that HRePML is about seven times faster than MLMM (Table 2 and Table S2 and Figure 4).

HRePML can be flexibly applied to animal and human GWAS, not limited to plant research. Genetic architecture is more complex, and the genome is much larger in animals and human beings than that in plants. One important issue is allelic heterogeneity, which cannot be effectively handled by traditional single-locus methods. More importantly, genetic heterogeneity can lead to a noncausative marker being a better descriptor of the phenotype than a causative one [46]. Another common issue is rare variant architecture, which may not always be resolved by increasing the sample size. HRePML, as one multi-locus method, can consider the complex genetic architecture and deal with these two issues well. Although the current version HRePML can analyze large samples with quantitative traits in humans, animals, or plants (Figure 5), it is not available to the UK BioBank scale data [47]. We recommend BOLT-LMM [10] for analyzing biobank scale samples.

The current study in *Arabidopsis* real data analysis showed that the results have relatively low consistency among HRePML, MLMM, FarmCPU, and GEMMA (Table 4 and Table S5). There are several possible reasons for this phenomenon. Firstly, the genetic structure of real data is more complex compared with simulated data, and large errors exist in phenotypic measurements. This can lead to reduce statistical power. Secondly, different methods are based on different assumptions and different models. The first three methods adopted a multi-locus model, while GEMMA used a single-locus model. Besides that, HRePML and MLMM were based on infinitesimal genetic architectures under a linear mixed model, while FarmCPU iterated on a fixed model and a random model. Thirdly, different methods respond differently to the effects of sample size, marker numbers, allele frequency, and heritability. In our opinion, there is complementarity between the various methods, and real data analysis requires considering the results of several methods together.

5. Conclusions

We proposed an alternative for fast multi-locus GWAS, based on the integration of the restricted and penalized maximum likelihood. Both the simulated and real data analyses demonstrated that our method HRePML improved the statistical power significantly compared with MLMM, FarmCPU, and GEMMA. In addition, HRePML can provide a higher accuracy estimation of the marker effects. More importantly, we developed an efficient tool in pure C++ for the Windows and Linux platform.

With the aid of the optimized math kernel library (MKL), HRePML can compute more efficiently when handling large individuals and millions of markers.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/11/11/1286/s1, Table S1: Comparison of statistical power and mean squared errors (MSE) for each QTN among HRePML, MLMM, FarmCPU and GEMMA methods in the second simulation study, Table S2: Comparison of average statistical power, average mean squared errors (MSE) and running time among HRePML, MLMM, FarmCPU and GEMMA methods in the second simulation study, Table S3: Parameters settings including true effect for each QTN with different sample size in the third simulation study and true effect for each QTN with different number of markers in the fourth simulation study, Table S4: The numbers of SNPs significantly associated with four development related traits in Arabidopsis thaliana and the number of genes around these SNPs identified by HRePML, MLMM, FarmCPU and GEMMA methods, Table S5: GWAS for four development related traits in *Arabidopsis thaliana* using HRePML, MLMM, FarmCPU and GEMMA methods.

Author Contributions: W.R. and J.X. conceived this work and designed the experiments. Z.L. and S.H. carried out the simulated experiments and plotted figures. W.R. and Z.L. analyzed real data. W.R. developed the program. W.R. and J.X. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Number 81803330), the Natural Science Foundations of Jiangsu Province (Grant Number BK20180950), and Nantong University Scientific Research Foundation for the Introduction of Talent (Grant Number 17R54).

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interests.

References

1. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; Harris, L.W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **2019**, *47*, D1005–D1012. [CrossRef] [PubMed]
2. Kichaev, G.; Bhatia, G.; Loh, P.R.; Gazal, S.; Burch, K.; Freund, M.K.; Schoech, A.; Pasaniuc, B.; Price, A.L. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **2019**, *104*, 65–75. [CrossRef]
3. Porcu, E.; Rueger, S.; Lepik, K.; eQTLGen Consortium; BIOS Consortium; Santoni, F.A.; Reymond, A.; Kutalik, Z. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **2019**, *10*, 3300. [CrossRef]
4. Ganjgahi, H.; Winkler, A.M.; Glahn, D.C.; Blangero, J.; Donohue, B.; Kochunov, P.; Nichols, T.E. Fast and powerful genome wide association of dense genetic data with high dimensional imaging phenotypes. *Nat. Commun.* **2018**, *9*, 3254. [CrossRef]
5. Xu, Y.; Xing, L.; Su, J.; Zhang, X.; Qiu, W. Model-based clustering for identifying disease-associated SNPs in case-control genome-wide association studies. *Sci. Rep.* **2019**, *9*, 13686. [CrossRef]
6. Lee, T.; Lee, I. araGWAB: Network-based boosting of genome-wide association studies in Arabidopsis thaliana. *Sci. Rep.* **2018**, *8*, 2925. [CrossRef]
7. Yang, J.; Zaitlen, N.A.; Goddard, M.E.; Visscher, P.M.; Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **2014**, *46*, 100–106. [CrossRef]
8. 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] [PubMed]
9. Zhou, X.; Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **2012**, *44*, 821–824. [CrossRef] [PubMed]
10. Loh, P.R.; Tucker, G.; Bulik-Sullivan, B.K.; Vilhjalmsson, B.J.; Finucane, H.K.; Salem, R.M.; Chasman, D.I.; Ridker, P.M.; Neale, B.M.; Berger, B.; et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **2015**, *47*, 284–290. [CrossRef] [PubMed]
11. Jiang, L.; Zheng, Z.; Qi, T.; Kemper, K.E.; Wray, N.R.; Visscher, P.M.; Yang, J. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **2019**, *51*, 1749–1755. [CrossRef]
12. Border, R.; Becker, S. Stochastic Lanczos estimation of genomic variance components for linear mixed-effects models. *Bmc Bioinform.* **2019**, *20*, 411. [CrossRef]
13. Hadfield, J.D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]

14. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Usinglme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
15. Lourenco, V.M.; Rodrigues, P.C.; Pires, A.M.; Piepho, H.P. A robust DF-REML framework for variance components estimation in genetic studies. *Bioinformatics* **2017**, *33*, 3584–3594. [CrossRef]
16. Cesarani, A.; Pocrnic, I.; Macciotta, N.P.P.; Fragomeni, B.O.; Misztal, I.; Lourenco, D.A.L. Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies. *J. Anim Breed. Genet.* **2019**, *136*, 40–50. [CrossRef]
17. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
18. Yuan, M. Model Selection and Estimation in Regression With Grouped Variables. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 49–67. [CrossRef]
19. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
20. Zhang, Y.M.; Xu, S. A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity* **2005**, *95*, 96–104. [CrossRef]
21. Hoffman, G.E.; Logsdon, B.A.; Mezey, J.G. PUMA: A unified framework for penalized multiple regression analysis of GWAS data. *PLoS Comput. Biol.* **2013**, *9*, e1003101. [CrossRef] [PubMed]
22. Tamuri, A.U.; Goldman, N.; dos Reis, M. A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **2014**, *197*, 257–271. [CrossRef]
23. Meyer, K. Simple Penalties on Maximum-Likelihood Estimates of Genetic Parameters to Reduce Sampling Variation. *Genetics* **2016**, *203*, 1885–1900. [CrossRef] [PubMed]
24. Gianola, D. Priors in whole-genome regression: The bayesian alphabet returns. *Genetics* **2013**, *194*, 573–596. [CrossRef]
25. Perez, P.; de los Campos, G. Genome-wide regression and prediction with the BGLR statistical package. *Genetics* **2014**, *198*, 483–495. [CrossRef] [PubMed]
26. Segura, V.; Vilhjalmsson, B.J.; Platt, A.; Korte, A.; Seren, U.; Long, Q.; Nordborg, M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **2012**, *44*, 825–830. [CrossRef]
27. Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet.* **2016**, *12*, e1005767. [CrossRef] [PubMed]
28. Sanyal, N.; Lo, M.T.; Kauppi, K.; Djurovic, S.; Andreassen, O.A.; Johnson, V.E.; Chen, C.H. GWASinlps: Non-local prior based iterative SNP selection tool for genome-wide association studies. *Bioinformatics* **2019**, *35*, 1–11. [CrossRef]
29. Sinoquet, C. A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies. *BMC Bioinform.* **2018**, *19*, 106. [CrossRef]
30. Sun, R.; Hui, S.; Bader, G.D.; Lin, X.; Kraft, P. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLoS Genet.* **2019**, *15*, e1007530. [CrossRef]
31. Hamazaki, K.; Iwata, H. RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Comput. Biol.* **2020**, *16*, e1007663. [CrossRef] [PubMed]
32. 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]
33. Xu, S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **2010**, *105*, 483–494. [CrossRef] [PubMed]
34. Rodrigue, N. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics* **2013**, *193*, 557–564. [CrossRef]
35. Zhu, C.; Byrd, R.H.; Lu, P.; Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560. [CrossRef]
36. Atwell, S.; Huang, Y.S.; Vilhjalmsson, B.J.; Willems, G.; Horton, M.; Li, Y.; Meng, D.; Platt, A.; Tarone, A.M.; Hu, T.T.; et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **2010**, *465*, 627–631. [CrossRef]

37. 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]
38. 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]
39. Schraudolph, N.N.; Yu, J.; Günter, S. A stochastic quasi-Newton method for online convex optimization. *AISTATS* **2007**, *2*, 436–443.
40. Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, *35*, 773–782. [CrossRef]
41. Schäling, B. *The Boost C++ Libraries*, 2nd ed.; XML Press: Laguna Niguel, CA, USA, 2014.
42. Cox, D.D.; O'Sullivan, F. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Stat.* **1990**, *18*, 1676–1695. [CrossRef]
43. 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]
44. 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] [PubMed]
45. The Arabidopsis Information Resource. Available online: https://www.arabidopsis.org/index.jsp. (accessed on 29 October 2020).
46. Platt, A.; Vilhjalmsson, B.J.; Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **2010**, *186*, 1045–1052. [CrossRef] [PubMed]
47. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209. [CrossRef] [PubMed]

Data Availability: The real datasets for this study can be found in the *Arabidopsis* Information Resource http://www.arabidopsis.org/. The C++ code implement of HRePML is available on https://github.com/wenlongren/HRePML.