A fast multi-locus random-SNP-effect EMMA for genome-wide association studies

Yang-Jun Wen¹, Hanwen Zhang², Jin Zhang¹, Jian-Ying Feng¹, Bo Huang¹, Jim M. Dunwell¹, Yuan-Ming Zhang¹, Yuan-Ming Zhang¹, Rongling Wu⁵

¹ State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China
² Faculty of Applied Science, The University of British Columbia, Vancouver V6T 1Z4, Canada
³ School of Agriculture, Policy and Development at University of Reading, Reading RG6 6AR, United Kingdom
⁴ Statistical Genomics Lab, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
⁵ Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA 17033-0850, USA

Corresponding authors:
Yuan-Ming Zhang, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China, E-mail: soyzhang@mail.hzau.edu.cn or State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China, E-mail: soyzhang@njau.edu.cn

Rongling Wu, Center for Statistical Genetics, Pennsylvania State University, Hershey, PA 17033, USA, E-mail: rwu@pbs.psu.edu
Abstract

Although the mixed linear model (MLM) such as efficient mixed model association (EMMA), has been widely used in genome-wide association studies (GWAS), relatively little is known about fast and efficient algorithms to implement multi-locus GWAS. To address this issue, we report a fast multi-locus random-SNP-effect EMMA (FASTmrEMMA). In this method, a new matrix transformation was constructed to obtain a new genetic model that includes only quantitative trait nucleotide (QTN) variation and normal residual error; letting the number of nonzero eigenvalues be one and fixing the polygenic-to-residual variance ratio was used to increase computing speed. All the putative QTNs with the ≤0.005 P-values in the first step of the new method were included in one multi-locus model for true QTN detection. Owing to the multi-locus feature, the Bonferroni correction is replaced by a less stringent selection criterion. Results from analyses of both simulated and real data showed that FASTmrEMMA is more powerful in QTN detection, model fit and robustness, has less bias in QTN effect estimation, and requires less running time than the current single- and multi-locus methodologies for GWAS, such as E-BAYES, SUPER, EMMA, CMLM and ECMLM. Therefore, FASTmrEMMA provides an alternative for multi-locus GWAS.

Key words genome-wide association study · mixed linear model · multi-locus mapping · random effect
Genome-wide association studies (GWAS) have been widely used in the genetic dissection of quantitative traits in human, animal and plant genetics, especially in combination with the output of genomic sequencing technologies. The most popular method for GWAS is the mixed linear model (MLM) method [1-2], because of its demonstrated effectiveness in correcting the inflation from many small genetic effects (polygenic background) and in controlling the bias of population stratification [3-7]. Since the MLM of Yu et al. [2] was published, many MLM-based methods have been proposed. However, most of them comprise a one-dimensional genome scan by testing one marker at a time, which is involved in multiple test correction for the threshold value of significance test. The widely-used Bonferroni correction is often too conservative to detect many important loci for quantitative traits.

Most quantitative traits are controlled by a few genes with large effects and numerous polygenes with minor effects. However, the current one-dimensional genome scan approaches for GWAS do not match the true genetic model for these traits. To overcome this issue, multi-locus methodologies have been recommended, for example, Bayesian LASSO [8], adaptive mixed LASSO [9], penalized Logistic regression [10-11], Elastic-Net [12], empirical Bayes [13] and empirical Bayes LASSO [14]. If the number of markers is several times larger than the sample size, all marker effects can be included in one single model and estimated in an unbiased way. If the number of markers is much many times larger than the sample size, however, these shrinkage approaches will fail. In this situation, we should consider how to reduce the number of marker effects in the multi-locus genetic model. For example, Zhou et al. [15] developed a Bayesian sparse linear mixed model and Moser et al. [16] proposed a Bayesian mixture model. Under these models, two to four common components in the mixture distribution were considered and only a few variance components were estimated. Although about 500 effects in the genetic model are finally considered after several rounds of Gibbs sampling, the computing time becomes a major concern for these Bayesian approaches. Recently, Segura et al. [17] and Wang et al. [7] have proposed multi-locus MLM approaches. However, further refinement for fast algorithm is needed.

In the MLM methods of Zhang et al. [1], the QTN effect is viewed as random so that three component variances (QTN, polygenic and residual variances) need to be estimated. If the number
of markers is large, this calculation takes a long time. To reduce computing time and increase power in QTN detection, a compressed MLM (CMLM) with a P3D algorithm [18] and an enriched CMLM (ECMLM) [19] have been proposed. On the other hand, Kang et al. [3] proposed an efficient mixed model association (EMMA), and other authors suggested alternatives, such as EMMA eXpedited (EMMAX) [20], FaST-LMM [21], FaST-LMM-Select [22], genome-wide EMMA [4] and GRAMMA-Gamma [23]. Recently, SUPER [24] has been developed based on FaST-LMM. Among the above fast methods, the SNP effect was viewed as fixed. It should be noted that Goddard et al. [25] treated marker effects as random, due to several advantages of the random model approach over the fixed model treatment [7,25-27]. For example, the random model approach will shrink the estimated SNP effects toward zero. However, Goddard et al. [25] did not provide an efficient computational algorithm to estimate marker effects. Although the random-SNP-effect MLM [7,27] can solve this issue, a new and efficient algorithm for fast calculation is also required.

In this paper, first, we treated the SNP effect as random. Although three component variances need to be estimated, a fast and new matrix transformation in the new method was constructed in order to quickly scan each marker on the genome. Then, all the putative QTN with a ≤0.005 P-value in the first step of the new method were placed into one multi-locus genetic model and these QTN effects were estimated by EM empirical Bayes (EMEB) [28] for true QTN identification. This new method, called fast multi-locus random-SNP-effect EMMA (FASTmrEMMA), was validated by analysis of real data from Arabidopsis [29] and by a series of simulation studies, as compared with the other methods: E-BAYES (multi-locus model) [30], SUPER, EMMA, CMLM and ECMLM (single-locus model).

The new methods have been implemented in R. The R software FASTmrEMMA is available.

RESULTS

Fast multi-locus random-SNP-effect EMMA (FASTmrEMMA)

Estimation of the QTN variance. FASTmrEMMA is a new algorithm that can approximate
the estimation of QTN variance (see Materials and Methods). Thus, we need to know whether this approximation has a significant effect on the estimate of QTN variance. To answer this question, four flowering time traits in \textit{Arabidopsis} [29] were analyzed by FASTmrEMMA and an exact method implemented by PROC MIXED in SAS. The estimates for QTN variance are listed in Figure 1 and Table S1. As a result, the relative error between the two methods ranged from 0.0 to 24.09%, and the average was 1.60%, indicating no effect on the QTN variance estimate using FASTmrEMMA under the conditions of this simulation.

To confirm the effectiveness of FASTmrEMMA, three Monte Carlo simulation experiments were carried out and the simulation procedures were same as those in Wang et al. [7]. Each sample in these simulation experiments was analyzed by six methods. In the six methods, FASTmrEMMA is also a new multi-locus algorithm within the framework of MLM, E-BAYES [30] is an existing multi-locus approach under the framework of Bayesian statistics, and SUPER, EMMA, ECMLM and CMLM are the existing single-locus GWAS methods.

\textbf{Statistical power for QTN detection.} In the above three simulation experiments, the power for each QTN was defined as the proportion of samples where the QTN was detected (the P-value is smaller than the designated threshold). When only six QTNs were simulated in the first experiment, the power in the detection of each QTN was higher for FASTmrEMMA than for the others (Figure 2a; Table S2). When a polygenic background ($h^2_{pq} = 0.075$) was added to the first experiment, a similar trend was observed (Figure 2b; Table S2). On one occasion (QTN number 1), E-BAYES method was slightly more powerful than FASTmrEMMA method. When the polygenic background was changed into an epistatic background ($h^2_{pq} = 0.15$), the results were also similar to those in the first experiment (Figure 2c; Table S2). These results demonstrate the highest power of FASTmrEMMA across all the approaches under various genetic backgrounds, although the other methods are also robust under these backgrounds.

\textbf{Accuracy for estimated QTN effects.} We used the average, MSE and MAD to measure the accuracy of an estimated QTN effect. We evaluated the accuracies for the estimates of all the six
simulated QTNs across all the six methods. As a result, the estimate of each QTN effect from FASTmrEMMA was much closer to the true value than the estimates obtained from the other methods. On these occasions (QTN numbers 1 and 4), the averages from E-BAYES were closer to the true value than those from FASTmrEMMA in three simulation experiments (Table S2). The MSE and MAD for each QTN effect were significantly less from FASTmrEMMA than from the others with an exception for QTN number 4 where E-BAYES has slightly less MSE and MAD than FASTmrEMMA in the third simulation experiment (Figures 2d-2i; Table S2). These results indicate that a higher accuracy for the estimate of QTN effect can be achieved using FASTmrEMMA than using the other methods.

**False Positive Rate and ROC curve.** All the false QTN, detected by the six methods, in three simulation experiments were used to calculate the empirical false positive rates of the six methods. These results are listed in Table S3. In these three simulation experiments, the empirical false positive rates of the six methods were between 0.096 and 5.181 (× 1E-4), and had the same order of magnitude. EMMA has the lowest false positive rate followed by ECMLM, CMLM and FASTmrEMMA methods, and SUPER has the maximum false positive rate followed by E-BAYES method.

A receiver operating characteristic (ROC) curve is a plot of the statistical power against the controlled type I error. This curve is frequently used to compare different methods for their efficiencies in the detection of significant effects; the higher the curve, the better is the method. When eleven probability levels for significance, between 1e-8 to 1e-3, were inserted, the corresponding powers were calculated in the second simulation experiment. The results are shown in Figure 3. Among the six approaches, clearly, FASTmrEMMA method is the best one and the next one is E-BAYES.

**Computing time.** In each of three simulation experiments, computing times for the six methods were recorded and are listed in Table S4. In summary, FASTmrEMMA has the least computing time followed by ECMLM, E-BAYES, CMLM and SUPER methods, and EMMA has the maximum computing time.
**Real data analysis in Arabidopsis**

To validate FASTmrEMMA, this new method along with E-BAYES, SUPER, EMMA, CMLM and ECMLM was used to re-analyze the *Arabidopsis* data [29] for days to flowering under long days (LD), days to flowering under short days with vernalization (SDV), leaf number at flowering with 8 wks vernalization, greenhouse (8W GH LN), and days to flowering, 8 wks vernalization, greenhouse (8W GH FT), and the results are listed in Table S5.

The numbers of SNPs significantly associated with the above four traits were 20, 17, 14 and 17, respectively, for traits LD, SDV, 8W GH LN and 8W GH FT, from FASTmrEMMA method. The corresponding numbers of the associated SNPs were 2, 6, 1 and 5 from E-BAYES; 21, 0, 0 and 0 from SUPER; 1, 5, 0 and 2 from EMMA; and 0, 1, 0 and 0 from both CMLM and ECMLM. Clearly, the number of significantly associated SNPs was much larger from FASTmrEMMA than from the other methods. These significantly associated SNPs for each trait were used to conduct a multiple linear regression analysis and the corresponding Bayesian information criteria (BIC) were calculated. For example, the BIC value for the model of 8W GH LN was -103.47 for FASTmrEMMA, 77.76 for E-BAYES and 117.50 for the others. FASTmrEMMA method shows the lowest BIC values for all the four traits (Table 1), indicating the best model fit among the six approaches.

Based on the SNPs detected by FASTmrEMMA, 6, 11, 5 and 7 genes were previously reported to be associated with the above four traits [32-34]. In the vicinity of the SNPs detected by E-BAYES, the corresponding numbers of the known genes are 2, 1, 0 and 1, respectively, for the above four traits [32]. Only four known genes for LD (SUPER), two known genes for LD (EMMA), and three known genes for SDV (EMMA) are in the neighborhood of the detected SNPs [32,34] (Table 2). Clearly, FASTmrEMMA method detected more known genes than did the other methods.

We also compared all the known genes detected in this study with all the candidate genes in Atwell *et al.* [29]. For example, among seven known genes (*At1g03457, At2g27380, At2g47230, At3g56900, At3g57000, At5g06550 and At5g06590*) for 8W GH FT in this study, no genes were within the 133 candidate genes in Atwell *et al.* [29] (2010). Among 11 known genes for SDV in
this study, only three genes (At5g04240, At5g57360 and At5g57390) were within the 153 candidate genes in Atwell et al. [29].

**Discussion**

When SNP effect is viewed as random, three variance components will be estimated. Generally, polygenic variance is larger than zero while variance components for most SNPs are zero, because these markers are not associated to the trait of interest. In other words, as in most mixed model approaches, variance components in FASTmrEMMA are also estimated under the assumption that one variance component is zero.

FASTmrEMMA is a new algorithm, and different from the widely-used one-dimensional genome scan approaches, such as SUPER, EMMA, CMLM and ECMLM. First, the SNP effects are viewed as random in FASTmrEMMA while they are viewed as fixed in SUPER, EMMA, CMLM and ECMLM. Because the random model approach will shrink the estimated SNP effects toward zero when the simulated QTN effects are small, leading to maximum correlation between observed and predicted phenotypic values [25,35]. Meanwhile, the power in the detection of QTN with random effect is higher than that with fixed effect [36].

Then, a quick single marker genome scan method was proposed to estimate the three variance components in the above mixed model. Here several techniques have been incorporated into the quick algorithm. The first technique is to fix the polygenic-to-residual variance ratio, which was adopted in CMLM/P3D [18] and EMMAX [20]. Although this algorithm is approximate, it has almost no effect on the estimate of SNP-effect variance, even if there is a large difference in the above ratios between the approximate and exact algorithms (Table S1). Clearly, this provides evidence for fixing the ratio in FASTmrEMMA. The second technique is to use a quick matrix calculation algorithm, such as, the eigen decomposition of matrix $X_cX_c^T$ is the same as that of $X_c^TX_c$ (a positive number). Thus, eigen decomposition, determinant and derivatives in the estimation of $\lambda_{\beta}$ can be quickly calculated. The final technique is to estimate residual variance along with the estimation of fixed effects. In the single marker genome scan, therefore, only one
parameter $\lambda_\beta$ needs to be estimated so that running time is obviously decreased. Although GCTA
algorithm [37] may be used to estimate the above three variance components, running time is a
major concern. A similar situation is also apparent when using PROC MIXED in SAS in Zhang et
al. (2005) [1].

Finally, our matrix transformation algorithm in FASTmrEMMA is different from those in SUPER,
EMMA, CMLM, ECMLM and mrMLM. For example, when many random effects are included
simultaneously in one genetic model and polygenic background also needs to be controlled, at
present there are no methods available. However, the new matrix transformation algorithm can
transfer polygenic background plus residual error into a normal residual error. This new model can
be easily treated by a Bayesian method. The applied study will be reported in the near future.

The multi-variance-component algorithm, E-BAYES [30], was also used to conduct multi-locus
GWAS, especially for the situation where the number of markers is several times larger than
sample size. However, results from simulation experiments showed that FASTmrEMMA is more
powerful in QTN detection and higher accurate in QTN effect estimation than is E-BAYES. Note
that E-BAYES method detected no genes in the real data analyses if all the 180,000 SNPs were
simultaneously included in one genetic model (Table S5). To solve this issue, only a proportion of
the SNPs, for example 2000 to 8000 SNPs, is included in one model at a time, so these SNP
effects may be estimated by E-BAYES. FASTmrEMMA is different from the adaptive mixed
LASSO [9]. If the number of markers is many larger than sample size, the adaptive mixed LASSO
does not work. FASTmrEMMA is also different from both the Bayesian sparse linear mixed model
[15] and the Bayesian mixture model [16]. The latter two operate under the framework of
Bayesian statistics, and the computing time becomes a major concern.

FASTmrEMMA is different from MLMM of Segura et al. [17] in two aspects. First, MLMM is a
simple, stepwise mixed-model regression with forward inclusion and backward elimination and
FASTmrEMMA is a two-step combined method. In MLMM, the computationally-intensive
forward-backward inclusion of SNPs is clearly a limiting factor in exploring the huge model space
[17]. Second, matrix transformation algorithm in MLMM is different from that in FASTmrEMMA.
This difference also exists between FASTmrEMMA and mRMLM of Wang et al. [7].

As described by Wang et al. [7], single-locus genome scan approaches for GWAS require Bonferroni correction for multiple tests. However, this correction often is too conservative to detect important loci for quantitative traits when the number of markers is extremely large. Clearly, FASTmrEMMA is based on a multi-locus model. Due to the multi-locus nature, Bonferroni correction is replaced by a less stringent selection criterion. Results from analysis of simulated and real data further validated the idea of a less stringent selection criterion in this study.

FASTmrEMMA is a combined method with two steps, each of which needs a critical P-value. In the first step, we compared the effect of three critical P-values (0.01, 0.005 and 0.001) on the power of QTN detection (Table S6) and found that the 0.005 critical P-value is the best. In the second step, a less stringent selection criterion between 0.05 and 0.05/$p$ was adopted. The two critical P-values in FASTmrEMMA are confirmed by our simulated and real data analysis.

**Conclusion**

In FASTmrEMMA algorithm, random-SNP-effect and multi-locus model methods are used to improve the power for QTN detection and to decrease the false positive rate, a new matrix transformation in the first step of FASTmrEMMA is constructed to obtain a new genetic model that includes only QTN variation and normal residual error. Additionally, letting the number of nonzero eigenvalues be one and fixing the polygenic-to-residual variance ratio are used to save running time. As a result, FASTmrEMMA has the highest power and accuracy for QTN detection and the best fit for a genetic model, as compared with E-BAYES, SUPER, EMMA, CMLM and ECMLM.

**Materials and Methods**

**Genetic model**

We consider the following standard MLM:

$$y = W\alpha + X\beta + Zu + \epsilon$$

(1)
where \( \mathbf{y} \) is an \( n \times 1 \) phenotypic vector of quantitative trait, and \( n \) is the number of individuals; \( \mathbf{W} = (w_1, w_2, \cdots, w_c) \) is an \( n \times c \) matrix of covariates (fixed effects) including a column vector of 1, population structure [2] or principle component [39] may be incorporated into \( \mathbf{W} \), and \( \mathbf{a} \) is a \( c \times 1 \) vector of fixed effects including the intercept; \( \mathbf{X} \) is an \( n \times 1 \) vector of marker genotypes, and \( \beta \sim N(0, \sigma_\beta^2) \) is random effect of putative QTN; \( \mathbf{Z} \) is an \( n \times m \) design matrix, \( \mathbf{u} \sim \text{MVN}_n(0, \sigma_u^2 \mathbf{K}) \) is an \( m \times 1 \) vector of polygenic effects; \( \mathbf{K} \) is a known \( m \times m \) relatedness matrix; and \( \mathbf{e} \sim \text{MVN}_n(0, \sigma_e^2 \mathbf{I}_n) \) is an \( n \times 1 \) vector of residual errors, \( \sigma^2_e \) is the variance of residual error, \( \mathbf{I}_n \) is an \( n \times n \) identity matrix, and MVN denotes multivariate normal distribution.

In animal datasets, \( m \) is the number of strains, \( n \) is the number of animals and \( \mathbf{Z} \) indicates which strain each animal belongs to (\( z_{ij} = 1 \) if individual \( i \) comes from strain \( j \) and \( z_{ij} = 0 \) otherwise); and in the \textit{Arabidopsis thaliana} data set of Atwell \textit{et al.} [29], \( m = n \) and \( \mathbf{Z} = \mathbf{I}_n \).

In the current methods, including EMMA [3], CMLM/P3D [18], ECMLM [19], EMMAX [20], FaST-LMM [21], FaST-LMM-Select [22], SUPER [24], GEMMA [4], and GRAMMA-Gamma [23], \( \beta \) is treated as a fixed effect, from which it is relatively easy to estimate \( \sigma^2_\beta \) and \( \sigma^2_e \). In this study we treat \( \beta \) as random in order to make the model more realistic [25,35,36]. In this case, three variance components need to be estimated under the assumption that QTN variance is zero, because most SNPs are not associated with the trait of interest. So the variance of \( \mathbf{y} \) in the model (1) is

\[
\text{Var}(\mathbf{y}) = \sigma_\beta^2 \mathbf{X}^T \mathbf{X} + \sigma_e^2 \mathbf{ZKZ}^T + \sigma_e^2 \mathbf{I}_n
= \sigma_e^2 (\lambda_\beta \mathbf{X}^T \mathbf{X} + \lambda_e \mathbf{ZKZ}^T + \mathbf{I}_n)
= \sigma_e^2 \mathbf{H}
\]

where \( \lambda_\beta = \sigma_\beta^2 / \sigma_e^2 \) and \( \lambda_e = \sigma_e^2 / \sigma_e^2 \).

**Fast multi-locus random-SNP-effect EMMA (FASTmrEMMA)**

The key to solve the model (1) is to estimate \( \sigma^2_\beta \), \( \sigma^2_e \) and \( \sigma^2_e \). Although many algorithms or estimations are available, such as analysis of variance, maximum likelihood (ML), restricted
maximum likelihood (REML), minimum norm quadratic unbiased, spectral decomposition [40] and average information [41], they are not feasible for a very high number of SNPs. Hence, we proposed a fast and efficient approximation algorithm in this study.

In the first step, we considered the reduced form of the model (1), which deleted \( X\beta \),

\[ y = W\alpha + Zu + \epsilon \]  

(3)

The variance of \( y \) is:

\[ \text{Var}(y) = \sigma^2 I_n + \sigma^2 ZKZ^T + \sigma^2 I_n \]

\[ = \sigma^2 (\lambda_y ZKZ^T + I_n) \]  

(4)

Using EMMA algorithm of Kang et al. [3], the estimate of \( \lambda_y \), denoted by \( \hat{\lambda}_y \), can be easily obtained.

In the second step, we considered the model (1), and replaced \( \lambda_y \) in (2) by the \( \hat{\lambda}_y \), so

\[ \text{Var}(y) = \sigma^2 (\lambda_{\beta} XX^T + \hat{\lambda}_y ZKZ^T + I_n) \]

\[ = \sigma^2 (\lambda_{\beta} XX^T + B) \]  

(5)

where \( B = \hat{\lambda}_y ZKZ^T + I_n \). An eigen (or spectral) decomposition of the positive semi-definite matrix \( B \) was

\[ B = Q_B \Lambda_B Q_B^T \]

\[ = (Q_1, Q_2) \begin{pmatrix} \Lambda_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \end{pmatrix} \]

(6)

\( (\text{Appendix A}) \), where \( Q_B \) is orthogonal, \( \Lambda_r \) is a diagonal matrix with positive eigenvalues, \( r = \text{Rank}(B) \). \( Q_1 \) and \( Q_2 \) are the \( n \times r \) and \( n \times (n - r) \) block matrices of \( Q_B \), \( 0 \) is the corresponding block zero matrix.

Let \( C = Q_1 \Lambda_r^{-\frac{1}{2}} Q_1^T \) , the model (1) was changed into

\[ y_c = W_c \alpha + X_c \beta + \epsilon_c \]  

(7)
where \( \mathbf{y}_c = \mathbf{C} \mathbf{y} \), \( \mathbf{W}_c = \mathbf{C} \mathbf{W} \), \( \mathbf{X}_c = \mathbf{C} \mathbf{X} \) and \( \mathbf{e}_c = \mathbf{C} \mathbf{Z} \mathbf{u} + \mathbf{C} \mathbf{e} \). Clearly, the model (7) is a new MLM, \( \mathbf{e}_c \sim \text{MVN}_n(\mathbf{0}, \sigma_c^2 \mathbf{I}_n) \), and
\[
\text{Var}(\mathbf{y}_c) = \sigma_c^2(\mathbf{X}_c^T \mathbf{X}_c + \mathbf{I}_n) = \sigma_c^2 \mathbf{H}_c = \mathbf{V}_c
\] (Appendix B). Once the ratio of \( \sigma_g^2 \) and \( \sigma_e^2 \), \( \lambda_g = \sigma_g^2/\sigma_e^2 \), was fixed at \( \hat{\lambda}_g \), it is possible to scan each marker on the genome. The evidence for the effectiveness of this approximation is shown in the results section.

Log-likelihood and restricted log-likelihood functions. According to the descriptions for the single-locus genome scan algorithm in previous GWAS studies [3,4,43,44], log-likelihood and restricted log-likelihood functions for the model (7) are
\[
l_e(\lambda_g, \sigma_g^2, \mathbf{a}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma_g^2) - \frac{1}{2} \log(\mathbf{H}_c) - \frac{1}{2\sigma_g^2} (\mathbf{y}_c - \mathbf{W}_c \mathbf{a})^T \mathbf{H}_c^{-1} (\mathbf{y}_c - \mathbf{W}_c \mathbf{a})
\] (9)
and
\[
l_r(\lambda_g, \sigma_g^2, \mathbf{a}) = -\frac{v}{2} \log(2\pi) - \frac{v}{2} \log(\sigma_g^2) + \frac{1}{2} \log(\mathbf{W}_c \mathbf{W}_c^T) - \frac{1}{2\sigma_g^2} (\mathbf{y}_c - \mathbf{W}_c \mathbf{a})^T \mathbf{H}_c^{-1} (\mathbf{y}_c - \mathbf{W}_c \mathbf{a})
\] (10)
respectively, where \( \mathbf{H}_c = \hat{\lambda}_g \mathbf{X}_c \mathbf{X}_c^T + \mathbf{I}_n \), \( v = n - a \), \( a = \text{rank}(\mathbf{W}_c) \leq \min(e, r) \), \( c = \text{rank}(\mathbf{W}) \), \( r = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{B}) \), supposing \( \mathbf{W} \) is column full rank.

Once \( \mathbf{a} \) and \( \sigma_g^2 \) are fixed, the ML and REML estimates for \( \hat{\lambda}_g \) is equivalent to maximizing the following target functions
\[
l_e(\lambda_g|\mathbf{a}, \sigma_g^2) \propto -\frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2} \log(\mathbf{H}_c) - \frac{n}{2} \log(\mathbf{y}_c^T \mathbf{P}_c \mathbf{y}_c)
\] (11)
and
\[
l_r(\lambda_g|\mathbf{a}, \sigma_g^2) \propto -\frac{v}{2} \log\left(\frac{2\pi}{v}\right) - \frac{v}{2} \log(\mathbf{W}_c \mathbf{W}_c^T) - \frac{1}{2} \log(\mathbf{H}_c) - \frac{v}{2} \log(\mathbf{y}_c^T \mathbf{P}_c \mathbf{y}_c)
\] (12)
where \( \mathbf{P}_c = \mathbf{H}_c^{-1} - \mathbf{H}_c^{-1} \mathbf{W}_c (\mathbf{W}_c^T \mathbf{H}_c^{-1} \mathbf{W}_c)^{-1} \mathbf{W}_c^T \mathbf{H}_c^{-1} \), and \(-\) denotes generalized inverse.

Since it is very slow to calculate determinant and inversion in the equations (11) and (12), a fast
computation algorithm should be considered. As described in GEMMA, Zhou and Stephens [4] first obtained the first and second derivatives for \( \lambda_\beta \), and then conducted eigen (or spectral) decomposition. In EMMA, however, Kang et al. [3] first conducted eigen decomposition, and then calculated the derivatives. These two ways are essentially the same. For simplicity, we adopted EMMA method.

It is possible to find \( \xi_i \) and \( \delta_s \), such that

\[
H_i = \lambda_\beta \mathbf{X}_i \mathbf{X}_i^T + \mathbf{I}_n = \mathbf{U}_n \text{diag} \left( \lambda_\beta \xi_i + 1, \cdots, \lambda_\beta \xi_n + 1 \right) \mathbf{U}_n^T = \mathbf{U}_n \text{diag} \left( \lambda_\beta \xi_i + 1, 1, \cdots, 1 \right) \mathbf{U}_n^T \tag{13}
\]

\[
M_i \mathbf{H}_i \mathbf{M}_i = \mathbf{M}_i \left( \lambda_\beta \mathbf{X}_i \mathbf{X}_i^T + \mathbf{I}_n \right) \mathbf{M}_i
\]

\[
= \left( \mathbf{U}_n \mathbf{W}_h \right) \text{diag} \left( \lambda_\beta \delta_1 + 1, \cdots, \lambda_\beta \delta_1 + 1, 0, \cdots, 0 \right) \left( \mathbf{U}_n \mathbf{W}_h \right)^T
\]

\[
= \mathbf{U}_n \text{diag} \left( \lambda_\beta \delta_1 + 1, 1, \cdots, 1 \right) \mathbf{U}_n^T \tag{14}
\]

where \( \mathbf{M}_i = \mathbf{I}_n - \mathbf{W}_i \left( \mathbf{W}_i^T \mathbf{W}_i \right) \mathbf{W}_i^T \), \( \mathbf{U}_n \) is an \( n \times n \) orthogonal matrix, and \( \mathbf{U}_h \) is an \( n \times v \) eigenvector matrix corresponding to the nonzero eigenvalues, and \( \mathbf{W}_h \) is an \( n \times (n - v) \) eigenvector matrix corresponding to zero eigenvalues.

Note that \( \xi_i = 0 \quad (i = 2, \cdots, n) \) and \( \delta_s = 0 \quad (s = 2, \cdots, v) \). This is because the nonzero eigenvalues of \( \mathbf{X}_i \mathbf{X}_i^T \) are the same as those of \( \mathbf{X}_i^T \mathbf{X}_i \), which is a positive number. For technical detail the reader is referred to Appendix C. It should be noted that \( \mathbf{U}_n \) and \( \mathbf{U}_h \) are independent of \( \lambda_\beta \). Let \( \mathbf{U}_n^T \mathbf{y}_s = (\eta_1, \cdots, \eta_v)^T \), then finding the ML and REML estimates for \( \lambda_\beta \) is equivalent to optimizing the following functions with respect to \( \lambda_\beta \) (Appendix D):

\[
l_\lambda (\lambda_\beta | \mathbf{a}, \sigma^2) = -\frac{n}{2} \log \left( \frac{2\pi}{n} \right) - \frac{n}{2} - \frac{1}{2} \log \left( \lambda_\beta \xi_1 + 1 \right) - \frac{n}{2} \log \left( \frac{\eta^2}{\lambda_\beta \delta_1 + 1 + \sum_{i=2}^v \eta_i^2} \right) \tag{15}\]

\[
l_\mu (\lambda_\beta | \mathbf{a}, \sigma^2) = -\frac{v}{2} \log \left( \frac{2\pi}{v} \right) - \frac{v}{2} - \frac{1}{2} \log \left( \lambda_\beta \delta_1 + 1 \right) - \frac{v}{2} \log \left( \frac{\eta^2}{\lambda_\beta \delta_1 + 1 + \sum_{i=2}^v \eta_i^2} \right) \tag{16}\]

**Estimation of parameter** \( \lambda_\beta \). At present three algorithms, Newton-Raphson (NR), Fisher scoring and expectation-maximization, are frequently used to obtain the ML and REML estimates [45]. In this study we adopted the NR algorithm, which has the form
where \( \lambda_{g,F}^{(1)} \) and \( \lambda_{g,R}^{(1)} \) are the ML and REML estimates at the \( t \)th iteration, respectively; and \( \lambda_{g,F}^{(1+1)} \) and \( \lambda_{g,R}^{(1+1)} \) are the update estimates, respectively, and the first derivatives of these two functions on \( \lambda_g \) were

\[
l'_t(\lambda_g|\sigma_g^2) = -\frac{\xi_t}{2(\lambda_g^2 + 1)} + n \frac{\delta_t \eta_t}{\eta_t^2/(\lambda_g^2 + 1) + \sum_{i=2}^n \eta_i^2}
\]

(18)

and the second derivatives on \( \lambda_g \) were

\[
l''_t(\lambda_g|\sigma_g^2) = \frac{\sigma^2}{2(\lambda_g^2 + 1)^2} - \frac{n}{2} \times \frac{\delta_t^2 \eta_t^2 \left[ \eta_t^2 + 2 \left( \sum_{i=2}^n \eta_i^2 \right) \left( \lambda_g^2 + 1 \right) \right]}{\eta_t^2 \left( \lambda_g^2 + 1 \right) + \sum_{i=2}^n \eta_i^2 \left( \lambda_g^2 + 1 \right)^2}
\]

(20)

Using the idea of EMMA [3], the range of \( \lambda_g \) is between 1E-05 (corresponding to almost pure environmental effect) to 1E+05 (corresponding to almost pure single-gene effect), and we divided this range evenly into 100 regions in logarithm scale to compute (18) or (19). The global ML or REML is searched for by applying the NR algorithm to all the intervals where the signs of derivatives change. This optimization technique for estimating \( \lambda_g \) has guaranteed the convergence as long as the kinship matrix \( K \) is positive semi-definite. Note that \( XX^T \) is always positive semi-definite.

**Estimation of fixed effects** \( \alpha \) and **residual variance** \( \sigma_e^2 \). Once \( \lambda_{g,F} \) is known, it is easy to
estimate $\alpha$ and $\sigma^2$. The ML estimates for $\alpha$ and $\sigma^2$ were

$$\hat{\alpha}_F = \left( W_F^T \hat{H}_{e,F}^{-1} W_F \right)^{-1} W_F^T \hat{H}_{e,F}^{-1} \mathbf{y},$$

$$\hat{\sigma}^2_{e,F} = \frac{1}{n} \left( \frac{\eta_i^2}{\hat{\lambda}_{\beta,F} \delta_i + 1} + \sum_{i=2}^{\sigma^2/\sigma_i^2} \right)$$

(22)

respectively, where $\hat{H}_{e,F}^{-1} = U_F \text{diag} \left( \frac{1}{\hat{\lambda}_{\beta,F} \xi_i + 1}, \ldots, 1 \right) U_F^T$. The REML estimates for $\alpha$ and $\sigma^2$ were

$$\hat{\alpha}_R = \left( W_R^T \hat{H}_{e,R}^{-1} W_R \right)^{-1} W_R^T \hat{H}_{e,R}^{-1} \mathbf{y},$$

$$\hat{\sigma}^2_{e,R} = \frac{1}{v} \left( \frac{\eta_i^2}{\hat{\lambda}_{\beta,R} \delta_i + 1} + \sum_{i=2}^{\sigma^2/\sigma_i^2} \right)$$

(23)

respectively, where $v = n - \alpha$ and $\hat{H}_{e,R}^{-1} = U_R \text{diag} \left( \frac{1}{\hat{\lambda}_{\beta,R} \xi_i + 1}, \ldots, 1 \right) U_R^T$ (Appendix D).

**Best linear unbiased prediction (BLUP) for parameter $\beta$.** Using $\text{Cov}(\mathbf{y}, \beta)^T$, the BLUP for the $\beta$, $\hat{\beta} = E(\beta | \mathbf{y})$, can be obtained. Based on the above fast and efficient algorithm, the ML or REML estimate is

$$\hat{\beta} = \hat{\lambda}_\beta X^T U_R \left( \frac{\eta_i}{\hat{\lambda}_{\beta} \delta_i + 1}, \eta_2, \ldots, \eta_{r} \right)^T$$

(24)

(Appendix E), where $\hat{\beta}$ denotes $\hat{\beta}_F$ or $\hat{\beta}_R$ while $\hat{\lambda}_\beta$ is $\hat{\lambda}_{\beta,F}$ or $\hat{\lambda}_{\beta,R}$, respectively.

**Likelihood ratio test (LRT).** Although the parameter on QTN in the model (1) is $\beta \sim N(0, \sigma_\beta^2)$, the estimation of $\hat{\lambda}_\beta$ is our concern in the above algorithm. Therefore, the null hypothesis might be $\lambda_\beta = 0$ [35]. The LRT statistic for the ML or REML estimate is

$$D = 2 \left( l(\hat{\lambda}_\beta) - l(0) \right)$$

(25)

where $l(\hat{\lambda}_\beta)$ is $l_F(\hat{\lambda}_{\beta,F})$ or $l_R(\hat{\lambda}_{\beta,R})$, and $l(0)$ is $l_F(0) = -n \frac{2 \pi}{n} \log(\frac{2 \pi}{n}) - n \frac{2}{n} \log \left( \sum_{i=2}^{\sigma^2/\sigma_i^2} \right)$ or
\( I_r(0) = -\frac{v}{2} \log\left(\frac{2\pi}{v}\right) - \frac{v}{2} \log\left(\sum_{i=1}^{v} \frac{1}{2} \eta_i^2\right) \) while \( \hat{\lambda}_p \) is \( \hat{\lambda}_p,F \) or \( \hat{\lambda}_p,R \), respectively.

Under the null hypothesis, the LRT statistic \( D \) follows approximately a mixture of two \( \chi^2 \) distributions with an equal weight, denoted by \( \frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2 \), where \( \chi_0^2 \) is just a fixed number of zero and \( \chi_1^2 \) is a \( \chi^2 \) distribution with one degree of freedom [35], and the P-values can be calculated accordingly. Let \( P \) be the P-value for each QTN, it was calculated using

\[
P = \begin{cases} 
1 & D = 0 \\
\frac{1}{2} \Pr(\chi_1^2 > D) & D > 0
\end{cases}
\]

(26)

**Kinship matrix.** Many methods for calculating kinship matrix \( K_{\text{mm}} \) from a large number of markers have been proposed, such as identical-by-state approach [2,3,7,26,46]. Here we adopted the method of Kang et al. [3]. Let \( S \) be an \( p \times m \) genotypic matrix with elements \( s_{oi} \in \{0, 0.5, 1\} \), the element of kinship matrix \( K \) is defined by

\[
k_{ij} = \begin{cases} 
1 & i = j \\
\frac{1}{p} \sum_{o=1}^{p} (s_{oi} \times s_{oj} + (1-s_{oi}) \times (1-s_{oj})) & j < i, i = 2, \ldots, m \\
k_{ji} & j > i
\end{cases}
\]

(27)

**Time complexity for the first step of FASTmrEMMA.** For the single-locus genome scan, FASTmrEMMA is involved in only one eigen decomposition at the beginning, and computational complexity is \( O(mn^2) \), where \( O \) is the big \( O \) notation. For each SNP tested, FASTmrEMMA effectively replaces the expensive additional eigen decomposition step in EMMA, and the computational complexity changes from \( O(mn^2) \) to \( O(n^3) \), because the nonzero eigenvalues of \( X'X \) are the same as those of \( X_1'X_1 \). After this, as in EMMA, each iteration of the optimization step requires inexpensive operations (computational complexity of \( O(n) \)) to evaluate both the first and second derivatives of target functions. Therefore, the overall time complexity for the first
step of FASTmrEMMA is $O(mn^2 + pn^2 + ptn)$, compared with $O(mn^2 + pmn^2 + ptn)$ for EMMA[4], where $t$ is the number of optimization iterations required for the Newton-Raphson method (quadratic rate of convergence).

In the GWAS, the number of SNPs is often 1000 times larger than the sample size, and actually most SNPs are not associated with the trait of interest. In this case, fitting all the genome markers in one model is not feasible. Once we delete these SNPs with zero effects, the reduced model is estimable. The described above can be considered as an initial screening step for FASTmrEMMA.

In the first step of FASTmrEMMA, a less stringent criterion for the initial stage screening was adopted, for example, all the SNPs with the $\leq 0.005$ P-values were selected to enter the next step for further evaluation. The majority of markers will be eliminated in the first step. Therefore, the number of markers left in the second stage analysis is often a small subset of all markers, say a few hundred or a few thousand at most, for example, no more than 600 significantly associated SNPs for the four traits in the $A.\ thaliana$ datasets [29].

In the multi-locus model, we proposed to use the EM empirical Bayes (EMEB) [28] because EMEB method is a random model approach in which each random marker effect is assigned an empirical distribution with a variance, and therefore is in accordance with treating a marker as a random effect in this study. The linear model is as followed:

$$\mathbf{y} = \mathbf{Wa} + \sum_{i=1}^{q} \mathbf{X}_i \beta_i + \mathbf{e}$$

(28)

where $\mathbf{y}$, $\mathbf{W}$, $\mathbf{a}$ and $\mathbf{e}$ are the same as model (1); $q$ is the number of the selected QTN in the first step of FASTmrEMMA; $\mathbf{X}_i$ and $\beta_i$ are an $n \times 1$ vector of marker genotypes and effect for the $i$th QTN, respectively. In the above model, polygenic background is not included, because all the potential QTN have been included in the model (28).

In the model (28), we adopt the normal prior for $\beta_i$, $P(\beta_i | \sigma^2) = N\left(0, \sigma^2_i\right)$, and the scaled inverse
$$\chi^2$$ prior for $$\sigma_i^2$$, $P(\sigma_i^2 | \tau, \omega) \propto (\sigma_i^2)^{(\tau+2)} \exp\left(-\frac{\omega}{2\sigma_i^2}\right)$, where we set $$(\tau, \omega) = (0,0)$$, which represents the Jeffreys’ prior, $P(\sigma_i^2 | \tau, \omega) = 1/\sigma_i^2$ [47]. The procedure for parameter estimation in EMEB [28] is as follows.

1) Initial-step: To initialize parameters with

$$\alpha = (W^T W)^{-1} W^T y$$

$$\sigma_i^2 = \frac{1}{n} (y - Wa)^T (y - Wa)$$

$$\sigma_i^2 = \left( X_i^T X_i \right)^T \left( X_i^T (y - Wa) \right)^2 + \left( X_i^T X_i \right)^2 \sigma_i^2$$

2) E-step: QTN effect can be predicted by

$$E(\beta_i) = \sigma_i^2 X_i^T V^{-1} (y - Wa)$$  (29)

where $V = \sum_{i=1}^{q} X_i X_i^T \sigma_i^2 + I \sigma_i^2$.

3) M-step: To update parameters $\sigma_i^2$, $\alpha$ and $\sigma_i^2$:

$$\sigma_i^2 = \frac{E(\beta_i^T \beta_i) + \omega}{\tau + 3}$$

$$\alpha = (W^T V^{-1} W)^{-1} W^T V^{-1} y$$  (30)

$$\sigma_i^2 = \frac{1}{n} (y - Wa)^T \left( y - Wa - \sum_{i=1}^{q} X_i E(\beta_i) \right)$$

where $E(\beta_i^T \beta_i) = E(\beta_i^T) E(\beta_i) + \text{tr} \left[ \text{var}(\beta_i) \right]$, $\text{var}(\beta_i) = I \sigma_i^2 - \sigma_i^2 X_i^T V^{-1} X_i \sigma_i^2$ and $$(\tau, \omega) = (0,0)$$.

Repeat E-step and M-step until convergence is satisfied.

Because the model is multi-locus in nature, Bonferroni correction is replaced by a less stringent selection criterion. Although the general 0.05 critical value may be used for the significance test, we decided to place a slightly more stringent criterion of LOD=3.0. The criterion is frequently adopted in linkage analysis and is the equivalent of $P = \Pr(\chi_i^2 > 3.0 \times 4.605) \approx 0.0002$, in which $\chi_i^2$ under the null hypothesis, follows a $\chi^2$ distribution with one degree of freedom.
Other five methods

E-BAYES (see the Additional Information) is an existing multi-locus Bayesian approach implemented by the SAS program[30], and was used as a gold standard for multi-locus model comparison.

EMMA, CMLM, ECMLM and SUPER (see the Additional Information) were used as gold standards for single-locus model comparisons. All the thresholds of P-values for significance tests were set at $0.05/p$, due to Bonferroni correction for multiple tests, where $p$ is the number of markers.

The Arabidopsis thaliana data

We analyzed the well-known A. thaliana datasets published by Atwell et al. [29]. Both phenotypes and genotypes were obtained from http://www.arabidopsis.usc.edu/. A total of 199 Arabidopsis lines and 216130 SNPs were used for analysis. Four flowering time traits (LD, SDV, 8W GH LN and 8W GH FT) with log-transformation were re-analyzed in this study. We excluded the individuals with missing phenotypes, non-polymorphic SNPs and SNPs with minor allele frequency (MAF) less than 10% and all the six methods (FASTmrEMMA, E-BAYES, SUPER, EMMA, CMLM and ECMLM) were used to analyze these four datasets. A total of approximate 180,000 SNPs for each trait were used to calculate the identity by state (IBS) matrix as the estimates of relatedness [3].

Simulation experiments

Three Monte Carlo simulation experiments were conducted to validate the new algorithms.

As described by Wang et al. [7] (2016), the SNP genotypes derived from the A. thaliana datasets [29] were also used to perform three simulation experiments. The purpose was to compare FASTmrEMMA with the single-locus model methods (SUPER, ECMLM, CMLM and EMMA) and the multi-locus model method (E-BAYES). In the first simulation experiment, 2000 SNPs on each chromosome were randomly sampled. As a result, all the SNPs between 11226256 and 12038776 bp on Chr. 1, between 5045828 and 6412875 bp on Chr. 2, between 1916588 and
3196442 bp on Chr. 3, between 2232796 and 3143893 bp on Chr. 4, and between 19999868 and 21039406 bp on Chr. 5 were used to conduct simulation studies. The sample size was 199, the number of lines from Atwell et al. [29]. Six QTNs were simulated and placed on the SNPs with allele frequencies of 0.30; their heritabilities of each effect size were set as 0.10, 0.05, 0.05, 0.15, 0.05 and 0.05, respectively; and their positions and effects are listed in Table S2. The residual variance was set at 10.0. The new phenotypes were simulated by the model: 
$$y = \mu + \sum_{i=1}^{6} x_i b_i + \epsilon,$$
where $\epsilon \sim MVN_n(0, 10 \times I_n)$. Each sample was analyzed by the above six methods. For each simulated QTN, we counted the samples in which the LOD statistic exceeded 3.0 for FASTmrEMMA, the P-value was less than 0.05 for E-BAYES, and the P-value $\leq 5e-6 (0.05/p)$ for the others. A detected QTN within 2 kb of the simulated QTN was considered a true QTN. The ratio of the number of such samples to the total number of replicates (1000) represented the empirical power of this QTN. The Type I error was calculated as the ratio of the number of false positive effects to the total number of zero effects considered in the full model. To measure the bias of QTN effect estimate, mean squared error (MSE) and mean absolute deviation (MAD) were defined as
$$MSE = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\beta}_i - \beta)^2, \quad MAD = \frac{1}{1000} \sum_{i=1}^{1000} |\hat{\beta}_i - \beta|$$
where $\hat{\beta}_i$ is the estimate of $\beta$ for each QTN in the $i$th sample. A method with a small MSE (or MAD) is generally more preferable than a method with a large MSE (or MAD).

To investigate the effect of polygenic background on FASTmrEMMA, polygenic effect was simulated in the second simulation experiment by multivariate normal distribution $MVN_n(0, \sigma_{pg}^2 K)$, where $\sigma_{pg}^2$ is polygenic variance, and $K$ is the kinship coefficient matrix between a pair of lines. Here $\sigma_{pg}^2 = 5$, so $h_{pg}^2 = 0.075$. The QTN size ($r^2$), residual variance, and others were the same as those in the second simulation experiment. The new phenotypes were simulated by the model: 
$$y = \mu + \sum_{i=1}^{6} x_i b_i + u + \epsilon,$$
where polygenic effect $u \sim MVN_n(0, 5 \times K)$ and $\epsilon \sim MVN_n(0, 10 \times I_n)$. To investigate the effect of epistatic background on FASTmrEMMA,
three epistatic QTN each with $\sigma^2_{\text{epi}} = 1.25$ and $h^2_{\text{epi}} = 0.05$ were simulated in the third simulation experiment. The first one was placed between 3063784 bp on Chr. 4 and 5227063 bp on Chr. 2; the second one was placed between 5986135 bp on Chr. 2 and 2031781 bp on Chr. 3; and the third one was placed between 2668059 bp on Chr. 3 and 11824678 bp on Chr. 1. The QTN size ($r^2$), residual variance and others were also the same as those in the first simulation experiment. The new phenotypes were simulated by the model: 

$$
\mathbf{y} = \mathbf{\mu} + \sum_{i=1}^{6} \mathbf{x}_i \mathbf{b}_i + \sum_{j=1}^{3} \left( \mathbf{A}_j \# \mathbf{B}_j \right) \mathbf{b}_{jj} + \mathbf{\varepsilon},
$$

where $\mathbf{\varepsilon} \sim \text{MVN}_n(\mathbf{0}, 10 \times \mathbf{I}_n)$, $b_{jj}$ is epistatic effect, and $A_j \# B_j$ is its incidence coefficient.

References

1. Zhang, Y. M. et al. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (Zea mays L.). *Genetics* 169, 2267–2275 (2005).
2. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38, 203–208 (2006).
3. Kang, H. M. et al. Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723 (2008).
4. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44, 821–824 (2012).
5. Li, G. X. & Zhu, H. J. Genetic Studies: the linear mixed models in genome-wide association studies. *The Open Bioinformatics Journal* 7, 27–33 (2013).
6. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291–295 (2015).
7. Wang, S. B. et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep* 6, 19444 (2016).
8. Yi, N. & Xu, S. Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179(2), 1045–1055 (2008).
9. Wang, D., Eskridge, K. M. & Crossa, J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J Agr Biol Envir St* 16, 170–184 (2010).
10. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet* 4, e1000130 (2008).
11. Ayers, K. L. & Cordell, H. J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiol* 34, 879–891 (2010).
12 Cho, S. et al. Joint identification of multiple genetic variants via Elastic-Net variable selection in a genome-wide association analysis. *Ann Hum Genet* **74**, 416–428 (2010).

13 Lü, H. Y., Liu, X. F., Wei, S. P. & Zhang, Y. M. Epistatic association mapping in homozygous crop cultivars. *PLoS ONE* **6**, e17773 (2011).

14 Wen, J. et al. Genetic dissection of heterosis using epistatic association mapping in a partial NCII mating design. *Sci Rep* **5**, 18376 (2015).

15 Zhou, X., Carbonetto, P. & Stephen, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet* **9**, e1003264 (2013).

16 Moser, G. et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* **11**, e1004969 (2015).

17 Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**, 825–830 (2012).

18 Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355–360 (2010).

19 Li, M. et al. Enrichment of statistical power for genome-wide association studies. *BMC Biology* **12**, 73 (2014).

20 Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348–354 (2010).

21 Lippert, C. et al. Fast linear mixed models for genome-wide association studies. *Nat Methods* **8**, 833–835 (2011).

22 Listgarten, J. et al. Improved linear mixed models for genome-wide association studies. *Nat Methods* **9**, 525–526 (2012).

23 Svisheeva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**, 1166–1170 (2012).

24 Wang, Q., Tian, F., Pan, Y., Buckler, E.S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS ONE* **9**, e107684 (2014).

25 Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating effects and making predictions from genome-wide marker data. *Stat Sci* **24**, 517–529 (2009).

26 Xu, S. Mapping quantitative trait loci by controlling polygenic background effects. *Genetics* **195**, 1209–1222 (2013).

27 Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284–290 (2015).

28 Xu, S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **105**, 483–494 (2010).
29 Atwell, S. et al. Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).

30 Xu, S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**, 513–521 (2007).

31 Tsepilov, Y. A. et al. Development and application of genomic control methods for genome-wide association studies using non-additive models. *PLoS ONE* **8**, e81431 (2013).

32 Schmid, M. et al. A gene expression map of *Arabidopsis thaliana* development. *Nat Genet* **37**, 501–506 (2005).

33 Stenvik, G. E. et al. The EPIP peptide of INFLORESCENCE DEFICIENT IN ABSCISSION is sufficient to induce abscission in *Arabidopsis* through the receptor-like kinases HAESA and HAESA-LIKE2. *The Plant Cell* **20**, 1805–1817 (2008).

34 Heyndrickx, K. S. & Vandepoele, K. Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiology* **159**, 884–901 (2012).

35 Wei, J. & Xu, S. A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. *Genetics* **202**, 471–486 (2016).

36 Wang, S. B. et al. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. *Sci Rep* **6**, 29951 (2016).

37 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).

38 Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

39 Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459–463 (2010).

40 Wang, S. G., Shi, J. H., Yin, S. J. & Wu, M. X. *An Introduction to Linear Models*. Beijing: Science Press, 2004.

41 Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameters estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).

42 Schott, J. R. *Matrix Analysis for Statistics*. New York: John Wiley & Sons, 1997.

43 Searle, S. R., Casella, G. & McCulloch, C. E. *Variance Components*. New York: Wiley, 2006.

44 Searle, S. R. & Quaas, R. L. *A detailed description of recent methods of estimating variance components, with applications in animal breeding*. Second Draft, 1978.

45 Demidenko, E. *Mixed Models: Theory and Applications with R*. Second Edition, Wiley, 2013.

46 Zhao, K. et al. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet* **3**, e4 (2007).
Figueiredo, M. A. T. Adaptive sparseness for supervised learning. *IEEE Trans Pattern Anal Mach Intell* **25**, 1151–1159 (2003).

**Acknowledgements**

This work was supported by the National Natural Science Foundation of China (grants 31571268 and 31301229), and Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2014RC020).

**Authors' contributions**

Y.-M.Z. conceived and supervised the study, and revised the manuscript. R.W. assisted the supervision of the experiments and revised the manuscript. Y.-J.W. deduced the formulas and wrote the draft. Y.-J.W., S.-B.W., H.Z., B.H., J.Z. and J.-Y.F. performed the experiments and analyzed the data. J.M.D. and H.Z. improved the language within the manuscript. All authors read, revised and approved the final manuscript.

**Additional Information**

**Competing Financial Interests:** The authors declare no competing financial interests.

**Supplementary information** accompanies this manuscript in the file entitled with “Additional Information”.


Figure 1 | Comparison of the QTN-variance estimates between fast single-locus random-SNP-effect EMMA (FASTsrEMMA) and one exact algorithm implemented by PROC MIXED in SAS. LD: days to flowering under long days; SDV: days to flowering under short days with vernalization; 8W GH LN: leaf number at flowering with 8 wks vernalization, greenhouse; and 8W GH FT: days to flowering, 8 wks vernalization, greenhouse.
Figure 2 | Comparison of fast multi-locus random-SNP-effect EMMA (FASTmrEMMA) with the single- and multi-locus approaches under various genetic backgrounds. The single-locus model approaches include SUPER, EMMA, CMLM and ECMLM, and the multi-locus approach has E-BAYES. The powers are presented in a-c, mean squared errors are showed in d-f, and mean absolute deviations are listed in g-i. Six QTN (a, d and g), six QTN plus polygenes (b, e and h), and six QTN plus three epistasis (c, f and i) were simulated, respectively, in the second to fourth simulation experiments.
Figure 3 | Statistical powers for six simulated QTN in the first simulation experiment plotted against type I error (in a log₁₀ scale) for the six GWAS methods (FASTmrEMMA, E-BAYES, SUPER, EMMA, CMLM and ECMLM).
| Trait          | FASTmrEMMA | E-BAYES | SUPER | EMMA  | ECMLM | CMLM  |
|---------------|------------|---------|-------|-------|-------|-------|
| LD            | 39.54      | 287.00  | 396.65| 299.97| 382.07| 382.07|
| SDV           | -88.09     | 43.20   | 179.54| 100.69| 169.87| 169.87|
| 8W GH LN      | -103.47    | 77.76   | 117.50| 117.50| 117.50| 117.50|
| 8W GH FT      | -321.72    | -155.55 | -82.41| -101.83| -82.41| -82.41|

LD: days to flowering under long days; SDV: days to flowering under short days with vernalization; 8W GH LN: leaf number at flowering with 8 wks vernalization, greenhouse; 8W GH FT: days to flowering, 8 wks vernalization, greenhouse.
| Trait | Gene | Chr | SNP(bp) | FASTmrEMMA LOD | Effect | MAF | $r^2$(%) | E-BAYES LOD | Effect | MAF | $r^2$(%) | SUPER P-value | Effect | MAF | $r^2$(%) | EMMA P-value | Effect | MAF | $r^2$(%) | References |
|-------|------|-----|---------|----------------|--------|-----|---------|-------------|--------|-----|---------|-------------|--------|-----|---------|-------------|--------|-----|---------|------------|
| LD    | At1g22770 | 1   | 8045438 | 4.872          | -0.112 | 0.395 | 0.549   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At1g23000 | 1   | 8128350 | 9.006          | -0.197 | 0.461 | 1.767   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At2g22540 | 2   | 9588685 | 10.338         | -0.330 | 0.281 | 4.034   | 10.753      | -0.611 | 0.281 | 13.817  | 2.78E-09    | -0.815 | 0.281 | 24.607  | Schmid et al. 2005 |
|       | At2g22610 | 2   | 9588685 | 10.338         | -0.330 | 0.281 | 4.034   | 10.753      | -0.611 | 0.281 | 13.817  | 2.78E-09    | -0.815 | 0.281 | 24.607  | Schmid et al. 2005 |
|       | At3g01970 | 3   | 22949227| 5.919          | 0.149  | 0.413 | 0.986   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At5g10140 | 5   | 3188328 | 12.759         | -0.272 | 0.263 | 2.630   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At4g00310 | 4   | 153459  |               |        |      |        |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At4g00335 | 4   | 167142  |               |        |      |        |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At4g00450 | 4   | 196614  |               |        |      |        |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At4g01280 | 4   | 516758  |               |        |      |        |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
| SDV   | At1g05440 | 1   | 1595585 | 4.298          | 0.117  | 0.214 | 1.346   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At1g05470 | 1   | 1595585 | 4.298          | 0.117  | 0.214 | 1.346   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At1g77080 | 1   | 28965510| 10.817         | -0.177 | 0.484 | 4.576   | 4.020       | -0.170 | 0.484 | 4.221  |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At2g41890 | 2   | 17488070| 4.339          | 0.099  | 0.302 | 1.208   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At3g20260 | 3   | 7084425 | 3.309          | 0.068  | 0.302 | 0.570   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At3g49600 | 3   | 18385143| 4.529          | 0.118  | 0.321 | 1.774   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At4g05420 | 4   | 2748735 | 4.286          | -0.091 | 0.459 | 1.203   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At5g04240 | 5   | 1164843 | 4.479          | -0.137 | 0.220 | 1.884   |             |        |     |        |             |        |     |        |             |        |     |        | Schmid et al. 2005 |
|       | At5g09805 | 5   | 3055565 | 4.763          | -0.105 | 0.233 | 1.151   |             |        |     |        |             |        |     |        |             |        |     |        | Stenvik et al. 2008 |
| Gene   | Chromosome | Position   | Log2FC | Log10P | P-value | MAF     | Significance |
|--------|------------|------------|--------|--------|---------|---------|-------------|
| At5g57360 | 5          | 23249199   | 5.419  | -0.141 | 0.321   | 2.533   | Schmied et al. 2005 |
| At5g57390 | 5          | 23249199   | 5.419  | -0.141 | 0.321   | 2.533   | Schmied et al. 2005 |
| At5g46880 | 5          | 19044037   | 3.55E-08 | -0.141 | 0.321   | 2.533   | 3.55E-08 0.408 0.107 9.296 Schmied et al. 2005 |
| At5g61700 | 5          | 26794176   | 3.55E-08 | -0.141 | 0.321   | 2.533   | 1.79E-07 -0.292 0.321 10.864 Schmied et al. 2005 |
| At5g62790 | 5          | 26794176   | 3.55E-08 | -0.141 | 0.321   | 2.533   | 1.79E-07 -0.292 0.321 10.864 Heyndrickx et al. 2012 |

8W GH LN
| Gene   | Chromosome | Position   | Log2FC | Log10P | P-value | MAF     | Significance |
|--------|------------|------------|--------|--------|---------|---------|-------------|
| At1g77080 | 1          | 28965510   | 3.857  | -0.109 | 0.497   | 2.610   | Schmied et al. 2005 |
| At2g27380 | 2          | 11703876   | 9.631  | -0.153 | 0.325   | 4.514   | Heyndrickx et al. 2012 |
| At4g32980 | 4          | 15918498   | 4.651  | -0.147 | 0.147   | 2.384   | Schmied et al. 2005 |
| At5g15850 | 5          | 5196549    | 5.923  | -0.106 | 0.319   | 2.145   | Schmied et al. 2005 |
| At5g45890 | 5          | 18600041   | 4.608  | -0.107 | 0.423   | 2.456   | Schmied et al. 2005 |

8W GH FT
| Gene   | Chromosome | Position   | Log2FC | Log10P | P-value | MAF     | Significance |
|--------|------------|------------|--------|--------|---------|---------|-------------|
| At1g04357 | 1          | 8612791    | 5.055  | 0.040  | 0.460   | 1.199   | Schmied et al. 2005 |
| At2g27380 | 2          | 11703876   | 4.744  | -0.043 | 0.323   | 1.122   | Heyndrickx et al. 2012 |
| At2g47230 | 2          | 19396129   | 4.208  | -0.038 | 0.298   | 0.911   | Schmied et al. 2005 |
| At3g56900 | 3          | 21079518   | 3.081  | -0.032 | 0.311   | 0.661   | Schmied et al. 2005 |
| At3g57000 | 3          | 21079518   | 3.081  | -0.032 | 0.311   | 0.661   | Schmied et al. 2005 |
| At5g06550 | 5          | 2002341    | 3.169  | -0.070 | 0.186   | 2.241   | Schmied et al. 2005 |
| At5g06590 | 5          | 2002341    | 3.169  | -0.070 | 0.186   | 2.241   | Heyndrickx et al. 2012 |

*MAF: minor allele frequency. The individuals with missing phenotypes and the SNPs with MAF ≤10% were excluded. The critical value for significance was LOD ≥3.0 for FASTmrEMMA and E-BAYES, and approximately 2.8E-07 P-value for SUPER, EMMA, CMLM and ECMLM. The results from CMLM and ECMLM were not listed in this table because no genes were detected. The dataset was derived from Atwell et al. (2010).*