Open access • Posted Content • DOI:10.1101/341784

A fast mrMLM algorithm for multi-locus genome-wide association studies — Source link

Cox Lwaka Tamba, Cox Lwaka Tamba, Yuan-Ming Zhang

Institutions: Egerton University, Huazhong Agricultural University

Published on: 07 Jun 2018 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: False positive rate

A fast mrMLM algorithm for multi-locus genome-wide association studies

Cox Lwaka Tamba[1, 2] and Yuan-Ming Zhang[1 *]

[1]Statistical Genomics Lab, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China; [2]Department of Mathematics, Egerton University, Egerton 536-20115, Kenya

Correspondence: soyzhang@mail.hzau.edu.cn

Abstract

Background: Recent developments in technology result in the generation of big data. In genome-wide association studies (GWAS), we can get tens of million SNPs that need to be tested for association with a trait of interest. Indeed, this poses a great computational challenge. There is a need for developing fast algorithms in GWAS methodologies. These algorithms must ensure high power in QTN detection, high accuracy in QTN estimation and low false positive rate.

Results: Here, we accelerated mrMLM algorithm by using GEMMA idea, matrix transformations and identities. The target functions and derivatives in vector/matrix forms for each marker scanning are transformed into some simple forms that are easy and efficient to evaluate during each optimization step. All potentially associated QTNs with P-values ≤ 0.01 are evaluated in a multi-locus model by LARS algorithm and/or EM-Empirical Bayes. We call the algorithm FASTmrMLM. Numerical simulation studies and real data analysis validated the FASTmrMLM. FASTmrMLM reduces the running time in mrMLM by more than 50%. FASTmrMLM also shows high statistical power in QTN detection, high accuracy in QTN estimation and low false positive rate as compared to GEMMA, FarmCPU and mrMLM. Real data analysis shows that FASTmrMLM was able to detect more previously reported genes than all the other methods: GEMMA/EMMA, FarmCPU and mrMLM.

Conclusions: FASTmrMLM is a fast and reliable algorithm in multi-locus GWAS and ensures high statistical power, high accuracy of estimates and low false positive rate.

Keywords: genome-wide association study; mixed linear model; multi-locus model; FarmCPU; mrMLM

Author Summary

The current developments in technology result in the generation of a vast amount of data. In genome-wide association studies, we can get tens of million markers that need to be tested for association with a trait of interest. Due to the computational challenge faced, we developed a fast algorithm for genome-wide association studies. Our approach is a two stage method. In the first step, we used matrix transformations and identities to quicken the testing of each random marker effect. The target functions and derivatives which are in vector/matrix forms for each marker scanning are transformed into some simple forms that are easy and efficient to evaluate during each optimization step. In the second step, we selected all potentially associated SNPs and evaluated them in a multi-locus model. From simulation studies, our algorithm significantly reduces the computing time. The new method also shows high statistical power in detecting significant markers, high accuracy in marker effect estimation and low false positive rate. We also used the new method to identify relevant genes in real data analysis. We recommend our approach as a fast and reliable method for carrying out a multi-locus genome-wide association study.

Background

Genome-wide association studies (GWAS) aim at investigating the genetic foundation of complex traits by focusing on the relationship between molecular markers and these traits [1, 2]. Initially, each accession was genotyped by SSR markers, and only hundreds of SSR markers were used to conduct GWAS in plants [3]. When restriction association site DNA sequencing (RAD-seq), specific-locus amplified fragment sequencing (SLAF-seq) and gene chip technology were adopted, then, there were tens of thousands single nucleotide polymorphisms (SNPs) available. When deep sequencing is implemented, millions of SNPs are obtained. As third generation sequencing technology generates, we can get tens of millions SNPs. Evidently, severe computational challenges are faced. Accordingly, there is a critical need for in-depth study of fast algorithm in GWAS methodologies.

Mixed linear model (MLM) method of GWAS was firstly established by Zhang et al. [4]. At each putative QTN scan in this approach, the pedigree-based coancestry matrices of QTN and polygenes are incorporated into the mixed linear model framework to estimate three variance components of QTN, polygenes and residual error. Its long running time makes this method unfashionable. Yu et al. [5] replaced the pedigree-based coancestry matrix with kinship matrix $(\mathbf{K})$ to define the degree of genetic covariance between pairs of individuals and view QTN effect as fixed. This method was improved by the spectral decomposition of $\mathbf{H} = \mathbf{ZKZ}' + \delta\mathbf{I}$ and joint maximum likelihood estimation of fixed effect $\boldsymbol{\beta}$ and σ_e^2, where $\mathbf{Z}$ is an incidence matrix mapping each observed phenotype to each inbred strain, $\mathbf{I}$ is an identity matrix and δ is the ratio of residual $\left(\sigma_e^2\right)$ to polygenic $\left(\sigma_g^2\right)$ variances. This is efficient mixed model association (EMMA) of Kang et al. [6]. Several approaches have been considered with the aim to reduce the computing time and increase power in QTN detection. Zhang et al. [7] suggested 'population parameters previously determined' (P3D) where individuals in polygenic effect are replaced by their corresponding groups, and kinship among individuals is replaced by the kinship among groups. The P3D eliminates the need to re-compute variance components. Kang et al. [8] also fixed the above δ in EMMA eXpedited (EMMAX). Li et al. [9] optimized the combination of kinship algorithms and grouping algorithms. Also, other authors suggested alternatives, such as FaST-LMM [10], FaST-LMM-Select [11], GRAMMAR-Gamma [12] and SUPER [13]. In genome-wide EMMA (GEMMA) of Zhou & Stephens [14], especially, target functions and derivatives in vector/matrix forms for each marker, despite their complicated appearance, are easy and efficient to evaluate during each optimization step. Among the above fast methods, the SNP effect was treated as being fixed. Although a random marker model has several advantages [15, 16], an efficient computational algorithm to estimate SNP effect needs to be addressed.

Multi-locus model has become the state-of-the-art GWAS procedure. Previous studies have reported higher power of QTN detection in multi-locus models as compared to the single-marker GWAS analysis [17, 16]. Bonferroni multiple test correction in the single-marker analysis is replaced by a less stringent selection criterion in multi-locus GWAS analysis [16, 18]. Therefore significant loci for complex traits are not missed out. Initially, many statistical approaches were used to estimate all the SNP effects in multi-locus model, such as Bayesian LASSO [19], penalized logistic regression [20, 21], adaptive mixed LASSO [22], elastic net [23], empirical Bayes [24], and empirical Bayes LASSO [25]. If the number of SNPs is large, the above methods are not feasible, even through the Bayesian sparse linear mixed model of Zhou et al. [26] and Bayesian mixture model of Moser et al. [27]. To overcome this shortcoming, Segura et al. [17] proposed multi-locus mixed-model (MLMM), which is a simple, stepwise mixed-model regression with forward inclusion and backward elimination. In FarmCPU of Liu et al. [28], the fixed effect model and random effect model are used iteratively until a stage of convergence is reached. The fixed effect model contains the testing marker and pseudo QTNs to control false positives. The pseudo QTNs are selected from associated markers and evaluated by the random effect model, with κ defined by the pseudo-QTNs. However, the computationally intensive forward-backward inclusion of SNPs is clearly a limiting factor in exploring the huge model space. Recently, we have proposed several multi-locus two-stage GWAS approaches, such as mrMLM [16], FASTmrEMMA [18] and ISIS EM-BLASSO [29]. In the first stage, single-locus methods are used to scan all the markers on the genome. In the second stage, a few SNPs potentially associated with the trait are selected and placed into the multi-locus model, and all the effects in the model are estimated by empirical Bayes for true QTN detection. Among our three methods, it is possible to quicken mrMLM.

In this study, we accelerated mrMLM algorithm of Wang et al. [16] using the GEMMA idea and matrix transformation of Miller [30]. In other words, target functions and derivatives in vector/matrix forms for each marker are transformed into some simple forms that are easy and efficient to evaluate during each optimization step. We call this method FASTmrMLM. A series of Monte Carlo simulation experiments and real data analyses were used to validate FASTmrMLM. FASTmrMLM significantly reduces the running time of the mrMLM algorithm. High power and low false positive rate (FPR) in QTN detection and high accuracy in QTN effect estimation were also observed as compared to GEMMA, FarmCPU and mrMLM.

Results

Computational efficiency

To confirm the effectiveness of FASTmrMLM, a series of Monte Carlo simulation experiments were carried out. Each sample was analyzed by five methods: FASTmrMLM, mrMLM, FarmCPU, GEMMA, and EMMA. In the first Monte Carlo simulation experiment where no polygenic variance was simulated, the running times (Intel Core i5-4570 CPU 3.20GHz, Memory 7.88G) for the above five methods are 6.25, 13.77, 5.12, 2.57 and 68.77 (hours), respectively (Fig 1 and Table 1). FASTmrMLM takes less than 50% of the running time needed when mrMLM is used. The same trend is observed in the rest of the simulations (Fig 1 and Table 1). It indicates that FASTmrMLM significantly quickens mrMLM. Although GEMMA and FarmCPU have lower computational time than FASTmrMLM, their performances in statistical power and parameter estimation accuracy are worse than those of FASTmrMLM.

Table 1. Comparison of the time taken (Hrs) in the detection of QTNs in three simulation experiments using FASTmrMLM, GEMMA/EMMA, FarmCPU and mrMLM methods (199 individuals each with 10000 SNPs, 1000 replicates)

Simulation	FASTmrMLM	mrMLM	FarmCPU	GEMMA	EMMA
I	6.25	13.77	5.12	2.57	68.77
II	6.23	13.25	4.95	2.84	68.25
III	6.28	13.93	5.08	2.59	68.93

Statistical power

Statistical power was used to evaluate the effectiveness of FASTmrMLM when compared with the other four methods. All the statistical powers in the three simulation experiments are presented in Fig 2 and Table S1. In the first simulation experiment (Fig 2a and Table S1), the average powers across six simulated QTN from FASTmrMLM, mrMLM, FarmCPU, GEMMA and EMMA were 68.8, 68.6, 41.9, 46.0 and 46.0 (%), respectively. It means that FASTmrMLM has the highest power in QTN detection. When paired t-test was conducted between FASTmrMLM and the other methods, FASTmrMLM has significantly higher power than FarmCPU, GEMMA, and EMMA (P-value = 0.004 ~ 0.012). Although there is no significant difference between FASTmrMLM and mrMLM (P-value = 0.688), FASTmrMLM has slightly higher power than mrMLM (Table S2). Therefore, FASTmrMLM is the most effective method for QTN detection.

Mean squared errors of estimated QTN effects

Mean squared error (MSE) was used to measure the accuracy of each estimated QTN effect for all the five methods in the three simulation experiments. All the MSE values for all the six simulated QTN effects in all the three simulation experiments are shown in Fig 3 and Table S3, and the average value for each simulated QTN effect in all the three simulation experiments is listed in Table S4. In the first simulation experiment (Fig 3a and Table S3), average MSE values across six simulated QTN from FASTmrMLM, mrMLM, FarmCPU, GEMMA, and EMMA were 0.0775, 0.0933, 0.2824, 0.5467 and 0.5432, respectively. It means that FASTmrMLM has the highest accuracy in the estimation of QTN effect. When paired t-test was conducted between FASTmrMLM and the other four methods, the MSE value was at least significantly lower from FASTmrMLM than from GEMMA and EMMA (P-value = 0.009 ~ 0.020). Although there is no significant difference between FASTmrMLM and the other two (mrMLM and FarmCPU) methods (P-value = 0.110 ~ 0.806), FASTmrMLM has slightly lower MSE than mrMLM and FarmCPU (Table S2). Therefore, FASTmrMLM has the highest accuracy in the estimation of QTN effect.

False positive rate

All the single-locus GWAS approaches are involved in multiple test issue so that Bonferroni correction is frequently used to control false positive rate (FPR) or empirical type 1 errors. In the current multi-locus GWAS method, a less stringent selection criterion was used to identify true QTN. In this situation, it is important to show the FPR values in the three simulation experiments. All the FPR values are presented in Fig 4 and Table 2. In the first simulation experiment (Fig 4a and Table 2), the FPR values for the above five methods were

1.80E-2, 1.99E-2, 1.78E-2, 3.25E-2 and 3.25E-2 (%), respectively (Table 2). It indicates that FASTmrMLM has almost the lowest FPR, although a less stringent selection criterion is adopted. Therefore, FASTmrMLM has controlled the FPR in QTN detection.

Table 2. Comparison of false positive rate (‰) in the detection of QTNs in three simulation experiments using FASTmrMLM, mrMLM, FarmCPU, and GEMMA/EMMA methods

Simulation	FASTmrMLM	mrMLM	FarmCPU	GEMMA	EMMA
I	0.1799	0.1990	0.1780	0.3250	0.3250
II	0.2276	0.2340	0.1740	0.1660	0.1660
III	0.2611	0.2660	0.2310	0.2530	0.2530

Receiver operating characteristic curve

Receiver operating characteristic (ROC) curve is obtained when statistical power is plotted against controlled Type 1 error. This curve is used to compare the efficiencies of different methods in the detection of significant effects. A method is considered the best if its ROC curve lies above all the other curves of the methods being compared. We simulated 100 various probability levels of significance between 1E-8 to 1E-2. We calculated statistical powers corresponding to these levels in the first simulation experiment. Fig 5 depicts the ROC curves for the four methods for each simulated QTN in the first experiment. As a result, FASTmrMLM performs the best for the simulated QTNs 1, 2, 4 and 6; mrMLM has the highest curve for QTNs 3 and 5 though its curves lie slightly above ROC curves of FASTmrMLM.

Real data analysis

We used FASTMRMLM, mrMLM, FarmCPU, and GEMMA/EMMA methods to re-analyze six Arabidopsis flowering time traits (LD, LDV, SD, 0W, 2W, and 4W) in Atwell et al. [31].

FASTmrMLM identified 17, 15, 14, 17, 14 and 15 SNPs to be significantly associated respectively with the six traits above. The identified SNPs for each trait were used to conduct a multiple linear regression analysis, and we calculated the corresponding AIC and BIC values. Table S5 shows the AIC and BIC values for all the methods in all the six traits. FASTmrMLM has low AIC and BIC values for nearly all the traits. mrMLM compares almost equally with FASTmrMLM. This indicates that SNPs found to be significant by FASTmrMLM fit the data better than the other methods.

We obtained 14, 10, 7, 10, 10, and 11 known genes in the proximity of detected SNPs when we used FASTmrMLM method to analyze the six traits respectively. The mrMLM method identified 4, 5, 1, 2, 5, and 4 known genes respectively for the above six traits. Known genes identified by FarmCPU for the six traits were 2, 3, 2, 0, 0, 4 and 2 respectively. GEMMA/EMMA was able to determine 5, 2, 0, 0, 3, and 3 known genes for the above six traits. FASTmrMLM detected more known genes than all the other methods (Table S6).

The new method was able to detect 7, 7, 6, 8, 7 and 5 new genes for the corresponding six traits considered in this study (Table S7). GEMMA/EMMA was not able to detect any gene for the traits SD and 0W. Indeed the Bonferroni correction is so stringent and may make significant genes to be missed it. The same observations are made when FarmCPU is used to analyze the 0W trait, and no gene is identified. From the results obtained in this study, we observe that the new approach can detect more associated genes when used in GWAS study. FASTmrMLM does reduce the computational time as well as ensures that the associated genes are not missed out. Based on these findings, we note that *Arabidopsis thaliana* GWAS results obtained in this study are reliable.

Discussion

FASTmrMLM in this study is different from mrMLM of Wang et al. [16] in two aspects. First, computations during each optimization stage in the first stage of FASTmrMLM are quicker than those in mrMLM. This is because FASTmrMLM uses simplified analytic forms of gradient functions and Hessian matrices, and mrMLM uses numerical gradients and Hessian matrices during optimization stage. The simplified analytic forms in target functions and derivatives quicken computation during each optimization stage of FASTmrMLM. Our results confirm the significant reduction in running time in FASTmrEMMA. Second, the LARS algorithm of Efron et al. [32] is used in FASTmrMLM to select the $n-1$ variables that are most likely associated with the quantitative trait of interest if the number of markers passing the 1% level of significance is more than n in the second stage of FASTmrMLM. The $n-1$ selected markers are then included in a multi-locus model and detected by EM-Empirical Bayes for true QTN identification. This can explain why slight improvements in statistical power and accuracy are observed in this study. In this study we also confirmed why we used EM-Empirical Bayes to estimate QTN effects in the multi-locus model. Here we compared three methods: EM-Empirical Bayes, adaptive Lasso, and SCAD. As a result, EM-Empirical Bayes has higher statistical power and more accurate than SCAD and adaptive Lasso (Tables S8 to S10).

In the past ten years, many approaches have been used to reduce running time in GWAS. First, QTN effect is viewed as fixed [5, 7, 8] rather than as previously viewed as random [4]. As such, the number of variance components decreases from three (QTN, polygenic and residual error) to two (polygenic and residual error). Second, the polygenic-to-residual variance ratio, obtained at null hypothesis, is fixed in the single-marker genome scan [7, 8]. In this case, the number of variance components decreases further from two (polygenic and residual error) to one (residual error). Finally, some matrix transformations and identities are adopted. One such matrix transformation is spectral decomposition, which lets target functions and derivatives be expressed by simple forms, such as in Kang et al. [6] and Wen et al. [18]. The first stage of the new method (FASTmrMLM) considers a QTN effect as random. Then, the polygenic-to-residual variance ratio obtained at null hypothesis is fixed in the single-marker genome testing. With a simple matrix transformation, the matrix $\mathbf{H}$ turns to be a sum of two matrices (an identity matrix ($\mathbf{I}$) and a matrix of rank one ($\mathbf{Z}_{ci}\mathbf{Z}_{ci}'\omega_i$)). More importantly, the results in Miller [30] are then used to compute quickly $\mathbf{H}^{-1}$, $|\mathbf{H}|$ and quadratic terms in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$, which frequently appear in the target functions, derivatives and parameter optimization in Newton-Raphson iterations.

In the single-locus scanning step of FASTmrMLM, all estimates are based on restricted maximum likelihood estimation (REML). This is because FASTmrMLM is an extension of mrMLM, which only considers REML estimates [16]. Of course, we preferred REML over maximum likelihood estimation because it produces unbiased estimates of the variance components by taking into account the degrees of freedom that result from evaluating the fixed effects.

As showed in this study, FASTmrMLM is better than GEMMA. The possible reasons are described below. First, FASTmrMLM considers QTN effect as random rather than as fixed as in GEMMA of Zhou & Stephens [14]. This confirms the advantages of a random marker model as outlined in Goddard et al. [15]. Then, the significance level for each test is LOD=3 in FASTmrMLM rather than 0.05/*m* in GEMMA. In theory, a less stringent selection criterion simultaneously increases statistical power and FPR. However, our new method increases not only statistical power but also FPR. Finally, FASTmrMLM is a multi-locus model method while GEMMA is a single-locus model approach. This can explain why various significance levels are adopted between FASTmrMLM and GEMMA.

Conclusion

We accelerated our previous multi-locus GWAS method: mrMLM, with the help of a new matrix transformation and matrix identities. As a result, the computational time of estimating variance components in the first step is significantly reduced. We implemented the LARS algorithm of Efron et al. [32] between the first step and EM-Empirical Bayes estimation in the second step. This makes slight improvements in statistical power and estimation accuracy as compared to mrMLM. We confirmed that EM-Empirical Bayes is the best method for the estimation of parameters in the multi-locus model. The proposed method, named FASTmrMLM, significantly increased the statistical power and decreased FPR compared with other methods: FarmCPU, GEMMA and EMMA. In real data analysis, more previously reported genes were detected by FASTmrMLM.

Materials and Methods

Genetic model

We consider a mixed linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}_i\beta_i + \boldsymbol{\phi} + \boldsymbol{\varepsilon} \tag{1}$$

with $\mathbf{y}$ being a $n \times 1$ phenotypic vector of quantitative trait of n individuals, $\mathbf{X}$ is an $n \times q$ incident matrix of fixed effects $\boldsymbol{\alpha}$ including the overall mean, $\mathbf{Z}_\mathbf{i}$ is an $n \times 1$ vector of the ith SNP, β_i is a random effect of the ith marker, it is assumed to be a normal distribution with zero mean and each marker prior variance σ_i^2, $\boldsymbol{\phi} \sim \text{MVN}\left(\mathbf{0}, \mathbf{K}\sigma_g^2\right)$ is the polygenic effect with a multivariate normal distribution with zero mean and variance σ_g^2 described by a covariance matrix $\mathbf{K}$ (kinship matrix), and $\boldsymbol{\varepsilon} \sim \text{MVN}\left(\mathbf{0}, \mathbf{I}\sigma_e^2\right)$ is the residual error $\mathbf{I}$ with being $n \times n$ identity matrix. In this study, the kinship matrix $\mathbf{K}$ is marker inferred kinship [33] defined as $\mathbf{K} = \sum_{i=1}^{m} \mathbf{Z}_i\mathbf{Z}_i' \Big/ m$.

From equation (1) we have that,

$$\begin{aligned} \mathrm{E}(\mathbf{y}) &= \mathbf{X}\boldsymbol{\alpha} \\ \mathrm{Var}(\mathbf{y}) &= \mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 + \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2 = \sigma_e^2\left(\mathbf{Z}_i\mathbf{Z}_i'\omega_i + \mathbf{K}\omega_g + \mathbf{I}\right) = \sigma_e^2\mathbf{H} \end{aligned} \tag{2}$$

where $\omega_i = \sigma_i^2 / \sigma_e^2$, $\omega_g = \sigma_g^2 / \sigma_e^2$ are the variance ratios and $\mathbf{H} = \mathbf{Z}_i\mathbf{Z}_i'\omega_i + \mathbf{K}\omega_g + \mathbf{I}$. In this model, we have three random effects: the ith marker effect, polygenic effect, and residual error. Under the pure polygenic model, we have that $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\alpha}$ and $\mathrm{Var}(\mathbf{y}) = \sigma_e^2\left(\mathbf{K}\omega_g + \mathbf{I}\right)$. We pre-estimate the value of $\omega_g = \sigma_g^2 / \sigma_e^2$ under the pure polygenic model and fix it when testing each SNP effect in the genome-wide scanning [7]. Using spectral decomposition, we can find a diagonal matrix $\mathbf{D} = \mathrm{diag}\left(\delta_1, \mathrm{L}\ , \delta_m\right)$ and a matrix $\mathbf{U}$ such that $\mathbf{K} = \mathbf{UDU}'$. Notice that spectral decomposition can be performed on $\mathbf{K}$ since as defined it is a square, symmetric matrix. Transforming $\mathbf{y}$ in Equation (1) by multiplying by $\mathbf{U}'$, we have,

$$\mathbf{U}'\mathbf{y} = \mathbf{U}'\mathbf{X}\boldsymbol{\alpha} + \mathbf{U}'\mathbf{Z}_i\beta_i + \mathbf{U}'\left(\boldsymbol{\phi} + \boldsymbol{\varepsilon}\right) \tag{3}$$

Let $\mathbf{y}^* = \mathbf{U}'\mathbf{y}, \mathbf{X}^* = \mathbf{U}'\mathbf{X}, \mathbf{Z}_i^* = \mathbf{U}'\mathbf{Z}_i$, then Equation (3) becomes

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\alpha} + \mathbf{Z}_i^*\beta_i + \mathbf{U}'\left(\boldsymbol{\phi} + \boldsymbol{\varepsilon}\right)$$

It follows that,

$$\begin{aligned} \mathrm{Var}(\mathbf{y}^*) &= \mathbf{Z}_i^*\mathbf{Z}_i^{*\prime}\sigma_i^2 + \mathbf{U}'\mathbf{K}\mathbf{U}\sigma_g^2 + \mathbf{U}'\mathbf{I}\mathbf{U}\sigma_e^2 = \sigma_e^2\left(\mathbf{Z}_i^*\mathbf{Z}_i^{*\prime}\omega_i + \mathbf{D}\hat{\omega}_g + \mathbf{I}\right) \\ &= \sigma_e^2\left(\mathbf{Z}_i^*\mathbf{Z}_i^{*\prime}\omega_i + \mathrm{diag}\left(\delta_1\hat{\omega}_g + 1, \delta_2\hat{\omega}_g + 1, \mathrm{L}\ , \delta_m\hat{\omega}_g + 1\right)\right) \end{aligned} \tag{4}$$

For simplicity let $\mathbf{D}_1 = \mathrm{diag}\left(\delta_1\hat{\omega}_g + 1, \delta_2\hat{\omega}_g + 1, \mathrm{L}\ , \delta_m\hat{\omega}_g + 1\right)$. $\hat{\omega}_g$ is fixed and $\mathbf{D}_1$ as defined is a positive semi-definite matrix. Therefore, we can obtain $\mathbf{D}_1^{1/2}$. Further transforming $\mathbf{y}^*$ by multiplying equation (3) through by $\mathbf{D}_1^{-1/2}$ and letting $\mathbf{y}_c = \mathbf{D}_1^{-1/2}\mathbf{y}^*, \mathbf{X}_c = \mathbf{D}_1^{-1/2}\mathbf{X}^*, \mathbf{Z}_c = \mathbf{D}_1^{-1/2}\mathbf{Z}_i^*$ and $\left(\boldsymbol{\phi} + \boldsymbol{\varepsilon}\right)_c = \mathbf{D}_1^{-1/2}\mathbf{U}'\left(\boldsymbol{\phi} + \boldsymbol{\varepsilon}\right)$. We have that,

$$\begin{aligned}\mathbf{D}_1^{-1/2}\mathbf{y}^* &= \mathbf{D}_1^{-1/2}\mathbf{X}^*\boldsymbol{\alpha} + \mathbf{D}_1^{-1/2}\mathbf{Z}_i^*\beta_i + \mathbf{D}_1^{-1/2}\mathbf{U}'\left(\boldsymbol{\phi}+\boldsymbol{\varepsilon}\right)\\ \mathbf{y}_c &= \mathbf{X}_c\boldsymbol{\alpha} + \mathbf{Z}_{ci}\beta_i + \left(\boldsymbol{\phi}+\boldsymbol{\varepsilon}\right)_c\end{aligned} \qquad (5)$$

Notice that the transformation in Equation (5) is equivalent to multiplying the original $\mathbf{y}$ by $\mathbf{U}'\mathbf{D}_1^{-\frac{1}{2}}$ i.e. $\mathbf{y}_c = \mathbf{U}'\mathbf{D}_1^{-1/2}\left(\mathbf{y}\right)$. Now,

$$\begin{aligned}&\mathrm{E}\left(\mathbf{y}_c\right) = \mathbf{X}_c\boldsymbol{\alpha}\\ &\mathrm{Var}\left(\mathbf{y}_c\right) = \sigma_e^2\left(\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\omega_i + \mathbf{I}\right) = \sigma_e^2\left(\mathbf{A}+\mathbf{B}\right) = \sigma_e^2\mathbf{H}\end{aligned} \qquad (6)$$

with $\mathbf{A} = \mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\omega_i$, $\mathbf{B} = \mathbf{I}$, and $\mathbf{H} = \mathbf{A}+\mathbf{B}$. Thus the distribution of our transformed data vector $\mathbf{y}_c$ is normal with the mean $\mathbf{X}_c\boldsymbol{\alpha}$ and variance-covariance matrix, $\sigma_e^2\mathbf{H}$. The parameters to be estimated are σ_e^2 and ω_i for $i = 1, 2, \mathrm{L}\;, \mathrm{m}$.

Parameter estimation

The profiled residual log likelihood (REML) of $\mathbf{y}^*$ after absorbing other terms in the constant term C is,

$$\mathrm{l}_R = C - \frac{1}{2}\left\{(n-q)\log\sigma_e^2 + \log\left|\mathbf{H}\right| + \log\left|\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right| + \mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c/\sigma_e^2\right\} \qquad (7)$$

where, $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}_c\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)^{-1}\mathbf{X}_c{}'\mathbf{H}^{-1}$, and $q = \mathrm{rank}\left(\mathbf{X}_c\right)$. We differentiate Equation (7) to obtain REML estimates as shown below.

Estimation of residual variance σ_e^2: We have

$$\frac{\partial \mathrm{l}_R}{\partial\sigma_e^2} = -\frac{1}{2}\left[\frac{n-q}{\sigma_e^2} - \frac{\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c}{\left(\sigma_e^2\right)^2}\right] \quad \Rightarrow \quad \hat{\sigma}_e^2 = \frac{\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c}{n-q}$$

This can be simplified to

$$\hat{\sigma}_e^2 = \frac{\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c}{n-q} = \frac{1}{n-q}\left[\left(\mathbf{y}_c{}'\mathbf{H}^{-1}\mathbf{y}_c\right) - \left(\mathbf{y}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)^{-1}\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{y}_c\right)\right] \qquad (8)$$

Estimation of the variance ratio ω_i: We have that,

$$\frac{\partial \mathrm{l}_R}{\partial \omega_i} = -\frac{1}{2}\left[\mathrm{tr}(\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}') - \frac{1}{\hat{\sigma}_e^2}\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}'_{ci}\mathbf{P}\mathbf{y}_c \right] \tag{9}$$

The second derivative of the residual log likelihood function with respect to ω_i is given below, and it is used to obtain the Hessian matrix.

$$\frac{\partial^2 \mathrm{l}_R}{\partial \omega_i^2} = -\frac{1}{2}\left\{ -\mathrm{tr}\left(\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\right) - (n-q)\frac{2\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{y}_c\right)\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c\right) - \left(\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{y}_c\right)^2}{\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c\right)^2} \right\} \tag{10}$$

Evaluation of REML equation and its derivatives require $\mathbf{H}^{-1}$ and $|\mathbf{H}|$. $\mathbf{H}$ is a sum of two matrices. The inverse and determinant of $\mathbf{B}$ can easily be computed because it is an identity matrix, and $\mathbf{A}$ is a matrix of rank 1. In mrMLM of Wang et al. [16], $\mathbf{H}$ is also a sum of two matrices but in this new algorithm (FASTmrMLM) we have simplified the $\mathbf{H}$ further to be the sum of an identity matrix and a matrix of rank one. Therefore,

$$\mathbf{H}^{-1} = \mathbf{B}^{-1} - \frac{1}{1+g}\mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1} = \mathbf{I} - \frac{1}{1+g}\mathbf{A} = \mathbf{I} - \frac{\omega_i}{1+g}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}' \tag{11}$$

and, $|\mathbf{H}| = (1+g)|\mathbf{B}| = 1+g$, where $g = \mathrm{tr}(\mathbf{A}\mathbf{B}^{-1}) = \mathrm{tr}(\mathbf{A}) = \omega_i\, \mathrm{tr}\left(\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\right)$ [30]. REML requires many quadratic terms in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$, which can be expressed as

$$\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau} = \boldsymbol{\eta}'\boldsymbol{\tau} - \frac{\omega_i}{1+g}\boldsymbol{\eta}'\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\boldsymbol{\tau} \tag{12}$$

where $\boldsymbol{\eta}'$ and $\boldsymbol{\tau}$ can be any vectors or matrices with $\mathbf{H}^{-1}$ such as $\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c$, $\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c$ or $\mathbf{X}_c{}'\mathbf{P}\mathbf{y}_c$. Note that $\boldsymbol{\eta}'\boldsymbol{\tau} = \sum_{j=1}^{n}\eta_j{}'\tau_j$, where η_j corresponds to the *j*th row (element) of the matrix (vector) $\boldsymbol{\eta}$ and τ_j corresponds to the *j*th row (element) of the matrix (vector) $\boldsymbol{\tau}$ for $j = 1,2,\mathrm{L}\ ,\mathrm{n}$. Notice that the value in Equation (8) above can easily be estimated because with the help of Equation (12) each term in the bracket in Equation (8) is in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$. For example,

$\mathbf{y}_c\mathbf{H}^{-1}\mathbf{y}_c = \mathbf{y}_c{}'\mathbf{y}_c - \frac{\omega_i}{1+g}\mathbf{y}_c{}'\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{y}_c$. Also, the gradient function/score function in Equation (9) and Hessian matrix in Equation (10) can also be expressed in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$. With these simplifications, the computation of the variance ratio ω_i for the *i*th SNP is less computationally intensive via the Newton-Raphson method. We estimate the variance ratio ω_i for the *i*th SNP by equating the gradient functions to be zero using the Newton-Raphson technique. With our simplifications,

$$\begin{aligned}\frac{\partial \mathrm{l}_R}{\partial \omega_i} &= -\frac{1}{2}\left[\mathrm{tr}(\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}') - \frac{1}{\hat{\sigma}_e^2}\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}'_{ci}\mathbf{P}\mathbf{y}_c\right] \\ &= -\frac{1}{2}\left\{\mathrm{tr}\left(\mathbf{Z}_{ci}\mathbf{H}^{-1}\mathbf{Z}_{ci}{}' - \mathbf{Z}_{ci}\mathbf{H}^{-1}\mathbf{X}_c\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)^{-1}\mathbf{X}_c{}'\mathbf{H}^{-1\prime}\mathbf{Z}_{ci}{}'\right) - \frac{1}{\hat{\sigma}_e^2}\left[\left(\left(\mathbf{y}_c{}'\mathbf{H}^{-1}\mathbf{Z}_{ci}\right) - \left(\mathbf{y}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)^{-1}\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{Z}_{ci}\right)\right)\right.\right. \\ &\quad \left.\left.\times\left(\left(\mathbf{Z}'_{ci}\mathbf{H}^{-1}\mathbf{y}_c\right) - \left(\mathbf{Z}'_{ci}\mathbf{H}^{-1}\mathbf{X}_c\right)\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{X}_c\right)^{-1}\left(\mathbf{X}_c{}'\mathbf{H}^{-1}\mathbf{y}_c\right)\right)\right]\right\}\end{aligned} \qquad (13)$$

which is in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$, and therefore its computation is so fast. We can also express the second derivative expression in Equation (10) in the form $\boldsymbol{\eta}'\mathbf{H}^{-1}\boldsymbol{\tau}$:

$$\frac{\partial^2 1_R}{\partial \omega_i^2} = -\frac{1}{2}\left\{-\mathrm{tr}(\mathbf{Z}_{ci}\mathbf{P}\mathbf{Z}_{ci}{}'\mathbf{Z}_{ci}\mathbf{P}\mathbf{Z}_{ci}{}') - (n-q)\frac{2\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{y}_c\right)\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c\right) - \left(\mathbf{y}_c{}'\mathbf{P}\mathbf{Z}_{ci}\mathbf{Z}_{ci}{}'\mathbf{P}\mathbf{y}_c\right)^2}{\left(\mathbf{y}_c{}'\mathbf{P}\mathbf{y}_c\right)^2}\right\}.$$

With these simplifications, the Newton-Raphson will converge smoothly to the estimate value of variance ratio ω_i for the *i*th SNP. The gradient function and the Hessian matrix are in a simplified form. Therefore, the computations in each simulation run are fast. We have implemented this algorithm in R software.

Empirical Bayes estimate of β_i: The joint distribution of $\mathbf{y}$ and β_i is a multivariate normal distribution

$$\begin{pmatrix}\mathbf{y} \\ \beta_i\end{pmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix}\mathbf{X}\boldsymbol{\alpha} \\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{Z}_i\mathbf{Z}'_i\omega_i + \mathbf{K}\omega_g + \mathbf{I} & \mathbf{Z}_i\sigma_i^2 \\ \mathbf{Z}_i{}'\sigma_i^2 & \sigma_i^2\end{bmatrix}\right)$$

The conditional distribution of β_i given $\mathbf{y}$ is

$$\beta_i \,|\, \mathbf{y} : \ \mathrm{N}\left[\mathbf{Z}_i'\hat{\sigma}_i^2\left(\mathbf{Z}_i\mathbf{Z}_i'\hat{\sigma}_i^2+\mathbf{K}\hat{\sigma}_g^2+\mathbf{I}\hat{\sigma}_e^2\right)^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\alpha}}\right),\ \hat{\sigma}_i^2-\hat{\sigma}_i^2\mathbf{Z}_i'\left(\mathbf{Z}_i\mathbf{Z}_i'\hat{\sigma}_i^2+\mathbf{K}\hat{\sigma}_g^2+\mathbf{I}\hat{\sigma}_e^2\right)^{-1}\mathbf{Z}_i\hat{\sigma}_i^2\right]$$

From a Bayesian analysis point of view, the conditional mean of β_i given $\mathbf{y}$ is an empirical Bayes estimate of β_i. Based on this framework, we obtain the Wald test statistic $\dfrac{\left(\mathrm{E}\left(\beta_i \,|\, \mathbf{y}\right)\right)^2}{\mathrm{Var}\left(\beta_i \,|\, \mathbf{y}\right)} \sim \chi_1^2$ (Chi-square test with 1 degree of freedom) and using this distribution we obtain P-values for each marker effect. We test each marker effect at 1% level of significance. We do not perform multiple test correction because we intend to include markers that pass this initial test in a multi-locus model.

Detection of true QTNs in multi-locus model

If the number of markers passing the 1% level of significance test is more than n, we invoke the LARS algorithm [32] to select the $n-1$ variables that are most likely associated with the quantitative trait of interest. LARS is a flexible method for variable selection which is conducted in lars package (http://cran.r-project.org/web/packages/lars/) in R language. The $n-1$ markers are then included in a multi-locus model. Note that if the number of markers passing the initial test is less than n, we skip the LARS step and proceed to include all the selected markers in a multi-locus model. We compared various multi-locus methods: SCAD [34], adaptive Lasso [35] and EM-Empirical Bayes [36]. EM-Empirical Bayes [36] has the highest statistical power and accuracy of the estimated marker effects (Tables S8, S9, and S10). EM-Empirical Bayes is a random model method given as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \sum_{i=1}^{s}\mathbf{Z}_i\beta_i + \boldsymbol{\varepsilon} \qquad (14)$$

where $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\alpha}$ are the same as in model in Equation (1), s is the number of potentially associated markers selected from the first step in FASTmrMLM, $\mathbf{Z}_i$ and β_i are $n \times 1$ incident vector and the random effect of the *i*th SNP, respectively. The polygenic variance is not included in the model because the model included all the potentially associated QTNs. We assume a normal prior for β_i, $\beta_i \mid \sigma_i^2 \sim N\left(0, \sigma_i^2\right)$ and a scaled χ^2 prior for σ_i^2, $f\left(\sigma_i^2 \mid \tau, \omega\right) \propto \left(\sigma_i^2\right)^{-\frac{1}{2}(\tau+2)} \exp\left(-\frac{\omega}{2\sigma_i^2}\right)$ and we set $(\tau, \omega) = (0,0)$, which is Jeffrey's prior $f\left(\sigma_i^2 \mid \tau, \omega\right) = \sigma_i^2/2$ [36]. The procedure for parameter estimation in EM-Empirical Bayes is as follows:

a) Initial step: We set initial values as,

$$\begin{aligned} &\sigma_i^2 = 1, \text{ for } i = 1, 2, \text{L} \ , s \\ &\boldsymbol{\alpha} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\sigma_e^2 = \frac{1}{2n}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\right) \end{aligned} \tag{15}$$

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

$$\text{E}\left(\beta_i\right) = \sigma_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\right) \tag{16}$$

where $\mathbf{V} = \sum_{i=1}^{s}\mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 + \mathbf{I}\sigma_e^2$.

c) M-step: To update parameters σ_i^2, $\boldsymbol{\alpha}$ and σ_e^2

$$\begin{aligned} &\sigma_i^2 = \frac{\text{E}\left(\beta_i'\beta_i\right) + \omega}{\tau + 3} \\ &\boldsymbol{\alpha} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &\hat{\sigma}_e^2 = \frac{1}{n}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\right)'\left\{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\right) - \sum_{i=1}^{s}\mathbf{Z}_i\text{E}\left(\beta_i\right)\right\} \end{aligned} \tag{17}$$

where $\text{E}\left(\beta_i'\beta_i\right) = \text{E}\left(\beta_i'\right)\text{E}\left(\beta_i\right) + \text{tr}\left[\text{Var}\left(\beta_i\right)\right]$, $\text{Var}\left(\beta_i\right) = \mathbf{I}\sigma_i^2 - \sigma_i^2\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_i\sigma_i^2$ and $(\tau, \omega) = (0,0)$.

We repeat E-step and M-step until convergence is satisfied. We select all SNPs with a score $\text{LOD} \geq 3$(log of odds) and regard them as significant. We term our algorithm as a fast multi-locus random-SNP-effect mixed linear model (FASTmrMLM).

mrMLM

mrMLM is a two-stage method which tests each random marker effects in the first stage before estimating the significant putative QTNs in a multi-locus model [16]. mrMLM and FASTmrMLM methods were implemented by the R software mrMLM, which is downloaded from http://cran.r-project.org/web/packages/mrMLM/index.html. In mrMLM, we select all markers with a score $\mathrm{LOD} \geq 3$ and regard them as significant.

GEMMA and EMMA

EMMA is a single-locus testing method for GWAS [6] and considers marker effects as fixed. The R codes for EMMA can be downloaded from http://mouse.cs.ucla.edu/emma/. GEMMA is a fast version of EMMA [14]. GEMMA can be run in Linux using source codes obtained from www.xzlab.org/software.html. For EMMA and GEMMA we select significant markers based on Bonferroni correction for multiple tests by setting a threshold for P-value at $0.05/m$, where m is the number of markers.

FarmCPU

FarmCPU was proposed by Liu et al. [28]. In this method, the fixed effect model and random effect model are used iteratively until a stage of convergence is reached. FarmCPU software package source codes are available at http://www.ZZLab.net/FarmCPU. We select significant markers based on Bonferroni correction for multiple tests by setting a threshold for P-value at $0.05/m$, where m is the number of markers.

SCAD

SCAD [34] is a shrinkage method that performs variables selection using concave penalties. SCAD can be run in R using the ncvreg package in R language downloaded from http://cran.r-project.org/web/packages/ncvreg/. Here we select all markers with a score $\mathrm{LOD} \geq 3$ and regard them as significant.

Adaptive Lasso

Adaptive Lasso [35] is a variable selection method that uses data-dependent weights for L_1-penalizing coefficients in the penalty by choosing the inverse of ordinary least-square estimates for the weights. Adaptive Lasso can be run in R using the parcor package in R language downloaded from http://cran.r-project.org/web/packages/parcor/. Here we select all markers with a score $\text{LOD} \geq 3$ and regard them as significant.

Monte Carlo simulation experiments

We carried out three Monte Carlo simulation experiments to validate FASTmrMLM. In all these simulation studies, we sampled the genetic marker values from 216,130 SNPs markers in Atwell et al. [31] (http://www.arabidopsis.org/), and we simulated all the phenotypes values for quantitative traits. We used the same sample size of 199 diverse inbred lines like that in Atwell et al. [31] dataset. We randomly sampled 2000 SNPs from each of the five chromosomes; in total, we obtained 10000 SNP genotypes. For the selected SNPs, we used the same positions and genotypes as those in Wang et al. [16].

The first experiment tested FASTmrMLM method on a model with no polygenic variances simulated. From the sampled SNPs, we set six QTN with effect sizes and positions as shown in Tables S8 to S10. We placed these QTNs on the SNPs with allelic frequencies of 0.30. The heritabilities for these QTNs were 0.10, 0.05, 0.05, 0.15, 0.05 and 0.05, respectively. In this case, the phenotypes were simulated from the model: $\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{6} \mathbf{Z}_i \beta_i + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} : \text{MVN}\left(\mathbf{0}, 10 \times \mathbf{I}\right)$. We simulated an overall mean of 10. In this simulation experiment, we carried out 1000 runs for each of the methods considered, i.e., FASTmrMLM, EMMA, GEMMA, FarmCPU, and mrMLM. For each simulated QTN, we counted the runs in which we obtained the QTN LOD ≥ 3 for FASTmrMLM and mrMLM and P-values less than $0.05/m$ (Bonferroni correction for multiple tests), where m is the total number of SNPs for GEMMA, EMMA, and FarmCPU. We considered QTNs within 1kb of the simulated QTN as true QTN. We counted the number of runs in which we deemed to have obtained each QTN. The proportion of such runs relative to the total number of runs (1000) represented the statistical power of this QTN. We calculated false positive rate (FPR) as the ratio of the number of false non-zero effects relative to the total number of zero effects considered in the full model. We measured the bias of each QTN effect estimate by calculating the mean squared error (MSE). We also carried out a paired t-test for the differences of statistical power or MSE between FASTmrMLM and other methods.

The second simulation experiment tested FASTmrMLM method on a model with an additive polygenic (small effect genes) background. We simulated polygenic effect from a multivariate normal distribution, $\mathrm{MVN}\left(\mathbf{0},\mathbf{K}\sigma_g^2\right)$ where σ_g^2 is the polygenic variance and $\mathbf{K}$ is the kinship matrix. We used $\sigma_g^2 = 2$ hence $h_g^2 = 0.092$. The QTN size (h^2), average, residual variance, and other values were the same as those in the first simulation study, i.e., the phenotypes were simulated from the model given as: $\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{6} \mathbf{Z}_i\beta_i + \boldsymbol{\phi} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\phi} \sim \mathrm{MVN}\left(\mathbf{0}, 2\times\mathbf{K}\right)$ and $\boldsymbol{\varepsilon} \sim \mathrm{MVN}\left(\mathbf{0}, 10\times\mathbf{I}\right)$.

The third simulation experiment examined FASTmrMLM on a model with an epistatic background. We simulated three epistatic QTN each with $\sigma_{epi}^2 = 1.25$ and $h_{epi}^2 = 0.05$. The details for the three epistatic QTN were described in Wang et al. [16]. The QTN size (h^2), average, residual variance, and other values were the same as those in the first simulation experiment. The phenotypes were simulated from the model:

$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{6} \mathbf{Z}_i \beta_i + \sum_{j=1}^{3} \left(\boldsymbol{C}_j \# \boldsymbol{D}_j\right) \beta_{jj} + \boldsymbol{\varepsilon}$, where, $\boldsymbol{\varepsilon} :$ $\text{MVN}\left(\mathbf{0}, 10 \times \mathbf{I}\right)$, β_{jj} is the epistatic effect and $\left(\boldsymbol{C}_j \# \boldsymbol{D}_j\right)$ is its incidence coefficient.

The *Arabidopsis thaliana* data

We re-analyzed six flowering time-related traits in *Arabidopsis* [31] by GEMMA, FASTmrMLM, EMMA, FarmCPU and mrMLM to validate our new method. These traits are days to flowering under long days (LD), days to flowering under long days with vernalization (LDV), days to flowering under short days (SD), days to flowering under LD with no vernalization (0W), days to flowering under long days with 2 weeks vernalized (2W), and days to flowering under long days with 4 weeks vernalized (4W). We downloaded these datasets from http://www.arabidopsis.usc.edu/. In the real data analyses, we determined the significantly associated SNPs by the critical threshold of $\text{LOD} \geq 3$ for FASTmrMLM and mrMLM, and with the P-values less than $0.05/m$ for GEMMA, EMMA, and FarmCPU. We mined candidate genes for the trait under study within 30 kb of the significantly relevant SNPs.

Additional file

Table S1. Comparison of the Statistical power in the detection of QTN in three simulation experiments using five GWAS approaches

Table S2. Paired t-test for the differences of statistical power and mean squared error (MSE) between FASTmrMLM (new) and other methods

Table S3. Comparison of the Mean squared error (MSE) for QTN effects across 1000 replicates in three simulation experiments using FASTmrMLM, mrMLM, FarmCPU, and GEMMA/EMMA GWAS approaches

Table S4. Comparison of the Mean for QTN effects across 1000 replicates in three simulation experiments using FASTmrMLM, mrMLM, FarmCPU and GEMMA/EMMA GWAS approaches

Table S5. Goodness of fit (AIC, BIC) for SNPs detected by FASTmrMLM, mrMLM, FarmCPU, and GEMMA/EMMA where a lower value indicates a better fit

Table S6. GWAS for six flowering time traits in Arabidopsis thaliana using FASTmrMLM, mrMLM, FarmCPU, and GEMMA methods

Table S7. New Genes detected only by FASTmrMLM in the GWAS of six flowering time traits in Arabidopsis thaliana

Table S8. Comparison of EM-Empirical Bayes, Adaptive Lasso, and SCAD with no polygenic variance simulated

Table S9. Comparison of EM-Empirical Bayes, Adaptive Lasso, and SCAD in the second simulation experiment with an additive polygenic background (explaining 0.092 of the phenotypic variance)

Table S10. Comparison of EM-Empirical Bayes, Adaptive Lasso, and SCAD in the third simulation experiment with three epistatic QTNs each explaining 0.05 of the phenotypic variance

Abbreviations

GWAS: genome-wide association study; SNP: single nucleotide polymorphisms; QTN: quantitative trait nucleotide; QTL: quantitative trait locus; EM: expectation and maximization; EMMA: efficient mixed-model association; GEMMA: Genome-wide efficient mixed-model association; FarmCPU: fixed and random model circulating probability unification; mrMLM: multi-locus random-SNP-effect mixed linear model; FASTmrMLM: a fast mrMLM multi-locus mixed linear model. ISIS EM-BLASSO; Iterative modified-Sure Independence

Screening EM-Bayesian LASSO; FASTmrEMMA: fast multi-locus random-SNP-effect EMMA.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Y.-M.Z. conceived and supervised the study, and revised the manuscript. C. L.T. performed the experiments and analysed the data, and wrote the draft. All the authors read, revised and approved the final manuscript.

Acknowledgements

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

Availability of data and materials

A single source file for all simulated datasets has been included as Additional files.

References

1. Wellcome Trust Case Control Consortium, Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature** 2007; 447:661–678.
2. Daly AK. Genome-wide association studies in pharmacogenomics. **Nat. Rev. Genet.** 2010, 11: 241–246.
3. Niu Y, Xu Y, Liu XF, Yang SX, Wei SP, Xie FT, Zhang YM. Association mapping for seed size and shape traits in soybean cultivars. **Mol. Breeding** 2013,31(4): 785-794.

4. Zhang YM, Mao Y, Xie C, Smith H, Luo L, Xu S. Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). **Genetics** 2005, 169, 2267– 2275.

5. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nat. Genet.** 2006, 38, 203–208.

6. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. **Genetics** 2008, 178: 1709–1723.

7. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. **Nat. Genet.** 2010;42:355–360.

8. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. **Nat. Genet.** 2010, 42:348–354.

9. Li M, Liu X, Bradbury P, Yu J, Zhang YM, Todhunter RJ, et al. Enrichment of statistical power for genome-wide association studies. **BMC Biol.** 2014, 12:73.

10. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. Fast linear mixed models for genome-wide association studies. **Nat. Methods** 2011, 8:833–5.

11. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. **Nat. Methods** 2012, 9(6):525–526.

12. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. **Nat. Genet.** 2012, 44:1166–1170.

13. Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. A SUPER powerful method for genome wide association study. **PLoS ON**E 2014; 9:e107684.

14. Zhou X, Stephens M. Genome-wide Efficient Mixed Model Analysis for Association Studies. **Nat. Genet.** 2012, 44(7): 821–824. doi:10.1038/ng.2310.

15. Goddard ME., Wray NR., Verbyla K, Visscher PM. Estimating effects and making predictions

from genome-wide marker data. **Stat. Sci.** 2009, 24, 517–529.

16. Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. **Sci. Rep.** 2016, 6: 19444.

17. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. **Nat. Genet.** 2012, 44: 825–830.

18. Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, et al. Methodological Implementation of mixed linear models in multi-locus genome-wide association studies. **Briefings in Bioinformatics** 2017, bbw145, DOI: 10.1093/bib/bbw145.

19. Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. **Genetics** 2008,179 (2):1045–55.

20. Hoggart CJ, Whittaker JC, De Iorio MD, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. **PLoS Genet.** 2008, 4:e1000130.

21. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. **Genet. Epidemiol.** 2010, 34:879–91.

22. Wang D, Eskridge KM, Crossa J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. **J. Agric. Biol. Environ. Stat.** 2010, 16:170–84.

23. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, et al. Joint identification of multiple genetic variants via Elastic-Net variable selection in a genomewide association analysis. **Ann. Hum. Genet.** 2010, 74:416–428.

24. Lü HY, Liu XF, Wei SP, Zhang YM. Epistatic association mapping in homozygous crop cultivars. **PLoS ONE** 2011, 6: e17773.

25. Wen J, Zhao X, Wu G, Xiang D, Liu Q, Bu HS, et al. Genetic dissection of heterosis using epistatic association mapping in a partial NCII mating design. **Sci. Rep.** 2015, 5: 18376.

26. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. **PLoS Genet.** 2013, 9:e1003264.

27. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. **PLoS Genet.** 2015, 11(4): e1004969.

28. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS Genet.** 2016, 12(2): e1005767.

29. Tamba CL, Ni YL, Zhang YM. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. **PLoS Comput Biol** 2017, 13(1): e1005357.

30. Miller K. On the Inverse of the Sum of Matrices. **Mathematics Magazine** 1981, 54(2): 67–72.

31. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. **Nature** 2010, 465: 627–631.

32. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. **Annals of Statistics** 2004, 32: 407–451.

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

34. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. **Journal of the American Statistical Association** 2001, 96 (456): 1348–1360.

35. Zou H. The adaptive lasso and its oracle properties. **Journal of the American Statistical Association** 2006, 101(476): 1418–1429.

36. Xu S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. **Heredity** 2010, 105:483–494.

Figure legends

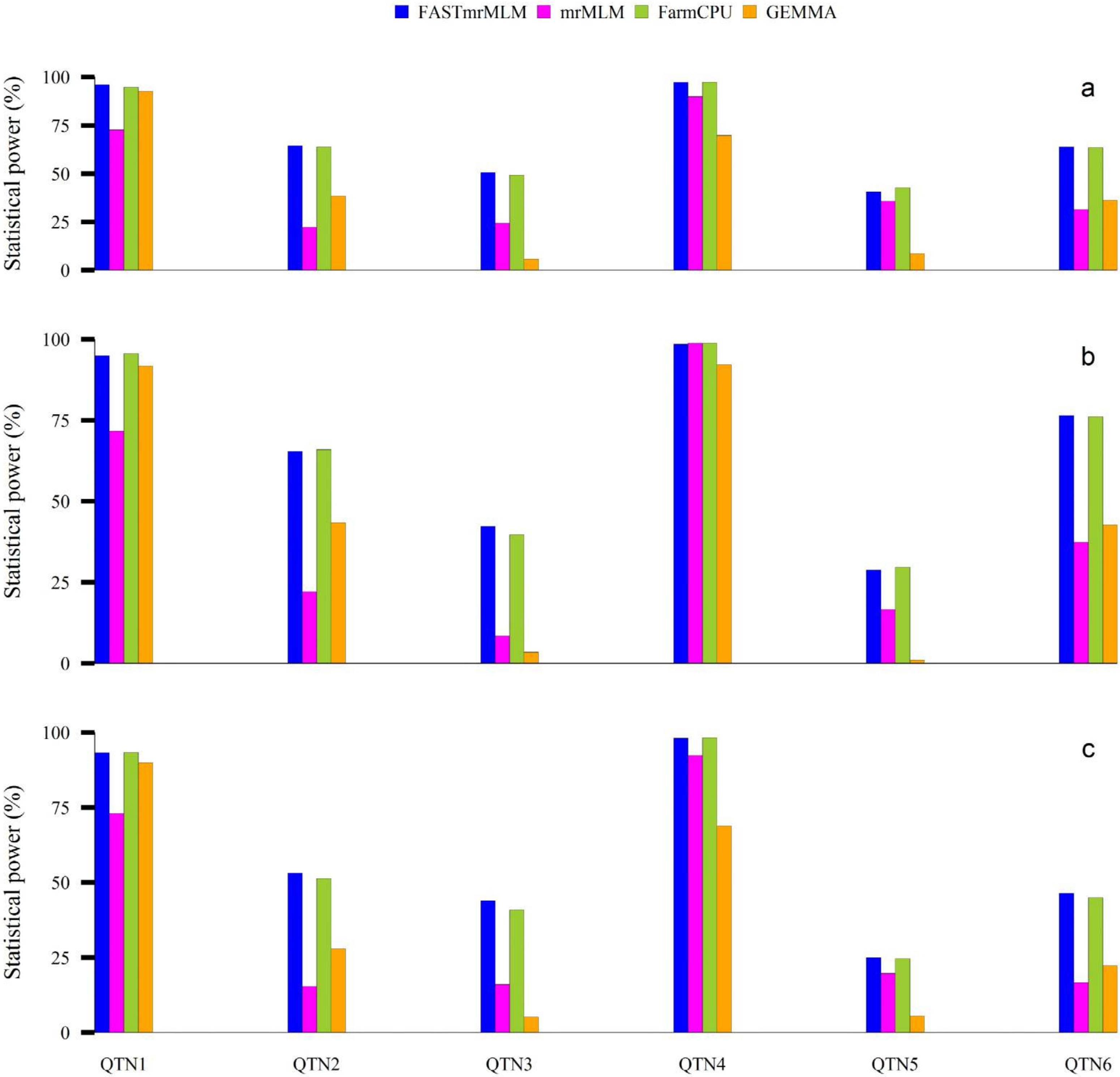
Fig 1. Computing times (hours) taken by FASTmrMLM, mrMLM, FarmCPU, GEMMA, and EMMA in the Monte Carlo simulation experiments I (**a**), II (**b**) and III (**c**).

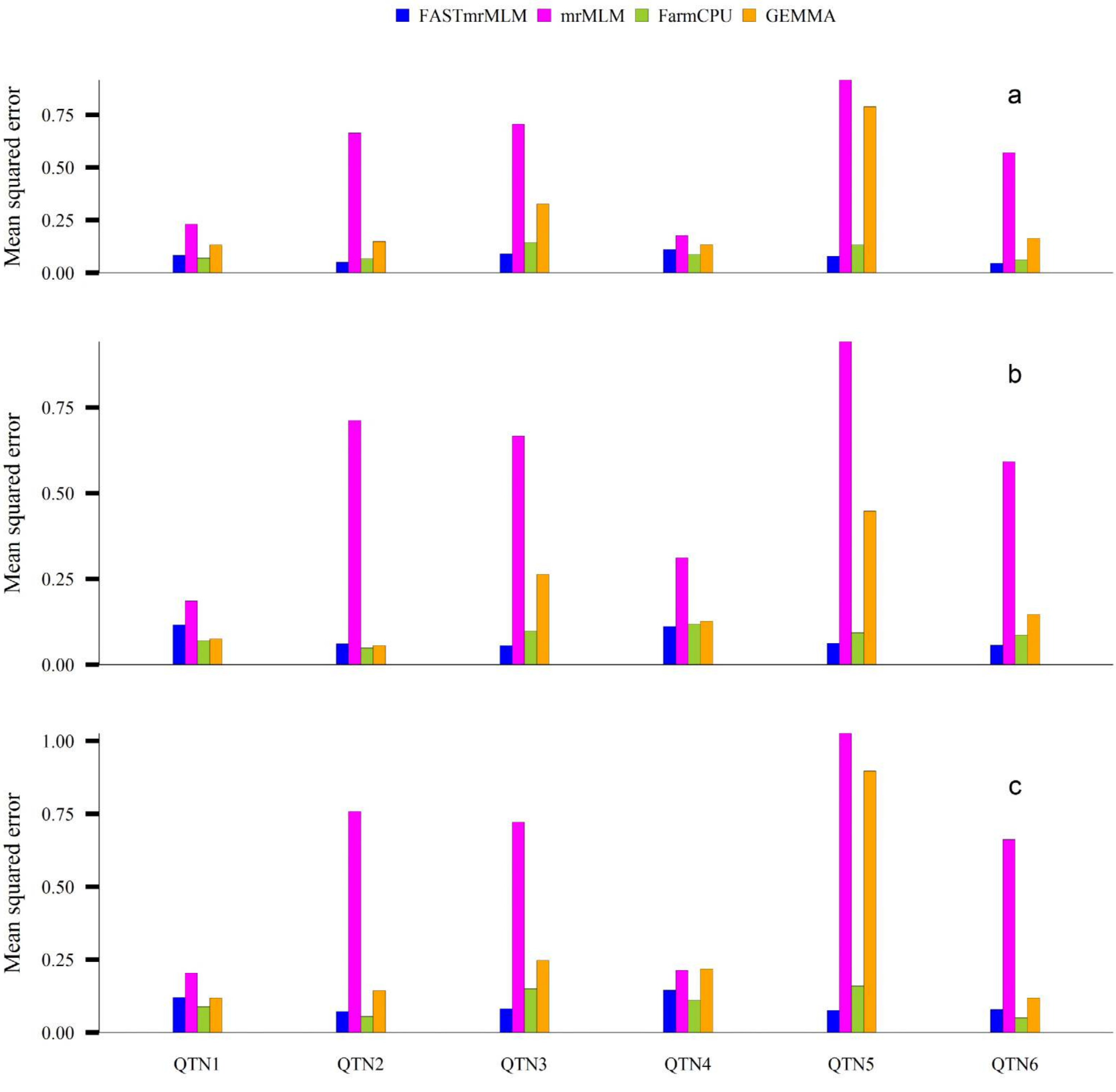
Fig 2. Statistical powers for all the simulated QTNs detected by FASTmrMLM, mrMLM, FarmCPU, and GEMMA in Monte Carlo simulation experiments I (**a**), II (**b**) and III (**c**)

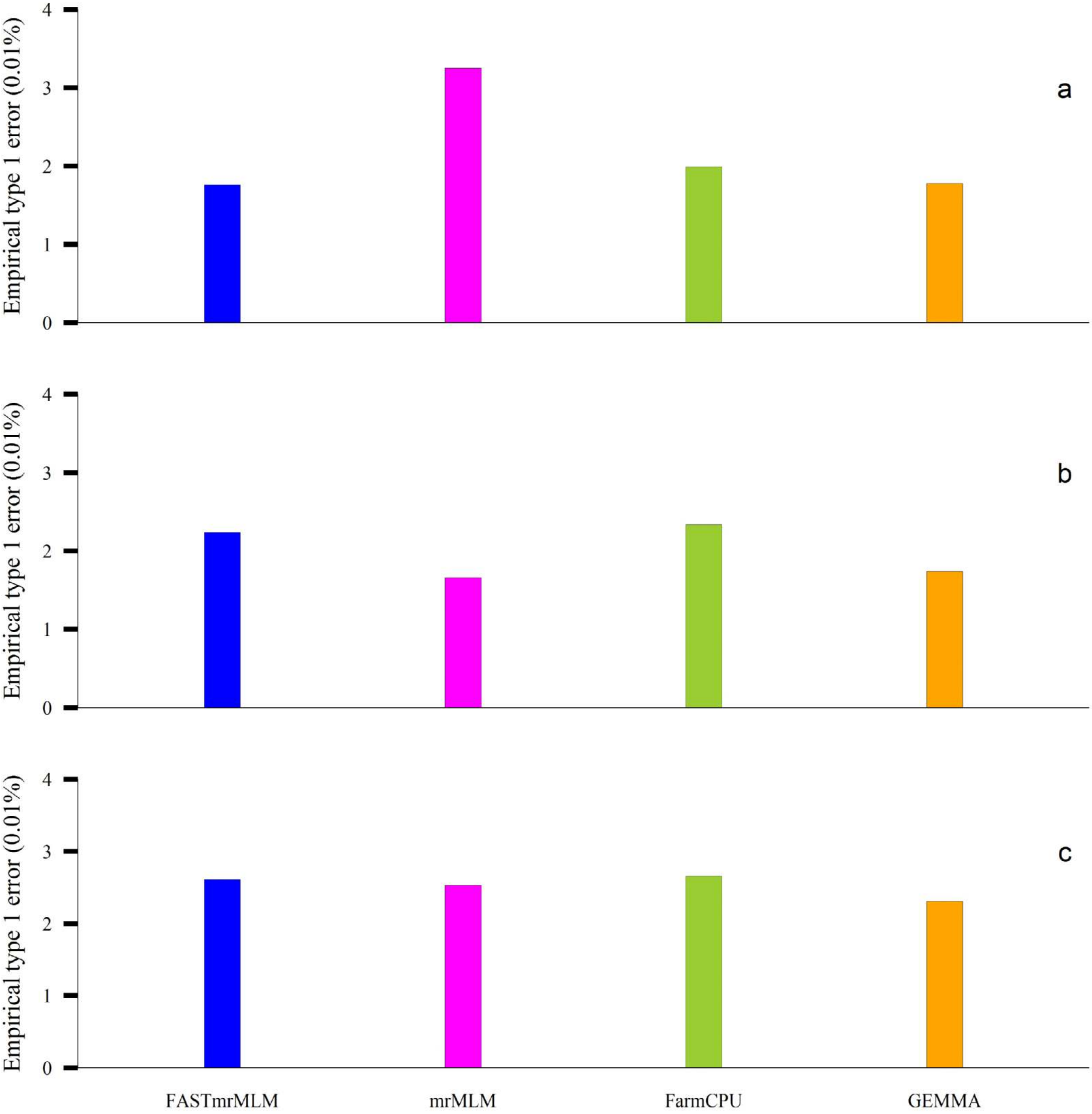
Fig 3. Mean square error (MSE) for all the simulated QTN effects estimated by F FASTmrMLM, mrMLM, FarmCPU, and GEMMA in Monte Carlo simulation experiments I (**a**), II (**b**) and III (**c**).

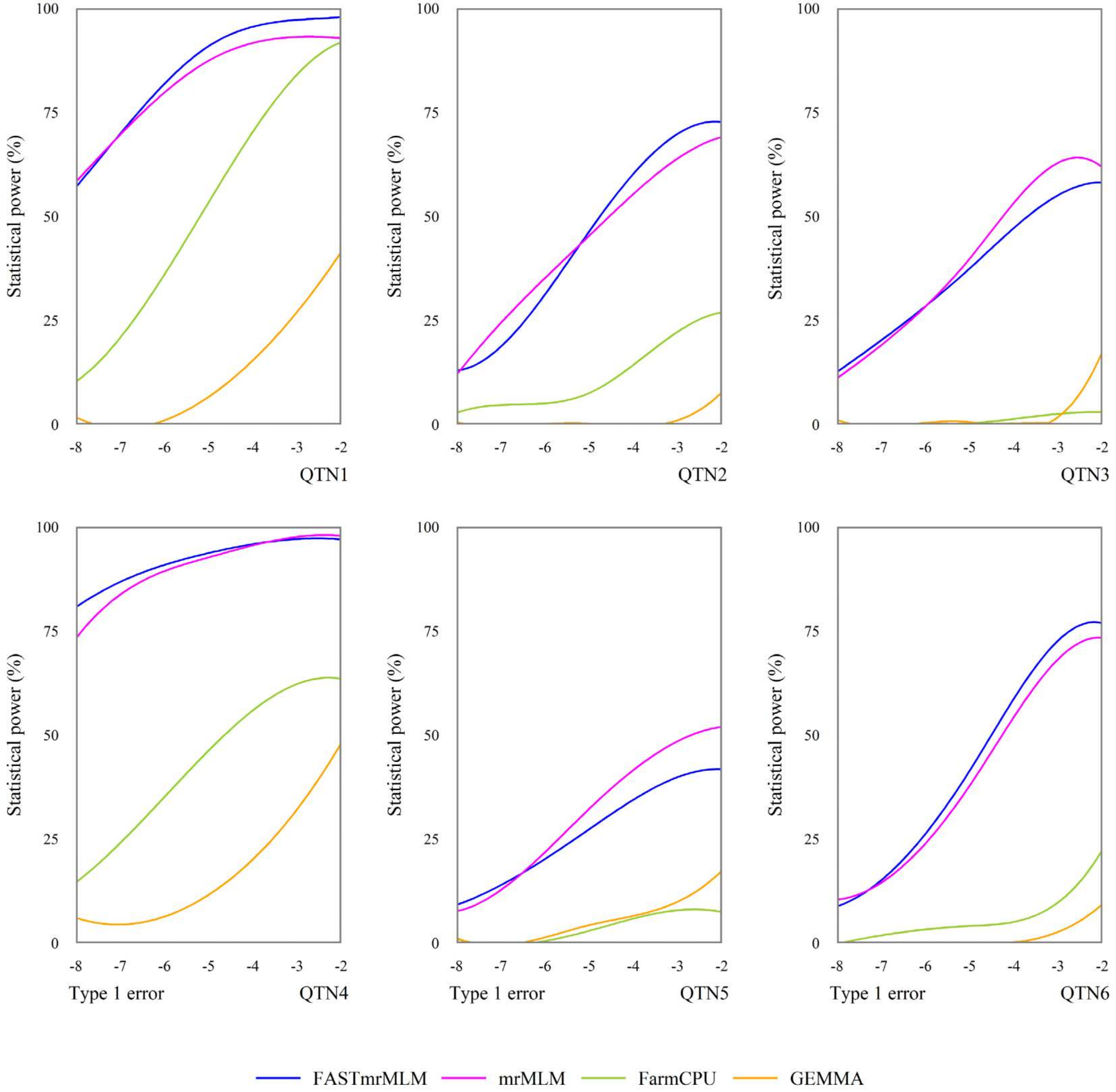
Fig 4. Empirical type 1 error rates derived from FASTmrMLM, mrMLM, FarmCPU, and GEMMA in the Monte Carlo simulation experiments I (**a**), II (**b**) and III (**c**).

Fig 5. Statistical powers of all the simulated QTNs in the first simulation experiment plotted against Type 1 error (in a log10 scale) for the five GWAS methods (FASTmrMLM, mrMLM, FarmCPU, and GEMMA).

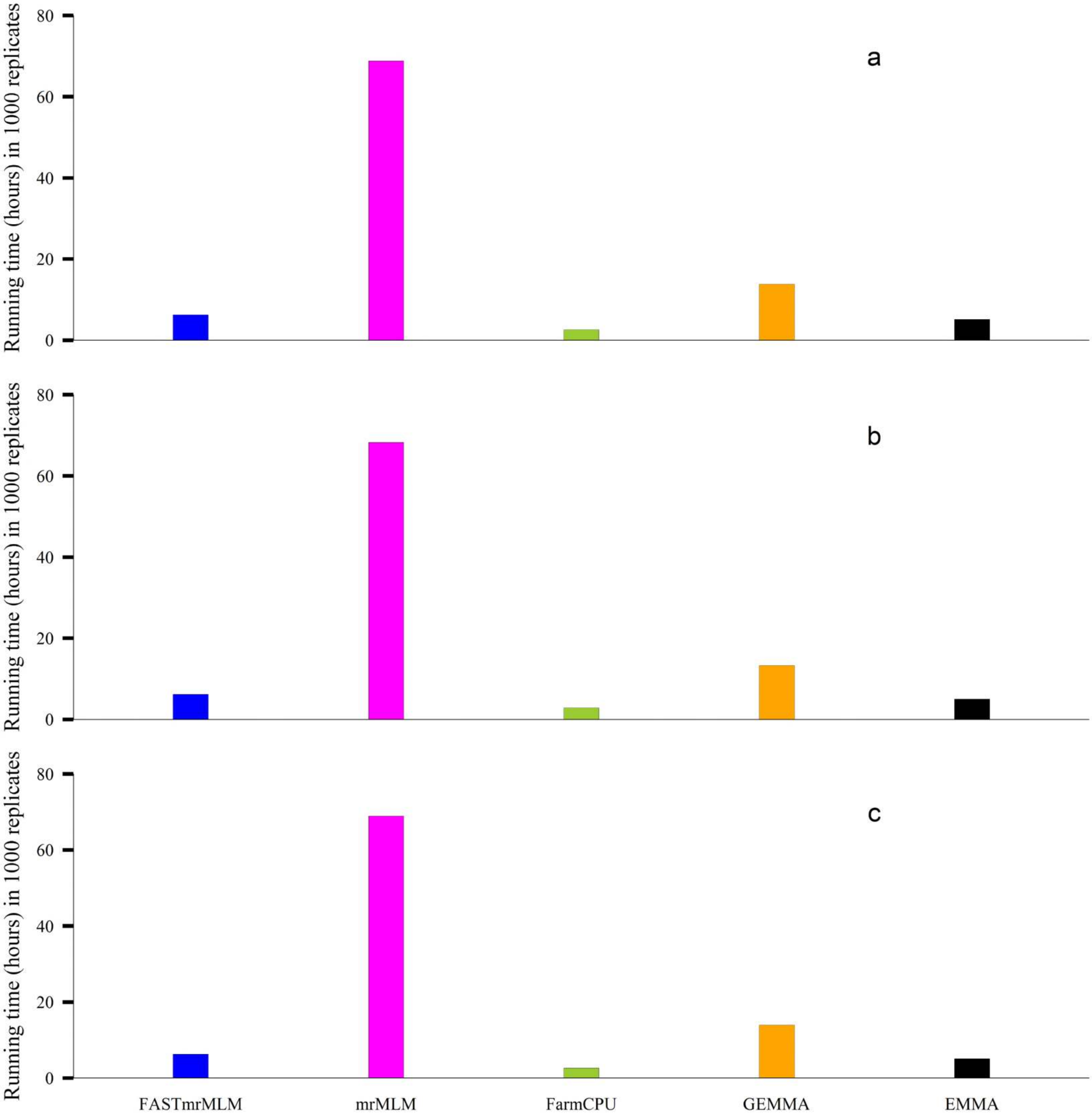

FASTmrMLM
mrMLM
FarmCPU
GEMMA
Statistical power (%)
0
25
50
75
100
a
b
c
QTN1
QTN2
QTN3
QTN4
QTN5
QTN6


FASTmrMLM
mrMLM
FarmCPU
GEMMA
a
b
c
Mean squared error
0.00
0.25
0.50
0.75
1.00
QTN1
QTN2
QTN3
QTN4
QTN5
QTN6


Empirical type 1 error (0.01%)
0
1
2
3
4
a
b
c
FASTmrMLM
mrMLM
FarmCPU
GEMMA



Statistical power (%)
100
75
50
25
0
-8
-7
-6
-5
-4
-3
-2
QTN1
QTN2
QTN3
Type 1 error
QTN4
QTN5
QTN6
FASTmrMLM
mrMLM
FarmCPU
GEMMA


Running time (hours) in 1000 replicates
80
60
40
20
0
a
b
c
FASTmrMLM
mrMLM
FarmCPU
GEMMA
EMMA