Computational and Structural Biotechnology Journal 23 (2024) 4357–4368

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

IIIVmrMLM.QEI: An effective tool for indirect detection of QTN-by-environment interactions in genome-wide association studies

Ya-Wen Zhang [a,b], Xue-Lian Han [a], Mei Li [a], Ying Chen [a], Yuan-Ming Zhang [a,*]

[a] College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
[b] International Genome Center, Jiangsu University, Zhenjiang 212013, China

ARTICLE INFO

Keywords:
QTN-by-environment interaction
IIIVmrMLM.QEI
Genome-wide association studies
Flowering time
Variation indicator
Environmental factor

ABSTRACT

Although 3VmrMLM-MEJA and several indirect indicators have been employed to identify QTN-by-environment interactions (QEIs) in genome-wide association studies (GWAS), there is no convenient, flexible, and accurate method to comprehensively identify QEIs. To address this issue, 3VmrMLM-random was first extended to 3VmrMLM-fixed. Next, the two single-environment QTN detection methods were integrated with trait differences and regression parameters to indirectly detect QEIs. Finally, these indirect indicators were extended to include environmental factors (EFs, such as temperature) and four environmental variation indicators. As a result, both 3VmrMLM-random and 3VmrMLM-fixed, alongside all the indirect indicators, were incorporated into a new tool, IIIVmrMLM.QEI, designed for effective QEI identification. Simulation studies demonstrated that 3VmrMLM-fixed showed significantly higher powers than existing fixed-SNP-effect methods (MLM and EMMAX) because it takes into account all the possible effects and controls for all the possible polygenic backgrounds. 3VmrMLM-random and 3VmrMLM-fixed exhibited superior combination power to 3VmrMLM-MEJA. In the re-analysis of *Arabidopsis* flowering time across three temperatures, 3VmrMLM-fixed (12) detected more known gene-by-environment interactions (GEIs) than both MLM (1) and EMMAX (1). Additionally, IIIVmrMLM.QEI (18) detected more known GEIs than 3VmrMLM-MEJA (6), when all indirect indicators were analyzed. All 18 GEIs were confirmed by haplotype analysis and associated with temperature variation in previous studies. Two and five GEIs were identified only by 3VmrMLM-fixed and 3VmrMLM-random, respectively, and 12 GEIs were identified only by indirect indicators, indicating the need to expand models and indirect indicators. This study provides a novel tool (https://github.com/YuanmingZhang65/IIIVmrMLM.QEI) for more comprehensive QEI detection.

1. Introduction

Gene-by-environment interaction (GEI) is an important genetic component of complex traits, reflecting the fact that the effects of a gene may vary between environments or may only be detectable in certain environments. With the escalating impacts of global climate change, sustainable crop production faces significant challenges. Developing climate-resilient crops necessitates identifying quantitative trait nucleotides (QTNs) and QTN-by-environment interactions (QEIs).

Over the past decade, several indirect approaches have been used to identify QEIs in genome-wide association studies (GWAS). First, QEIs are identified using multi-trait GWAS methods, where multiple environments are considered as multiple traits [1]. Their power is equal to the trait difference as phenotype [2,3]. Second, a number of indirect indicators have been used to detect QEIs via some widely used GWAS methods, such as genetic risk score [4,5], environmental score [6], and regression parameters [7]. However, the sample sizes for these approaches are equivalent to the number of individuals (n) in a single environmental experiment. This reduces the power of QEI detection and makes QTN detection infeasible [8]. To address this issue, Sul et al. [9] and Moore et al. [10] developed mixed model-based methods. However, their mixed models only considered allelic substitution effects and their polygenic backgrounds, and did not include additive and dominance effects, additive-by-environment (ae) and dominance-by-environment (de) interaction effects, and their corresponding polygenic backgrounds. To address the detection of interactions in human genetics, the

* Corresponding author.
E-mail addresses: yawen@ujs.edu.cn (Y.-W. Zhang), luluzi@webmail.hzau.edu.cn (X.-L. Han), 1286985431@qq.com (M. Li), chenying8861@163.com (Y. Chen), soyzhang@mail.hzau.edu.cn (Y.-M. Zhang).

https://doi.org/10.1016/j.csbj.2024.11.046
Received 17 August 2024; Received in revised form 24 November 2024; Accepted 29 November 2024
Available online 2 December 2024

Brown-Forsythe test [11,12] and the double generalized linear model [13] have been used to detect variance controlling loci, although their power is relatively low [14]. It should be noted that these mixed model based methods are based on the assumption of random mating. To overcome these problems, we have recently established a compressed variance component mixed model framework, namely 3VmrMLM [8]. In our 3VmrMLM method, the assumption of random mating is not required, all the effects to be detected and estimated in QTN, QEI, and QTN-by-QTN interaction detection are compressed into an effect-related vector, and all the possible polygenic backgrounds are compressed into an effect-vector-related polygenic background, and the sample size in multi-environment joint GWAS analysis is equal to *mn*, where *m* is the number of environments and *n* is the number of individuals. Although 3VmrMLM has been established, there are limited methods available to comprehensively identify QEIs.

To address the above issue, this study developed a novel tool, IIIVmrMLM.QEI, for more comprehensive QEI detection. In IIIVmrMLM.QEI, 3VmrMLM-random was first extended to 3VmrMLM-fixed, and these two methods were integrated with trait differences, regression intercept (RI), and regression coefficient (RC) as phenotypes for indirectly identifying QEIs. More importantly, the above three indirect indicators were extended to environmental factor (EF, such as temperature) and four environmental variation indicators: range, variance (Var), standard deviation (SD), and coefficient of variation (CV). These environmental variation indicators effectively capture phenotypic differences and variability across environments, providing a direct measure of phenotypic response in different environments. In addition, RI evaluates general performance, RC measures genotype-specific phenotypic changes across different environments, and EF assesses adaptability. QEIs, detected using these methods and indicators, are more comprehensive than those identified by existing approaches. Monte Carlo simulations validated all methods in the tool, which were also benchmarked against current methods [8,15,16], highlighting the need to extend the fixed model and indirect indicators. Researchers will benefit from the IIIVmrMLM.QEI for more comprehensive QEI and GEI identification, and the extension of year / location environments to meteorological factors and other environmental conditions. To distinguish these new methods, the integration of 3VmrMLM-random with RI as phenotypes is abbreviated as 3VmrMLM-random-RI, and the others are similar.

2. Materials and methods

2.1. Association mapping population

The well-known *Arabidopsis thaliana* flowering time dataset was downloaded from http://www.arabidopsis.usc.edu/ [17]. A total of 199 accessions were genotyped by 216130 SNP markers, and 192 accessions were phenotyped by flowering time under 16 h daylight at 10 ℃, 16 ℃, and 22 ℃ conditions, representing three distinct environments. The datasets were listed in Data Files S1-S5. 200 kb upstream and downstream regions of significant / suggested QEIs were used to mine known FT genes [18], while the 200 kb was changed as 50 kb for all the candidate GEIs.

2.2. Monte Carlo simulation study

The Monte Carlo simulation studies simulated two scenarios similar to those in our previous studies [19,20]. The first dataset was simulated by sampling 200 individuals from 1136 Simmental cattle (https://datadryad.org/stash/dataset/doi:10.5061/dryad.4qc06; Zhu et al. [21]), each with 20000 SNP markers selected from 597650 SNP markers. To simulate the phenotypic values in three environments, three QTNs, three QEIs, and four QTNs with main and environmental interaction effects (r^2: 2 %~8 %) were simulated on 10 chromosomes. The numbers, positions, and sizes of QTNs and QEIs are listed in Table S1. All the phenotypes were simulated based on the model $\mathbf{y} = \mathbf{X\alpha} + \mathbf{Z}_E e + \sum_{i=1}^{10} Z_i\gamma_i + \sum_{i=1}^{10} Z_{ei}\gamma_{ei} + \boldsymbol{\varepsilon}$, where e is environmental effect, $\mathbf{Z}_E$ is its design matrix; γ_i is the effect of the ith QTN, and Z_i is its design matrix; γ_{ei} is the effect of the ith QEI, and Z_{ei} is its design matrix; $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(0, 10 \times \mathbf{I}_{bn})$, $b = 3$, and $n = 200$. 1000 replicates were performed to evaluate the performances of all the methods, including statistical power of all the simulated loci, false positive rate (FPR) and false negative rate (FNR), and the accuracies for the estimates for positions and effects of QTNs and QEIs, as described in our previous studies. Significant loci were determined based on Bonferroni correction, while suggested loci for 3VmrMLM related methods were determined based on LOD score ≥ 3.0 [8]. All the simulation datasets are listed in Data Files S6-S9. The second dataset was simulated using PLINK 1.9 [22], consisting of 1000 (n) individuals with one million markers each. To simulate the phenotypic values in five (b) environments, three QTNs, three QEIs, and four QTNs with main and environmental interaction effects (r^2: 1 %~7 %) were simulated on one chromosome. The number of replicates was 20. The numbers, positions and sizes of the QTNs and QEIs are listed in Table S9, while all the simulation datasets are listed in Data Files S10-S13.

2.3. 3VmrMLM-fixed: Fixed SNP effect model of QTN detection in single environment data analysis

Genetic model As described in Li et al. [8], genetic model for complex trait $\mathbf{y}$ in QTN detection using 3VmrMLM method is

$$\mathbf{y} = \mathbf{X\alpha} + \mathbf{Z}_t\boldsymbol{\gamma}_t + \mathbf{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{\alpha}$ is a fixed effect vector, $\mathbf{X}$ is its design matrix for $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}_t$ is a 3×1 vector of QTN genotypic effects, and $\mathbf{Z}_t$ is the $n \times 3$ design matrix for $\boldsymbol{\gamma}_t$, in which marker genotypes are coded as (1, 0, 0) for AA, (0, 1, 0) for Aa, and (0, 0, 1) for aa; $\mathbf{u} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{K}\phi^2)$ is a vector of QTN genotype polygenic effects, where $\mathbf{K} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{Z}_t\mathbf{Z}_t^{\mathrm{T}}$ is kinship matrix, ϕ^2 is polygenic background variance, and m is the number of markers; $\boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is an independent and identically distributed vector of residual errors, σ_e^2 is residual variance, and $\mathbf{I}$ is unit matrix.

If the vector γ_t in model (1) is treated as fixed, the above model is transformed into

$$\mathbf{y} = (\mathbf{X} \quad \mathbf{Z}_t)\begin{pmatrix}\boldsymbol{\alpha}\\ \gamma_t\end{pmatrix} + \mathbf{u} + \boldsymbol{\varepsilon} = \mathbf{X}^*\boldsymbol{\alpha}^* + \mathbf{u} + \boldsymbol{\varepsilon} \tag{2}$$

Thus, we have

$$\mathbf{E}(\mathbf{y}) = \mathbf{X}^*\boldsymbol{\alpha}^*$$
$$\mathbf{Var}(\mathbf{y}) = \mathbf{K}\sigma_g^2 + \mathbf{I}\sigma_e^2 = \sigma_e^2\left(\mathbf{K}\lambda_g + \mathbf{I}\right)$$

where $\lambda_g = \sigma_g^2/\sigma_e^2$ is the variance ratios. The λ_g value can be pre-estimated under pure polygenic model, and λ_g is fixed as $\widehat{\lambda}_g$ when testing each SNP effect in genome-wide scanning. Using eigen decomposition for K, we can obtain $\mathbf{K} = \mathbf{UDU}'$, where $\mathbf{D} = diag\{\delta_1, \ldots, \delta_n\}$ is a diagonal matrix for the eigenvalues and $\mathbf{U}$ is an $n \times n$ matrix for the eigenvectors. Let $\mathbf{y}_c = \mathrm{U}'\mathbf{y}$, $\mathbf{X}_c^* = \mathrm{U}'\mathbf{X}^*$, $\mathbf{u}_c = \mathrm{U}'\mathbf{u}$, and $\boldsymbol{\varepsilon}_c = \mathrm{U}'\boldsymbol{\varepsilon}$, model (2) was changed into

$$\mathbf{y}_c = \mathbf{X}_c^*\boldsymbol{\alpha}^* + \mathbf{u}_c + \boldsymbol{\varepsilon}_c \tag{3}$$

$$\mathrm{Var}(\mathbf{y}_c) = \mathrm{Var}(\mathbf{U}'\mathbf{y}) = \mathbf{D}\lambda_g + \mathbf{I}.$$

Restricted log-likelihood function Restricted log-likelihood function of $\mathbf{y}_c$ is

$$\ell_R(\alpha^*, \sigma_e^2) \propto -\frac{1}{2}\left[(n-q)\log\sigma_e^2 + \log|\mathbf{H}| + \log|\mathbf{X}_c^{*\prime}\mathbf{H}^{-1}\mathbf{X}_c^*| + \frac{1}{\sigma_e^2}\left(\mathbf{y}_c - \mathbf{X}_c^*\alpha^*\right)'\mathbf{H}^{-1}\left(\mathbf{y}_c - \mathbf{X}_c^*\alpha^*\right)\right] \tag{4}$$

where $q = \text{rank}(\mathbf{X}_c^*)$. Thus, restricted maximum likelihood (REML) estimates for parameters of interest can be obtained as below.

Estimation of residual variance σ_e^2 and fixed effect $\widehat{\boldsymbol{\alpha}}^*$ Using REML function, we can obtain that:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^* &= \left(\mathbf{X}_c^{*\prime}\mathbf{H}^{-1}\mathbf{X}_c^*\right)^{-1}\mathbf{X}_c^{*\prime}\mathbf{H}^{-1}\mathbf{y}_c^* \\ \widehat{\sigma}_e^2 &= \frac{1}{n-q}\left(\mathbf{y_c}-\mathbf{X}_c^*\widehat{\boldsymbol{\alpha}}^*\right)^{\mathrm{T}}\mathbf{H}^{-1}\left(\mathbf{y_c}-\mathbf{X}_c^*\widehat{\boldsymbol{\alpha}}^*\right)\end{aligned} \tag{5}$$

where $\mathbf{H} = \mathbf{D}\widehat{\lambda}_g + \mathbf{I}$.

Tests for locus related effects Wald test statistic for locus-related effects $\boldsymbol{\gamma}_t$ under $H_0 : \boldsymbol{\gamma}_t = 0$ is

$$W_k = \widehat{\gamma}_{\mathbf{t}}^T[\text{var}(\widehat{\gamma}_{\mathbf{t}})]^{-1}\widehat{\gamma}_{\mathbf{t}} \tag{6}$$

where $\text{var}(\widehat{\boldsymbol{\gamma}}_t)$ can be obtained from$\text{var}(\widehat{\alpha}^*) = \widehat{\sigma}_e^2(\mathbf{X}_c^{*\prime}\mathbf{H}^{-1}\mathbf{X}_c^*)^{-1}$. The probability for H_0 is calculated from $1 - \Pr(\chi_\nu^2 < W_k)$, where ν is the degree of freedom for locus related effects.

Multi-locus model and its parameter estimation in QTN detection As described in previous studies [19,20], all the potentially associated markers have been selected from genome-wide scanning. All the effects are placed into one multi-locus model. In the model, all the effects are estimated by EM empirical Bayes [23]. The linear model is as followed:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \sum_{i=1}^{s}\mathbf{Z}_i\beta_i + \boldsymbol{\varepsilon} \tag{7}$$

where $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\varepsilon}$ are the same as model (1); s is the number of the selected effects in the first step; β_i is genetic effects, and $\mathbf{Z}_i$ is marker genotypic incident vector for β_i.

In the model (7), all the priors and parameter estimation can be found in Li et al. [8], here we only list some important formulas:

1) Initial step: initialize parameters with:

$$\begin{aligned}\boldsymbol{\alpha} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ \sigma_e^2 &= \frac{1}{n}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right) \\ \sigma_i^2 &= \left[\left(\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right)\right]^2 + \left(\mathbf{Z}_i^T\mathbf{Z}_i\right)^{-1}\sigma_e^2\end{aligned} \tag{8}$$

2) E step: the effect vector can be predicted by:

$$E(\beta_i) = \sigma_i^2\mathbf{Z}_i^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}) \tag{9}$$

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

3) M step, we update parameters σ_i^2, $\boldsymbol{\alpha}$, and σ_e^2

$$\begin{aligned}\sigma_i^2 &= \frac{E\left(\beta_i^T\beta_i\right)+\omega}{\tau+3} \\ \boldsymbol{\alpha} &= \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ \sigma_e^2 &= \frac{1}{n}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})^T\left[\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}-\sum_{i=1}^{s}\mathbf{Z}_iE(\beta_i)\right]\end{aligned} \tag{10}$$

All the non-zero effects were further identified by likelihood ratio test for significant or suggested QTNs. Bonferroni correction was used to determine significant QTNs. To avoid the loss of important loci, the threshold of LOD score ≥ 3.0 was used to determine suggested QTNs [8].

2.4. Regression intercept and coefficient, environmental factor and variation indicators used as phenotype for identifying QEIs via 3VmrMLM-random and 3VmrMLM-fixed

A total of n accessions in the association mapping population were measured for complex traits across m environments, with these observations listed in Table 1. The regression of trait phenotypes y_{ij} of the ith accession in the jth environment ($i = 1, \ldots, n$;$j = 1, \ldots, m$) on environmental averages $\bar{y}_{\cdot j}$ was carried out [7,24], and regression coefficient (b_{1i}) and intercept (a_{1i}) were calculated based on regression model of Li et al. [8].

$$y_{ij} = a_{1i} + b_{1i}\bar{y}_{\cdot j} + e_{ij} \tag{11}$$

where e_{ij} is residual error. The estimates for b_{1i} and a_{1i} were used as phenotypes to conduct genome-wide association studies for indirectly identifying QEIs [7]. Specifically, in this study, we carried out the regression of trait phenotypes y_{ij} on environmental factor (e.g. temperature or other meteorological factors, treatment, and soil fertility), then regression coefficient (b_{2i}) was used as phenotype to indirectly identify QEIs.

Four variation indicators were calculated, such as $R_i = \max\{y_{i1}, \ldots, y_{im}\} - \min\{y_{i1}, \ldots, y_{im}\}$ for range, $V_i = \frac{1}{n-1}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2$ for variance, $SD_i = \sqrt{V_i}$ for SD, and $CV_i = \frac{SD_i}{\bar{y}_{i\cdot}} \times 100\%$ ($i = 1, \ldots, n$) for CV (Table 1). These variation indicator values were used as phenotypes to associate with all the SNP markers for identifying QEIs.

2.5. F_1 score

The F_1 score is a weighted average of precision and recall, ranging from 0 to 1. A higher score indicates a better model. It is defined as

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

where Precision = TP/(TP + FP), Recall = TP/(TP + FN), and TP is true positive, FP is false positive, and FN is false negative.

2.6. 3VmrMLM-MEJA

3VmrMLM-MEJA is an existing multi-locus GWAS method [8], in which QTN effect is treated as random. Bonferroni correction was used to determine significant QTNs and QEIs, while LOD score ≥ 3.0 was used to determine suggested QTNs and QEIs. All the trait phenotypes of all the n accessions in m environments were jointly analyzed to associate with markers for identifying QTNs and QEIs using 3VmrMLM-MEJA.

2.7. MLM via software GAPIT v3

MLM is an existing single-locus genome scan GWAS method, in which QTN effect is treated as fixed. Bonferroni correction was used to determine significant QTNs. Each variation indicator/environmental index was treated as a trait to associate with all the markers for identifying QEIs of complex trait using software GAPIT v3 [16] (https://github.com/jiabowang/GAPIT3).

2.8. EMMAX

EMMAX is an existing single-locus and fast genome scan GWAS method [15], in which QTN effect is treated as fixed. Bonferroni correction was used to determine significant QTNs. Each variation indicator/environmental index was treated as a trait to associate with all the SNPs for identifying QEIs using EMMAX (http://csg.sph.umich.edu//kang/emmax/download/index.html).

Table 1
Regression intercept (a_1), regression coefficient (RC, b_1), RC (b_2) of the trait phenotype on environmental factor (temperature, t), and environmental variation indicators calculated from multi-environment phenotype datasets.

Accession	Observations for complex trait in m environments				a_1	b_1	b_2	Variation indicators			
	1	2	⋯	m				Range	Variance	SD	CV
1	y_{11}	y_{12}	⋯	y_{1m}	a_{11}	b_{11}	b_{21}	R_1	V_1	SD_1	CV_1
2	y_{21}	y_{22}	⋯	y_{2m}	a_{12}	b_{12}	b_{22}	R_2	V_2	SD_2	CV_2
⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	y_{n1}	y_{n2}	⋯	y_{nm}	a_{1m}	b_{1m}	b_{2m}	R_m	V_m	SD_m	CV_m
Environmental factor	t_1	t_2	⋯	t_m	SD: standard deviation; CV: coefficient of variation						
Environmental average	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	⋯	$\bar{y}_{\cdot m}$							

2.9. Haplotype analysis for true GEI-phenotype causality

All the SNPs with P ≤ 0.1 for indirect indicators and MEJA in single-marker genome scanning within each known FT gene and its upstream 2.5 kb were extracted and used for haplotype analysis, including one-way ANOVA for all the eight indirect indicators and two-way ANOVA for FT. In ANOVA, 199 accessions were grouped based on gene haplotypes. The significance threshold was set at 5 % probability level.

3. Results

3.1. An effective tool for more comprehensive detection of QEIs

Since the establishment of 3VmrMLM [8], 3VmrMLM-MEJA has been used to detect and estimate additive and dominant effects, as well as additive-by-environment (*ae*) and dominant-by-environment (*de*) interaction effects, while controlling for all the polygenic backgrounds. In real data analysis, this method can only identify a subset of QEIs. To address this issue, IIIVmrMLM.QEI was developed to more effectively detect QEIs in multi-environment GWAS, as outlined in Fig. 1. This new tool consists of two modules. The first is the 3VmrMLM-random in Li et al. [8], which is integrated with all the eight indirect indicators, in which trait difference, RI, RC, range, Var, SD, CV, and EF are defined in Fig. 1, to indirectly identify QEIs, and the other is that our extended 3VmrMLM-fixed in this study, which is also integrated with all the above eight indirect indicators to indirectly identify QEIs. Around the above QEIs, known and candidate GEIs are mined via bioinformatics and multi-omics data analysis.

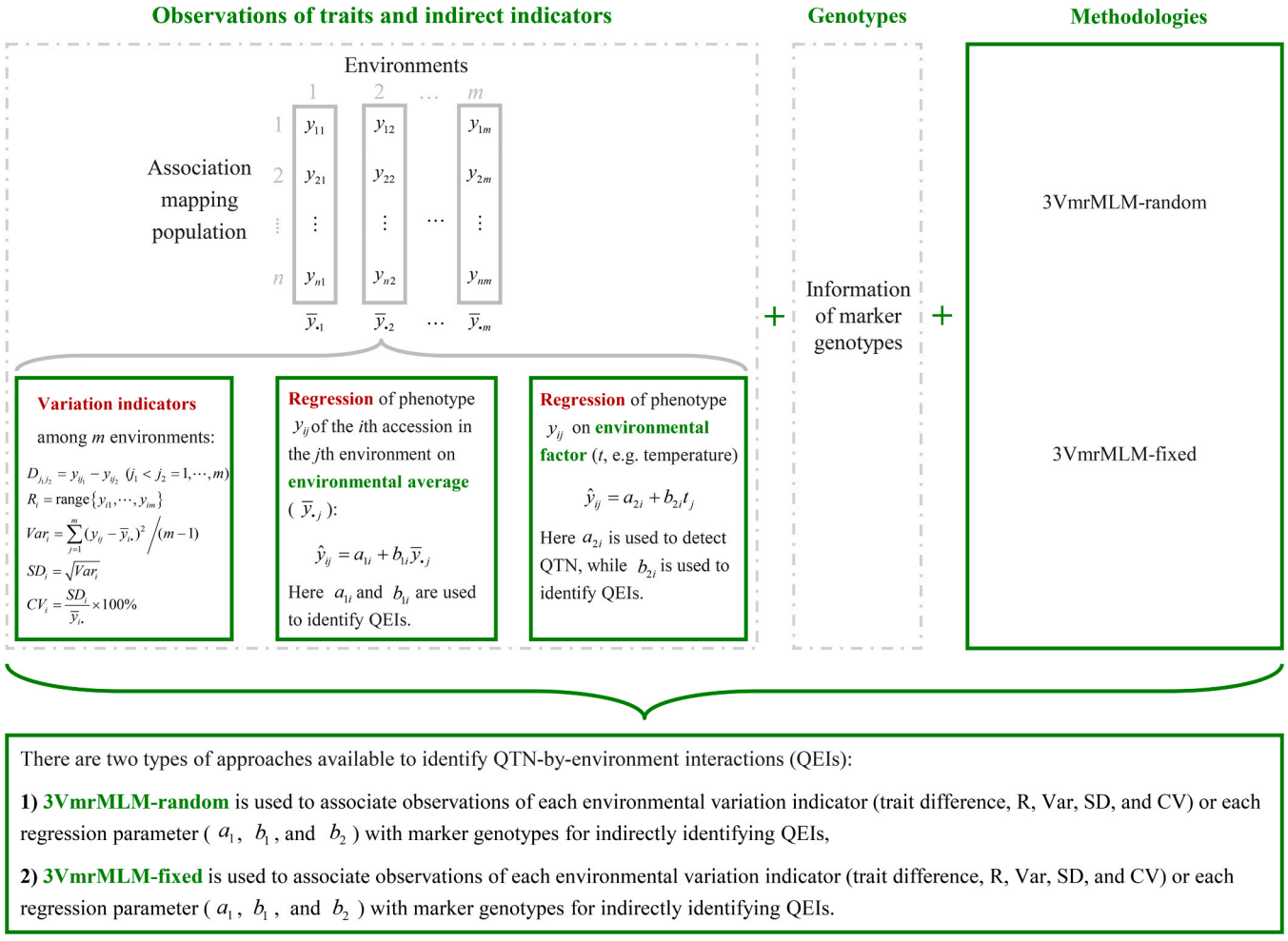

Fig. 1. The flow chart of the new tool IIIVmrMLM.QEI in the detection of QTN-by-environment interactions (QEIs) for complex traits in genome-wide association studies.

The software is written in R language. Before running it, we need to install the add-on packages and "IIIVmrMLM.QEI" (file A), in which the versions of add-on packages may be found in File A. When implementing it, we need to use the following two commands:

library (IIIVmrMLM.QEI).

IIIVmrMLM.QEI(fileGen="D:/Users/Genotype",filePhe="D:/Users/Phenotype.csv",fileKin=NULL,filePS="D:/Users/PopStr.csv", PopStrType="Q",fileCov=NULL,method=c("fixed","random"),indicator=c("difference","RI","RC","range","variance","SD","CV","EF"),trait= 1,n.en=c(3),SearchRadius= 80,svpal= 0.01,DrawPlot=FALSE,Plotformat= "jpeg", dir= "D:/Users/").

We also developed a user-friendly stand-alone tool with a Command Line Interface (CLI) that can be used via the command line or shell scripting as follows:

./run.IIIVmrMLM.QEI -dir /home/username/ -fileGen Genotype -filePhe Phenotype.csv -fileKin NULL -filePS PopStr.csv -PopStrType Q -fileCov NULL -method "c('random','fixed')" -trait "c(1)" -n.en "c(3)" -SearchRadius 80 -indicator "c('difference','RI','RC','range','variance','SD','CV','EF')" -DrawPlot FALSE -Plotformat "tiff" .

All the details of the parameters in the function are listed in file A.

3.2. Monte Carlo simulation studies

The Monte Carlo simulation studies in this study serve three purposes. The first is to confirm whether all seven indirect indicators except EF can be integrated with single-environment QTN detection methods to indirectly identify QEIs, the second is to confirm whether the newly extended 3VmrMLM-fixed are better in indirectly detecting QEIs than existing fixed-SNP-effect methods (MLM and EMMAX), and the last is to confirm whether 3VmrMLM-random and 3VmrMLM-fixed in the tool are better than 3VmrMLM-MEJA.

To demonstrate the first objective of the simulation studies, ten QTNs were simulated in three environments. Among the ten QTNs, the first three QTNs have not environmental interaction effects (EIEs), while the last seven QTNs have EIEs. All the phenotypes of each individual in three environments can be used to calculate the above seven indirect indicators, which are used to associate with marker genotypes using 3VmrMLM-random, 3VmrMLM-fixed, MLM, and EMMAX. Using 3VmrMLM-random at the significant probability level (Bonferroni correction), the average power for detecting QTNs with EIEs was notably high, ranging from 33.8 % ± 12.2 % to 75.9 % ± 13.2 %, while the power for QTNs without EIEs was near zero, indicating the capability of indicators to detect QEIs rather than QTNs (Table S1), notably, almost all detected QEIs matched their simulated positions. When estimating the effects of the last seven QTNs with EIEs, both additive and dominant effects were obtained with high accuracy and precision. Similar trends were observed with the other three methods (Tables S2-S8).

To demonstrate the second objective of the simulation studies, we compared the average powers of the last seven QTNs identified by 3VmrMLM-fixed, and two existing fixed-SNP-effect methods (MLM and EMMAX) at the significant probability level. The results showed that 3VmrMLM-fixed had superior average power across all seven indirect indicators (differences, RI, RC, range, Var, SD, and CV), with values ranging from 28.19 % to 68.8 %. In contrast, EMMAX and MLM exhibited considerably lower average powers, ranging from 5.59 % to 15.66 % for EMMAX and from 3.39 % to 9.27 % for MLM (Fig. 2A-2G;

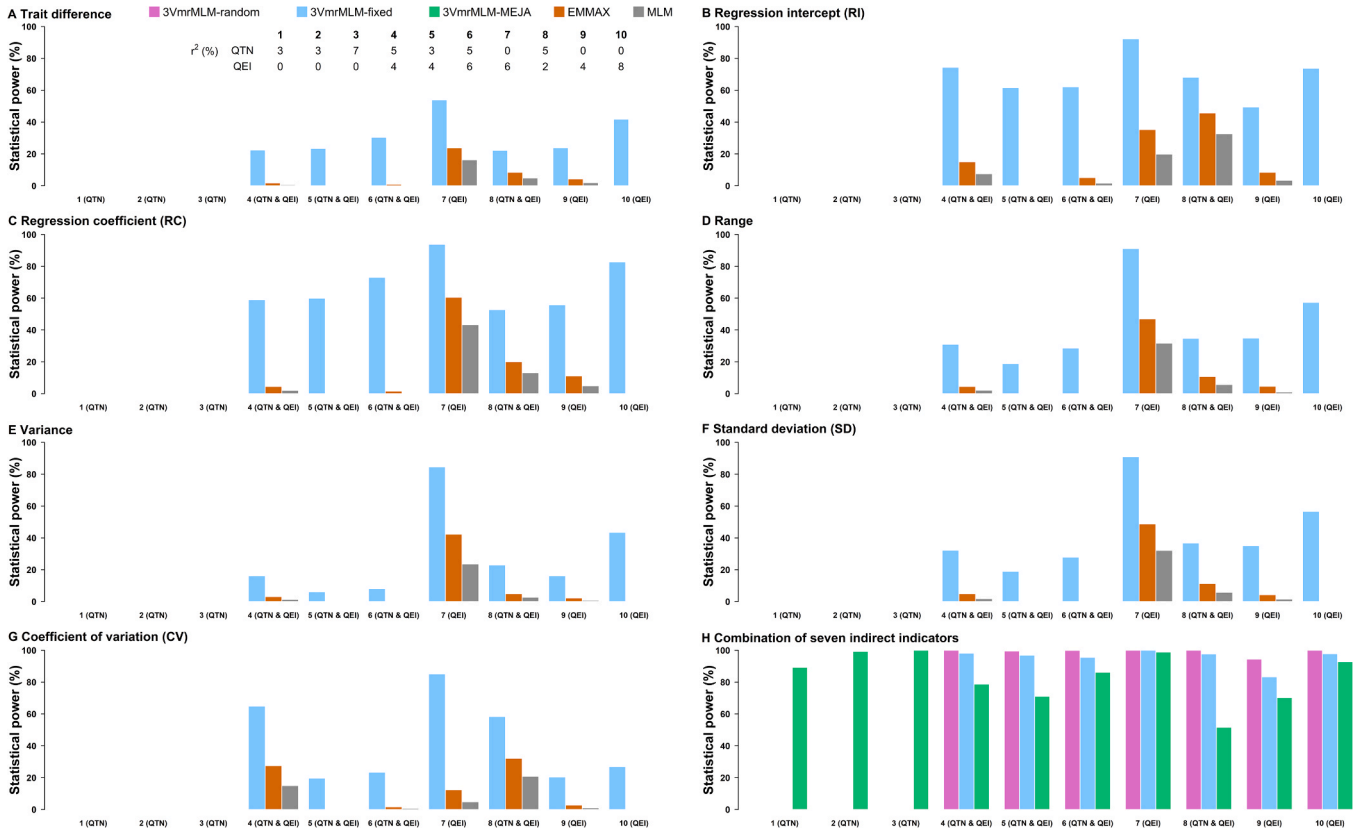

Fig. 2. The comparisons between the newly extended 3VmrMLM-fixed method and the existing fixed-SNP-effect methods (MLM and EMMAX) and between the 3VmrMLM-random and 3VmrMLM-fixed methods in the new tool and 3VmrMLM-MEJA in the detection of QTN and QTN-by-environment interactions (QEIs) in 200 individuals measured by the trait of interest in three environments. The numbers 1 to 10 represent QTNs, with the first three having no environmental interaction effects and the last seven having environmental interaction effects.
(A) Difference; (B) regression intercept; (C) regression coefficient; (D) range; (E) variance; (F) standard deviation; (G) coefficient of variation; and (H) the comparison of combination power between the 3VmrMLM-random and 3VmrMLM-fixed methods in the new tool and 3VmrMLM-MEJA.

Tables S2-S8). Moreover, the average FPRs for the indirect indicators were 1.10‰ to 2.66‰ for 3VmrMLM-fixed, 0.09‰ to 0.20‰ for EMMAX, and 0.01‰ to 0.08‰ for MLM (Tables S2-S8). Although 3VmrMLM-fixed had slightly higher FPR than the other two methods, all the FPRs are less than 3.00‰. To further evaluate the performance of these models, we examine the average powers under various *P*-values and FPR alongside F_1 scores in QEI detection across the above three methods, offering a more comprehensive assessment of its effectiveness.

As the *P*-value threshold for significant QEIs was changed, the corresponding power and FPR were obtained. Two types of curves were generated: power versus *P*-value threshold (Fig. 3A) and power versus observed FPR (Fig. 3B). As shown in Fig. 3A-3B, 3VmrMLM-fixed surpassed both MLM and EMMAX. This conclusion was also corroborated by the F_1 score (Fig. 3C), confirming its improved efficacy. To confirm the robustness of the newly extended 3VmrMLM-fixed in another case scenario, we added an additional simulation experiment with 1000 individuals each with one million markers (20 replicates) to compare the three tools. As a result, the F_1 score was 0.8861 for 3VmrMLM-fixed, 0.8755 for EMMAX, and 0.8235 for MLM (Table S9), validating its robustness.

To demonstrate the third objective of the simulation studies, all the significant QTNs with EIEs for all the above seven indirect indicators identified by 3VmrMLM-random or 3VmrMLM-fixed at the significant probability level were summarized, and the average combination powers of all the seven indicators among the last seven QTNs for each method were calculated. As a result, 3VmrMLM-random (99.1 %) and 3VmrMLM-fixed (95.6 %) had significantly higher powers to detect QEIs than 3VmrMLM-MEJA (78.5 %) (Fig. 2H; Table S10). The average SDs for additive and dominant effect estimates of significant QEIs showed the high accuracy of 3VmrMLM-random, 3VmrMLM-fixed, and 3VmrMLM-MEJA (Table S10). The average FPRs for all the indirect indicators were 1.70 ± 0.62 (‰) for 3VmrMLM-random, 1.38 ± 0.58 (‰) for 3VmrMLM-fixed, and 1.16 ± 0.17 (‰) for 3VmrMLM-MEJA, indicating the effectiveness of controlling FPRs for all the three methods (Table S10).

3.3. *Identification of GEIs for flowering time in Arabidopsis under three temperatures using the new tool IIIVmrMLM.QEI*

To illustrate the primary objective of the simulation studies, in which all indirect indicators for QEI detection are feasible, the observations of *Arabidopsis* flowering time (FT) of each line under 16 h of daylight at 10 ℃ (Env1), 16 ℃ (Env2), and 22 ℃ (Env3) in Atwell et al. [17] were used to calculate trait differences (Env1–Env2, Env1–Env3, and Env2–Env3), RI, RC, range, Var, SD, and CV, and all of them, along with the RC of flowering time on temperature (EF: 10 ℃, 16 ℃, and 22 ℃), were used as phenotypes to indirectly identify QEIs (Fig. 1). All the QEIs identified in this study are summarized in file B. Within 200 Kb around these QEIs, known and candidate GEIs were identified, with results presented in Table 2 (numbers of QEIs, genes and GEIs) and 3 (known GEIs) and Tables S11 (haplotype analysis by one-way ANOVA), S12 (haplotype analysis by two-way ANOVA), S13 and S15 (known GEIs), and S14 and S16 (candidate GEIs).

Using Var as an example, 17 and 15 significant QEIs were identified by 3VmrMLM-random and 3VmrMLM-fixed, respectively, to be significantly associated with Var, while 4 and 2 suggested QEIs were identified by 3VmrMLM-random and 3VmrMLM-fixed, respectively, to be associated with Var (Table 2; file B: Tables 4.4 and 5.4). Using the two

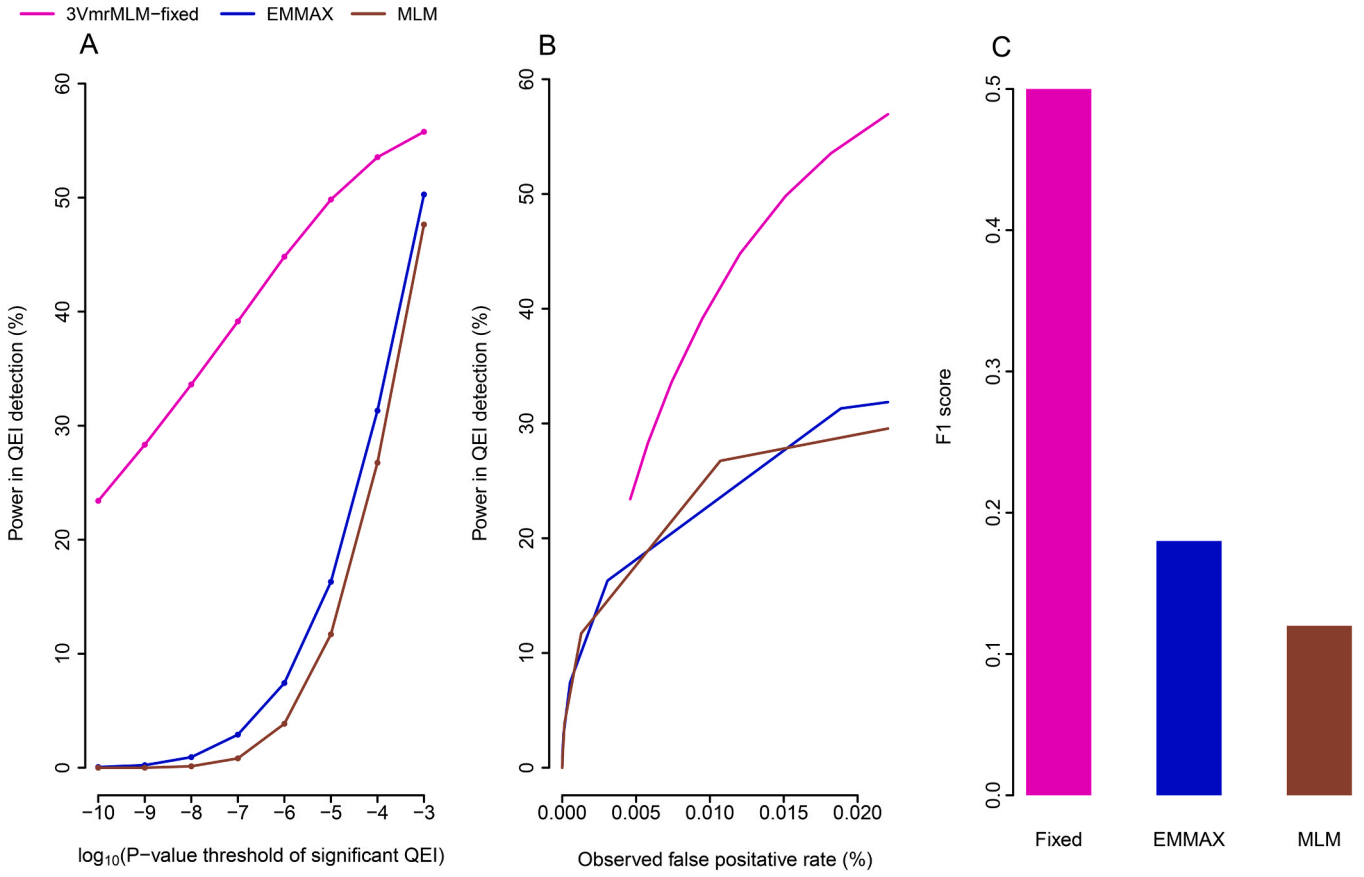

Fig. 3. Average powers at different *P*-value thresholds and observed FPRs alongside the F_1 scores in QEI detection across different indirect indicators using 3VmrMLM-fixed, EMMAX, and MLM.
(A) The x-axis indicates P-value threshold of significant QEI. (B) The x-axis indicates observed FPRs (%). The y-axis shows average power in (A) and (B) and the F1 score in (C).

Table 2
Comparison of the numbers of QEIs, known genes/GEIs, and candidate GEIs from 3VmrMLM-MEJA (MEJA), 3VmrMLM-random (random), 3VmrMLM-fixed (fixed), MLM, and EMMAX.

Indicator	No. of QEIs					No. of known genes					No. of known GEIs					No. of candidate GEIs				
	MEJA	Random	Fixed	MLM	EMMAX	MEJA	Random	Fixed	MLM	EMMAX	MEJA	Random	Fixed	MLM	EMMAX	MEJA	Random	Fixed	MLM	EMMAX
Env1–Env2		19 (16/3)	16 (11/5)	0	0		26 (22/4)	13 (9/4)	0	0		5 (4/1); 7 (5/2); 4(3/1)	2 (1/1); 7 (4/3); 2 (1/1)	0	0		5 (5/0)	5 (4/1)	0	0
Env1–Env3		28 (27/1)	10 (5/5)	0	0		29 (28/1)	12 (7/5)	0	0		3 (2/1); 2 (2/0); 1 (1/0)	3(2/1); 1 (0/1); 1 (0/1)	0	0		6 (6/0)	2 (1/1)	0	0
Env2–Env3		20 (17/3)	19 (17/2)	2	5		17 (13/4)	10 (9/1)	2	2		4(4/0); 3 (3/0); 1 (1/0)	0; 1 (1/0); 0	1;0;0	1;0;0		2 (2/0)	1 (1/0)	0	0
RI		19 (14/5)	20 (17/3)	2	3		21 (16/5)	30 (28/2)	2	2		3(2/1); 3 (3/0); 1 (1/0)	8 (8/0); 8 (8/0); 4 (4/0)	1;0;0	1;0;0		7 (5/2)	4 (4/0)	0	0
RC		24 (22/2)	18 (13/5)	1	3		15 (15/0)	15 (8/7)	2	2		6(6/0); 2 (2/0); 2 (2/0)	4 (2/2); 4 (3/1); 2 (1/1)	1;0;0	1;0;0		3 (3/0)	3 (1/2)	0	0
Range		16 (14/2)	12 (10/2)	0	2		18 (15/3)	16 (12/4)	0	4		3 (2/1); 3 (1/2); 2 (1/1)	6 (5/1); 1 (1/0); 1 (1/0)	0	1;2;1		6 (6/0)	3 (2/1)	0	1
Variance		21 (17/4)	17 (15/2)	1	3		25 (23/2)	21 (20/1)	2	4		4 (3/1); 4 (3/1); 2 (2/0)	5 (5/0); 6 (6/0); 4 (4/0)	1;1;1	1;1;1		5 (5/0)	3 (3/0)	0	1
SD		18 (15/3)	11 (9/2)	0	2		16 (13/3)	16 (10/6)	0	4		3 (3/0); 4 (3/1); 2 (2/0)	6 (4/2); 2 (1/1); 2 (1/1)	0	1;2;1		4 (4/0)	2 (2/0)	0	1
CV		25 (22/3)	12 (7/5)	0	0		22 (21/1)	20 (13/7)	0	0		2 (2/0); 1 (1/0); 1 (1/0)	5 (3/2); 1 (1/0); 1(1/0)	0	0		4 (4/0)	5 (4/1)	0	0
EF		28 (27/1)	10 (5/5)	0	0		29 (28/1)	12 (7/5)	0	0		3 (2/1); 2 (2/0); 1 (1/0)	3 (2/1); 1 (0/1); 1 (0/1)	0	0		6 (6/0)	2 (1/1)	0	0
Total (unique)	29 (22/7)	141	97	2	6	22 (19/3)	113	86	2	4	6 (5/1); 3 (2/1)	15	12	1	1	4 (3/1)	26	18	0	1

Env1, Env2, and Env3 indicated 10, 16, and 22 (℃), respectively. In the three columns for the numbers of QEIs, known genes and candidate GEIs, such as in "19 (16/3)", 16 and 3 is the number of significant and suggested QEIs, respectively, and 19 is their sum. In the column for "No. of known GEIs", such as in "6 (5/1); 3 (2/1)", the meanings of "3 (2/1)" are similar to those in "19 (16/3)", in which the haplotypes of these genes are significant for flowering time, but in "6(5/1)" the haplotypes of these genes are significant for flowering time or indirect indicators, while in "5 (4/1); 7 (5/2); 4(3/1)", the meanings of "5(4/1)" are similar to those in "6 (5/1)", and the meanings of "4(3/1)" are similar to those in "3 (2/1)", in which haplotype analysis was significant in one-way and two-way ANOVA, but in "7 (5/2)" haplotype analysis was significant only in one-way ANOVA.

methods, 23 and 20 known FT genes were found around these significant QEIs, respectively, while 2 and 1 known FT genes were found around these suggested QEIs, respectively (Table 2). To confirm whether these known genes are interacted with temperature change, the above known FT genes were used to perform haplotype analysis by one-way (gene haplotypes) ANOVA for indirect indicator (Var) and two-way (gene haplotype and environment) ANOVA for flowering time. In one-way ANOVA, three and six known FT genes around significant QEIs from 3VmrMLM-random and 3VmrMLM-fixed, respectively, were found to be significant, while one known FT gene around a suggested QEI from 3VmrMLM-random was found to be significant (Table S11). These known FT genes were further analyzed by two-way ANOVA. As a result, two and four known FT genes around significant QEIs from 3VmrMLM-random and 3VmrMLM-fixed, respectively, were found to be significant (Table S12). Notably, these known genes were found to be directly / indirectly associated with temperature change in previous studies, and the relevant evidence is presented in Table 3. Thus, the above two and four known FT genes are considered as known FT GEIs. Within 50 Kb around these QEIs, the remaining known FT genes were used to mine candidate FT GEIs, as a result, five and three known FT genes from 3VmrMLM-random and 3VmrMLM-fixed, respectively, were considered as candidate FT GEIs, pending further experimental validation to explore these novel GEI-FT associations (Table 2 and S14).

In summary, fourteen significant GEIs identified from indirect indicators were directly associated with temperature variation in previous studies, including *CSTF77*, *NF-YC9*, *PIE1*, *SPA3*, *VRN1*, *GA1*, *TSF*, *JMJ14*, *MOS1*, *FLC*, *MIR156D*, *MIR156E*, *COL5*, and *MAF4*, while *GA3OX2*, *LIF2*, *BBX19*, and *FPF1* were found to be indirectly associated with temperature variation in previous studies (Fig. 4A; Table 3, S13 and S15). In addition, 39 candidate GEIs such as *GRP2*, *FRI*, *SPL5*, and *TOE3* were mined (Tables S14 and S16).

To demonstrate the second objective of the simulation studies, in which the extended 3VmrMLM-fixed is better than existing fixed-SNP-effect methods (MLM and EMMAX), all the eight indirect indicators were used as phenotypes to indirectly identify QEIs using EMMAX and MLM in the above FT datasets. In summary, 6 and 2 significant QEIs were identified by EMMAX and MLM, respectively (Table 2; file B: Tables B6-B7), and one and one known GEIs around significant QEIs from EMMAX and MLM were found to be truly associated with FT (Tables 2, 3, S13 and S15; Fig. 4A). In addition, one candidate GEI was found around significant QEI from EMMAX (Tables S14 and S16). Clearly, the newly extended 3VmrMLM-fixed (97 QEIs and 12 known and 18 candidate GEIs) identifies many more QEIs and known/candidate GEIs than EMMAX (six QEIs and one known and one candidate GEIs) and MLM (two QEIs and one known GEIs) (Table 2).

To further demonstrate the third objective of the simulation studies, in which 3VmrMLM-random and 3VmrMLM-fixed are better than 3VmrMLM-MEJA, all the above trait phenotypes under three temperatures were jointly used to associate with 216130 SNP markers using 3VmrMLM-MEJA. As a result, 22 significant and 7 suggested QEIs were identified (Table 2; file B: Table B1), and 19 (3) known FT genes were found to be located around significant (suggested) QEIs (Table 2). To confirm whether these known genes are interacted with temperature change, the above known FT genes were used to perform haplotype analysis by two-way ANOVA for flowering times. As a result, two and one genes were significant around significant and suggested QEIs, respectively (Table 2; Fig. S1; Tables S12-S13). All the three significant GEIs were found to be directly associated with temperature change in previous studies, including *SPA3*, *MOS1*, and *MIR156D* (Table 3; Fig. 4B). Therefore, these known genes were considered as known FT GEIs. In addition, the other 4 known FT genes were considered as candidate FT GEIs in the same way, including *ARP4*, *SPL5*, *AGL24*, and *GRP2*, pending further experimental validation to explore these novel GEI-FT associations (Table 2 and Table S14).

Based on the above results, 3VmrMLM-random identified 141 QEIs along with 15 known and 26 candidate GEIs, while 3VmrMLM-fixed identified 97 QEIs as well as 12 known and 18 candidate GEIs. Therefore, both 3VmrMLM-random and 3VmrMLM-fixed identify substantially more QEIs and known/candidate GEIs than 3VmrMLM-MEJA (29 QEIs and 6 known and 4 candidate GEIs) (Table 2). More importantly, although 3VmrMLM-random can identify numerous QEIs and GEIs, it may still be incomplete, as some known FT GEIs, such as *GA3OX2*, *PIE1*, and *GA1*, were not identified by 3VmrMLM-random (Table 3; Tables S13 and S15). Therefore, the new tool, IIIVmrMLM.QEI, developed in this study, provides a more comprehensive approach to detecting QEIs in multi-environment GWAS.

4. Discussion

The genetic variations among individuals are derived from both recombinant in bi-parental segregation populations for linkage analysis and historical recombinant in association mapping populations for GWAS. Hence, the genetic basis for identifying genetic loci is recombinant for linkage analysis and linkage disequilibrium for GWAS, and recombinant-based linkage analysis has a higher repetition rate in detecting genetic loci across different models, indirect indicators, and environments than linkage disequilibrium-based GWAS [25–28]. Therefore, it is essential to extend models and indirect indicators in GWAS in this study.

In GWAS, many software packages were available to identify QTNs associated with complex traits [29]. Although 3VmrMLM-MEJA without random mating assumption was established to identify QTNs and QEIs with high power, the number of known and candidate GEIs was relatively limited in real data analysis, e.g., 3VmrMLM-MEJA (six known and four candidate GEIs) identified fewer known and candidate GEIs than IIIVmrMLM.QEI (18 known and 38 candidate GEIs) (Tables 2 and 3; Fig. 4; Fig. S1; Tables S13-S16). Meanwhile, 3VmrMLM-random-D (2), 3VmrMLM-random-VI (1), and 3VmrMLM-fixed-VI (1) in IIIVmrMLM.QEI can detect some method-specific known GEIs, and 5, 6, and 3 known GEIs were simultaneously identified by two, three, and more than four methods, respectively, demonstrating the repeatability and specificity between different methods (Fig. S2). The new tool provides a more comprehensive framework to identify QEIs and mines known and candidate GEIs, indicating the necessity of developing the new tool in this study. More importantly, multiple lines of evidence were provided to confirm the true GEI-phenotype causality in this study. In detail, thirty-four known FT genes were confirmed by haplotype analysis for indirect indicators, eighteen GEIs were further validated by two-way ANOVA (Tables S11 and S12), and all the eighteen significant GEIs were found to be directly (14) or indirectly (4) associated with temperature change in previous studies (Table 3).

Significant advances in methods and applications have been achieved in this study. First, 3VmrMLM-random was extended to 3VmrMLM-fixed for the first time. Unlike in Li et al. [8], where these methods were used to identify QTNs, we employed them to indirectly detect QEIs. As a result, when compared with 3VmrMLM-random, 3VmrMLM-fixed can identify three method-specific GEIs in real data analysis. Furthermore, 3VmrMLM-fixed outperformed MLM and EMMAX in identifying QEIs in both real and simulated datasets. This improvement is attributed to the fact that additive and dominant effects and their polygenic backgrounds are included in our mixed model in this study. In other words, all the possible effects are detected and estimated conditional on controlling for all the possible polygenic backgrounds in this study, while only one confounding effect and its polygenic background are considered in MLM and EMMAX. This conclusion is consistent with the quantitative trait locus mapping in Zhou et al. [30].

Second, 3VmrMLM-random and 3VmrMLM-fixed were applied to analyze trait difference, RC, RI, and extended EF and VIs to indirectly identify QEIs for the first time. We also compared 3VmrMLM-random and 3VmrMLM-fixed with mrMLM in detecting QEIs via indirect indicators. As a result, 3VmrMLM-random (38.9 %, 81.0 %, 80.2 %, 54.7 %, 44.7 %, 54.6 %, and 57.7 %) and 3VmrMLM-fixed (36.6 %,

Table 3
Eighteen known gene-by-environment interactions (GEIs) around significant and suggested QTN-by-environment interactions (QEIs) for flowering time (FT) using 3VmrMLM-random, 3VmrMLM-fixed, 3VmrMLM-MEJA, EMMAX, and MLM.

No.	QEI				Method/ indicator	Known genes	Evidence for GEIs				
	Chr	Posi (bp)	LOD	r^2 (%)			*P*-value1 (Indicator)	*P*-value2 (FT)	Environment	Indicator (or trait)	Differences of indicators under various environments
1	1	2907566 ~ 2907566	10.12 ~ 26.16	1.58 ~ 4.79	a ~ b/1,5	*NF-YC9*	6.45E−09 * **	7.41E−07 * *	Vernalization	Flowering time	Vernalization: WT-*nf-yc* < Non-vernalization: WT-*nf-yc*.
2	1	5916352 ~ 6106169	16.60 ~ 28.00	1.81 ~ 2.30	a/4,7	*CSTF77*	1.35E−07 * ** ~ 1.44E−04 * **	7.30E−05 * ** ~ 2.16E−03 * *	Elevated temperature	Expression	The activity of *CSTF77* is only compromised under elevated temperatures.
3	1	30201816	15.97	1.60	b/5	*GA3OX2*	1.35E−03 * *	4.01E−02 *	Temperature	Expression	At 16 °C, the level of *TEM* keeps higher and longer than those at 22 °C, and the expression of *GA3OX2* is regulated by *TEM*.
4	3	3867466 ~ 3943409	4.04 ~ 11.12	0.39 ~ 1.96	b ~ c/ 1,2,5	*PIE1*	2.07E−07 * ** ~ 5.17E−04 * **	2.92E−04 * ** ~ 3.27E−03 * *	Elevated temperature	Hypocotyl/ Petiole length	While hypocotyl/ petiole elongation and early flowering in the Col at warm temperatures, *pie1* mutants show reduced elongation response to higher temperature.
5	3	5102027 ~ 5264482	4.03 ~ 8.20	0.17 ~ 2.60	a ~ c/2 ~ 4,7	*SPA3*	6.60E−06 * ** ~ 7.53E−04 * **	2.15E−04 * ** ~ 1.46E−03 * *	Temperature	Primary root length	Primary root length: 28 °C WT-*spa123* > 22 °C WT-*spa123*.
6	3	6485151 ~ 6485151	15.30 ~ 16.46	1.71 ~ 2.04	a ~ b/1	*VRN1*	2.87E−04 * **	4.14E−02 *	Vernalization	*FLC* mRNA levels	Unlike *fca−1*, *FLC* levels in *vrn1 −1fca−1* increased upon return to normal temperatures, highlighting *VRN1*'s role in stably maintaining *FLC* repression during later development in warm temperatures.
7	4	227242 ~ 310657	8.58 ~ 24.08	1.60 ~ 5.49	a ~ b/ 3,4,7	*LIF2*	4.54E−03 * *	3.35E−02 *	Vernalization	Flowering time	Vernalization: *lhp1/ FRI*-Col - *FRI*-Col > Non-Vernalization: *lhp1/FRI*-Col - *FRI*-Col, *LIF2* primarily functions downstream of *LHP1*.
8	4	1261211 ~ 1271920	9.99 ~ 25.18	3.29 ~ 6.56	b/1,4,6 ~ 8	*GA1*	2.26E−04 * **	2.65E−02 *	Temperature	The accumulation of GFP-RGA	29 °C: WT-*ga1 −3* > 20 °C WT-*ga1 −3*
9	4	10804201 ~ 11166402	5.45 ~ 6.86	0.41 ~ 0.63	a/1,8	*TSF*	7.51E−07 * **	1.95E−03 * *	Vernalization	mRNA levels	In *fca−1* and *FRI*-Sf2 (Col), *TSF* mRNA levels were reduced. Vernalization treatment restored *TSF* expression in *fca−1* and *FRI*-Sf2 (Col).
10	4	11166402	6.86	0.63	a/1	*JMJ14*	3.83E−08 * **	8.66E−05 * **	Temperature	Expression of target gene	Seven selected genes were upregulated in Col−0 at 27 °C, but downregulated in the *icdh−1/jmj14 −1* mutant under the same conditions compared to Col−0.
11	4	12628000 ~ 12635782	8.90 ~ 12.53	0.44 ~ 0.98	a, c/1	*MOS1*	4.00E−14 * **	5.25E−08 * **	Cold treatment	Expression of target gene	In both *rpp4 −1* and *rpp4 −1mos1*, the expression of *PR1* increased with cold

(*continued on next page*)

Table 3 (*continued*)

No.	QEI				Method/ indicator	Known genes	Evidence for GEIs				
	Chr	Posi (bp)	LOD	r^2 (%)			*P*-value1 (Indicator)	*P*-value2 (FT)	Environment	Indicator (or trait)	Differences of indicators under various environments
											treatment, but in *rpp4 −1mos1* it did not rise as much as *rpp4 −1*.
12	4	18067898	3.45 ~ 11.10	0.73 ~ 2.50	a ~ c/1 ~ 3,5,8	*BBX19*	4.72E−04 * **	3.02E−02 *	High temperature	Flowering time	Under high temperature, the level of plant endogenous hormone GAs is downregulated, while the delayed flowering in of 35GS:*BBX19* is inhibited by GA.
13	5	3055565 ~ 3301110	3.95 ~ 22.36	0.48 ~ 3.36	a ~ e/1 ~ 8	*FLC*	1.42E−08 * ** ~ 2.29E−02 *	3.07E−04 * **	Temperature	Expression	The decrease in *FLC* expression between 9 and 13 DAG: 29 °C < 22 °C.
14	5	3284990 ~ 3585818	3.95 ~ 19.78	0.51 ~ 3.36	a ~ c/1 ~ 3,6 ~ 8	*MIR156D*	1.65E−06 * ** ~ 5.93E−05 * **	3.59E−04 * **	Heat stress	Expression	*miR156* isoforms are highly induced after heat stress, and *miR156* may avoid flowering during a hot spell.
15	5	3873896 ~ 3911823	4.39 ~ 4.70	0.70 ~ 0.96	a/1,5	*MIR156E*	2.62E−04 * ** ~ 5.67E−04 * **	1.46E−02 *	Heat stress	Expression	*miR156* isoforms are highly induced after heat stress, and *miR156* may avoid flowering during a hot spell.
16	5	8508968 ~ 8609857	4.96 ~ 31.15	0.71 ~ 4.91	a ~ b/2 ~ 4,6	*FPF1*	1.13E−08 * ** ~ 1.46E−07 * **	3.22E−06 * ** ~ 5.35E−05 * **	High temperature	Expression level	Under high temperature, the level of plant endogenous hormone GA is inhibited, while the expression of *FPF1* is increased by GA.
17	5	23373062 ~ 23490084	4.94 ~ 18.64	1.52 ~ 2.26	a ~ b/ 2,5,6	*COL5*	1.70E−05 * ** ~ 2.36E−05 * **	1.26E−02 *	Low Temperature	The activity of proteins	The activity of COL proteins (COL1/2/3/ 5) is repressed by TOE proteins at low temperatures, which leads to the repression of *FT* expression.
18	5	26137944 ~ 26137944	7.61 ~ 12.05	1.64 ~ 2.15	a/2,5	*MAF4*	1.36E−13 * ** ~ 4.66E−11 * **	2.19E−10 * ** ~ 2.54E−09 * **	Cold treatment	mRNA expression levels	In the early stages of cold exposure, the mRNA expression levels of *MAF4* are significantly increased.

3VmrMLM-random, 3VmrMLM-fixed, 3VmrMLM-MEJA, MLM in GAPIT, and EMMAX were indicated by a to e, respectively; indirect indicators trait difference, regression intercept, regression coefficient, range, variance, standard deviation (SD), coefficient of variation (CV), and RC of flowering time on temperature (EF) were indicated by 1 to 8, respectively. The known GEIs in Supplementary Tables 12 and 14 were summarized in this table, where the QEIs were identified by one of methods or indicators, and their GEIs were confirmed in haplotype analysis using one- (indirect indicators) and two-way (environments and flowering time) ANOVA. *P*-value1 (indicator) and *P*-value2 (FT): The probabilities of haplotype analysis for indirect indicator in one-way ANOVA and for flowering time in two-way ANOVA, respectively. All the references were cited in Supplementary Table S12 and S14.

73.6 %, 73.0 %, 51.4 %, 37.8 %, 51.8 %, and 54.9 %) exhibited higher average powers for difference, RI, RC, range, variance, SD, and CV than mrMLM (23.5 %, 56.1 %, 48.0 %, 36.9 %, 31.9 %, 36.1 %, and 44.9 %) (Table S17), which were observed in Li et al. [8,29], further confirming the above reason. In conclusion, combining diverse indirect indicators that capture phenotypic changes across environments with optimal QTN detection methods (3VmrMLM-random and 3VmrMLM-fixed) enhances the accuracy of indirect QEI detection.

In our 3VmrMLM method, the nominal FPR at the genome-wide level was set at 0.05, and the *P*-value threshold of significant loci was set at $0.05/m$ based on the widely used Bonferroni correction. The Bonferroni correction is known to be too stringent in crop GWAS [25]. To mitigate the risk of missing important loci, an additional threshold for suggested loci was applied. In our previous multi-locus GWAS methodologies, the LOD score of 3.0 was deemed an effective threshold [26]. Therefore, this standard was also adopted in this study. Once we obtain more significant and suggested QEIs, multi-omics, bioinformatics, haplotype analysis, and ATAC-seq dataset analysis can be employed to mine candidate functional genes. In this case, we focus on the QEIs with some important candidate functional genes, while all the significant and suggested loci contribute to enhancing the accuracy of genomic selection [31–33].

When there are more environments, more general indicators of environmental variation are more appropriate. In the real data analysis, 2 known and 15 candidate GEIs were exclusively identified by the four environmental VIs (Table 3, S13 and S16). Therefore, in this study, the trait difference was extended to four environmental VIs in this study.

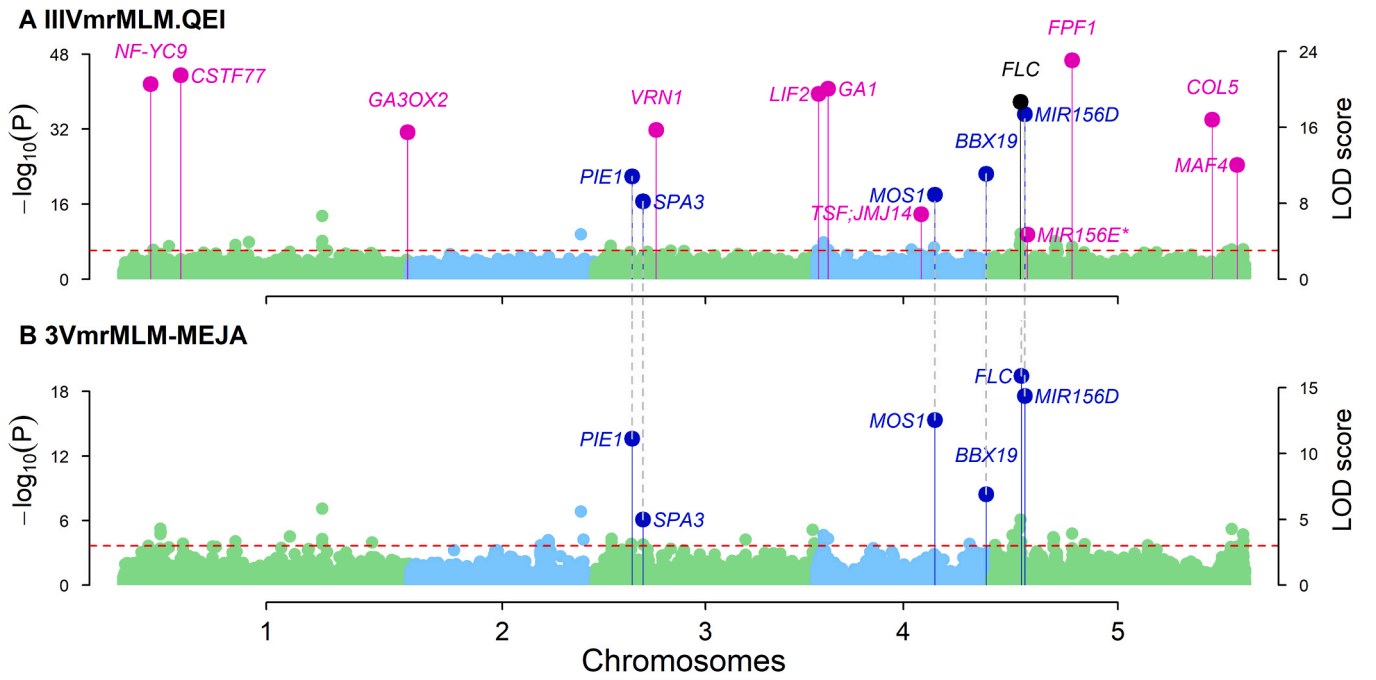

Fig. 4. Manhattan plots of identifying QTN-by-environment interactions (QEIs) for flowering time in 199 *Arabidopsis thaliana* lines using the new tool IIIVmrMLM.QEI.

(A) All the QEIs and their associated gene-by-environment interactions (GEIs) for flowering time and (B) QEIs identified only by multi-environment joint analysis (3VmrMLM-MEJA) and their associated GEIs. The left y axis represents the medians of −log10(P) values across all the approaches, which are obtained from single-marker genome-wide scanning in the first step, and the right *y* axis represents LOD scores, which are obtained from a likelihood ratio test for QEIs in the second step. The threshold of LOD score for suggested QEIs was 3.0 (red dashed line) in likelihood ratio test. If LOD score ≥ 15, the LOD scores are transformed as LOD=15+(LOD−15)/15. These LOD scores, along with their known GEIs, are shown in points with straight lines. All the GEIs, identified by one (pink), at least two indicators/methods (blue), and also detected by MLM/EMMAX (black), are marked with different colors. In addition, the GEIs with asterisk are around suggested QEIs.

Additionally, 3VmrMLM-random and 3VmrMLM-fixed were applied for the first time to analyze the four environmental VIs. Although VI-based methods were inferior to regression parameter-based methods in simulation studies, the former can identify some method-specific GEIs compared to the latter in real data analysis, e.g., 3VmrMLM-random-VI identified a specific GEI and 3VmrMLM-fixed-VI identified a specific GEI to be truly associated with flowering time (Table 3), indicating the fact that these new methods can be used to not only demonstrate the repeatability of QEI detection, but also to identify new QEIs. Hence, extending these indicators was deemed essential in this study.

In this study, we also conducted regression analysis of *Arabidopsis* flowering time on environmental temperatures (10 ℃, 16 ℃, and 22 ℃), and the regression coefficients were used as phenotypes to identify QEIs. Environmental factors (EF), encompassing specific meteorological factors (e.g. temperature), treatment levels, and soil fertility levels, plays an important role in studying environmental plasticity and predicting crop yields. Compared with RI and RC, EF detected 2 known and 2 candidate method-specific GEIs (Tables S13 and S14), more importantly, the year/location environments are expanded to include meteorological factors, treatment levels, soil fertilizer content, and other environmental conditions, indicating the need to analyze EF by the three methods in a new tool.

The newly extended 3VmrMLM-fixed and all the indirect indicators provide a robust theoretical framework for identifying QEIs. First, the statistical foundation of 3VmrMLM-fixed is detailed in the Materials and Methods section, with supporting evidence presented in the Results section. Second, trait differences and regression parameters serve as phenotypes to indirectly identify QEIs, as described in Korte et al. [1], Zan and Carlborg [3], Kerin and Marchini [6], Li et al. [7], and Fu and Wang [34]. Finally, although QTNs and QEIs were simulated in Monte Carlo simulation studies, only QEIs were detected by all the indirect indicators as phenotypes (Tables S1-S9), providing solid evidence for this method. In addition, we calculated correlation coefficients of previously proposed indirect indicators (trait difference, RI, and RC) with newly proposed indirect indicators (range, Var, SD, CV and EF). As a result, very significant correlations (Table S18) support the detection of QEIs by the new indirect indicators in this study.

In almost all the existing GWAS methods, marker genotypes are coded as 0, 1, and 2, the effect to be detected is the allelic substitution effect, and the polygenic background to be controlled is also related to the allelic substitution effect. In this case, association mapping population is assumed to be randomly mated [35]. However, this assumption does not exist in crop association mapping populations. 3VmrMLM-related methods do not rely on this assumption. Thus, this tool based on 3VmrMLM-related methods is expected to be more applicable in future GWAS studies for animals, plants, and humans. In particular, the Multi_env module of software IIIVmrMLM [29] can be used to identify QEIs in human genetics, although there are no same individuals in different environments [8]. In addition, if different individuals of each mice family are measured under multiple environments for trait phenotypes, IIIVmrMLM.QEI in this study is feasible to identify QEIs in human genetics.

5. Conclusion

To comprehensively detect QEIs, trait phenotype difference, regression parameters (RI and RC), variation indicators (range, variance, standard deviation, and coefficient of variation), and environmental factors (EF) were first integrated with previous 3VmrMLM-random and extended 3VmrMLM-fixed to develop an effective tool IIIVmrMLM.QEI for QEI detection. The new tool (18), validated by Monte Carlo simulation studies, identified much more known GEIs for flowering time than multi-environment joint analysis (6).

Funding

This work was supported by the National Natural Science Foundation of China (32070557; 32200500; 32270673; 32470657), Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation (2014RC020), Natural Science Foundation of Hubei Province (2022CFB780), and Postdoctoral Innovative Research Position of Hubei Province. We thank Mr. Hanwen Zhang (hywenzhang@henceedu.com; Hence Education Ltd., Vancouver, Canada) for improving the language within the manuscript.

CRediT authorship contribution statement

Yuan-Ming Zhang: Writing – review & editing, Supervision, Methodology. **Chen Ying:** Data curation. **Mei Li:** Data curation. **Xue-Lian Han:** Data curation. **Ya-Wen Zhang:** Writing – original draft, Software, Methodology.

Conflict of interest

We are pleased to submit our manuscript entitled "IIIVmrMLM.QEI: An effective tool for indirect detection of QTN-by-environment interactions in genome-wide association studies" to be considered for publication in Computational and Structural Biotechnology Journal. No conflict of interest exits in the submission of this manuscript, all authors acknowledge the content of this manuscript and consent to its publication.

Declaration of Competing Interest

All authors acknowledge the content of this manuscript and consent to its publication.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.11.046.

Data Availability

The new tool can be downloaded from https://github.com/YuanmingZhang65/IIIVmrMLM.QEI and has been presented in this manuscript as IIIVmrMLM.QEI_1.0.zip (R code) and IIIVmrMLM.QEI.zip (CLI). All the simulated and real datasets in this study are included in the IIIVmrMLM.QEI software and are also presented as Data Files S1 to S5 (real datasets) and S6 to S9 (simulated datasets).

References

[1] Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 2012;44(9):1066–71.
[2] Casale FP, Horta D, Rakitsch B, Stegle O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. PLoS Genet 2017;13(4):e1006693.
[3] Zan Y, Carlborg Ö. A polygenic genetic architecture of flowering time in the worldwide Arabidopsis thaliana population. Mol Biol Evol 2019;36(1):141–54.
[4] Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, et al. Sugar-sweetened beverages and genetic risk of obesity. N Engl J Med 2012;367:1387–96.
[5] Huang T, Hu FB. Gene-environment interactions and obesity: recent developments and future directions. BMC Med Genom 2015;8(1):S2.
[6] Kerin M, Marchini J. Inferring gene-by-environment interactions with a Bayesian whole-genome regression model. Am J Hum Genet 2020;107(4):698–713.
[7] Li X, Guo T, Wang J, Bekele WA, Sukumaran S, et al. An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. Mol Plant 2021;14(6):874–87.
[8] Li M, Zhang YW, Zhang ZC, Xiang Y, Liu MH, et al. A compressed variance component mixed model for detecting QTNs and QTN-by-environment and QTN-by-QTN interactions in genome-wide association studies. Mol Plant 2022;15(4):630–50.
[9] Sul JH, Bilow M, Yang WY, Kostem E, Furlotte N, et al. Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. PLoS Genet 2016;12(3):e1005849.
[10] Moore, Casale R, Jan Bonder FP, Horta M, BIOS Consortium D, et al. A linear mixed-model approach to study multivariate gene-environment interactions. Nat Genet 2019;51(1):180–6.
[11] Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several. Technometrics 1974;16:129–32.
[12] Rönnegård L, Valdar W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. BMC Genet 2012;13:63.
[13] Lee Y, Nelder JA. Hierarchical generalized linear models. J R Stat Soc B 1996;58:619–56.
[14] Murphy MD, Fernandes SB, Morota G, Lipka AE. Assessment of two statistical approaches for variance genome-wide association studies in plants. Heredity 2022;129(2):93–102.
[15] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42(4):348–54.
[16] Wang J, Zhang Z. GAPIT Version 3: boosting power and accuracy for genomic association and prediction. Genom Prote Bioinf 2021;19(4):629–40.
[17] Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 2010;465(7298):627–31.
[18] Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, et al. The extent of linkage disequilibrium in Arabidopsis thaliana. Nat Genet 2002;30(2):190–3.
[19] Wang SB, Feng JY, Ren WL, Huang B, Zhou L, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 2016;6:19444.
[20] Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform 2018;19(4):700–12.
[21] Zhu B, Zhu M, Jiang J, Niu H, Wang Y, et al. The impact of variable degrees of freedom and scale parameters in Bayesian methods for genomic prediction in Chinese Simmental Beef Cattle. PLoS One 2016;11(5):e0154118.
[22] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559–75.
[23] Xu S. An expectation-maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 2010;105(5):483–94.
[24] Finlay KW, Wilkinson GN. The analysis of adaptation in a plant-breeding programme. Aust J Agric Res 1963;14(6):742–54.
[25] Zhang YM, Jia Z, Dunwell JM. Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. Front Plant Sci 2019;10:100.
[26] Zhang YM, Jia Z, Dunwell JM. Editorial: The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. *Lausanne*: Frontiers Media,; 2019.
[27] Zhang YW, Wen YJ, Dunwell JM, Zhang YM. QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked QTLs for quantitative traits in bi-parental segregation populations. Comput Struct Biotechnol J 2019;18:59–65.
[28] Zhang YM, Jia Z, Xie SQ, Wen J, Wang S, et al. Editorial: advances in statistical methods for the genetic dissection of complex traits in plants. Front Plant Sci 2024;15:1357564.
[29] Li M, Zhang YW, Xiang Y, Liu MH. IIIVmrMLM: The R and C++ tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. Mol Plant 2022;15(8):1251–3.
[30] Zhou YH, Li G, Zhang YM. A compressed variance component mixed model framework for detecting small and linked QTL-by-environment interactions. Brief Bioinform 2022;23(2):bbab596.
[31] He L, Xiao J, Rashid KY, Jia G, Li P, et al. Evaluation of genomic prediction for pasmo resistance in flax. Int J Mol Sci 2019;20(2):359.
[32] Yin L, Zhang H, Zhou X, Yuan X, Zhao S, et al. KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome Biol 2020;21(1):146.
[33] Yu GN, Cui YR, Jiao YX, Zhou K, Wang X, et al. Comparison of sequencing-based and array-based genotyping platforms for genomic prediction of maize hybrid performance. Crop J 2023;11(2):490–8.
[34] Fu R, Wang X. Modeling the influence of phenotypic plasticity on maize hybrid performance. Plant Commun 2023;4(3):100548.
[35] Bernardo R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. Heredity 2020;125(6):375–85.