

pKWmEB: integration of Kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study

Article

Accepted Version

Ren, W.-L., Wen, Y.-J., Dunwell, J. and Zhang, Y.-M. (2018) pKWmEB: integration of Kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity*, 120. pp. 208-218. ISSN 1365-2540 doi: <https://doi.org/10.1038/s41437-017-0007-4> Available at <https://centaur.reading.ac.uk/72952/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

To link to this article DOI: <http://dx.doi.org/10.1038/s41437-017-0007-4>

Publisher: Nature Publishing Group

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

1 **pKWmEB: Integration of Kruskal-Wallis test with**
2 **empirical Bayes under polygenic background control**
3 **for multi-locus genome-wide association study**

4

5 Wen-Long Ren^{1,§}, Yang-Jun Wen^{1,§}, Jim M. Dunwell², Yuan-Ming Zhang^{1,*}

6

7 1 State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural
8 University, Nanjing 210095, China / Statistical Genomics Lab, College of Plant Science and
9 Technology, Huazhong Agricultural University, Wuhan 430070, China

10 2 School of Agriculture, Policy and Development, University of Reading, Reading RG6 6AR,
11 United Kingdom

12

13 §: These authors contributed equally to this work.

14

15 * **Correspondence:** Dr. Yuan-Ming Zhang, College of Agriculture, Nanjing Agricultural
16 University, Nanjing 210095, China. E-mail: soyzhang@njau.edu.cn / College of Plant Science and
17 Technology, Huazhong Agricultural University, Wuhan 430070, China. E-mail:
18 soyzhang@mail.hzau.edu.cn

19

20 **Abstract**

21 Although non-parametric methods in genome-wide association studies (GWAS) are robust in
22 quantitative trait nucleotide (QTN) detection, the absence of polygenic background control in
23 single-marker association in genome-wide scans results in a high false positive rate. To overcome
24 this issue, we proposed an integrated non-parametric method for multi-locus GWAS. First, a new
25 model transformation was used to whiten the covariance matrix of polygenic matrix K and
26 environmental noise. Using the transferred model, Kruskal-Wallis test along with least angle
27 regression was then used to select all the markers that were potentially associated with the trait.
28 Finally, all the selected markers were placed into multi-locus model, these effects were estimated
29 by empirical Bayes, and all the nonzero effects were further identified by a likelihood ratio test for
30 true QTN detection. This method, named pKWmEB, was validated by a series of Monte Carlo
31 simulation studies. As a result, pKWmEB effectively controlled false positive rate, although a less
32 stringent significance criterion was adopted. More importantly, pKWmEB retained the high power
33 of Kruskal-Wallis test, and provided QTN effect estimates. To further validate pKWmEB, we
34 re-analyzed four flowering time related traits in *Arabidopsis thaliana*, and detected some
35 previously reported genes that weren't identified by the other methods.

36 **Keywords:** genome-wide association study, Kruskal-Wallis test, multi-locus model, empirical
37 Bayes, polygenic background control

38

39 Introduction

40 The genome-wide association study (GWAS) has become a very effective approach to identifying
41 the genetic loci associated with complex traits (Sladek *et al.*, 2007; WTCCC, 2007; Li *et al.*, 2013).
42 Since the establishment of mixed linear model (MLM) based GWAS methods (Zhang *et al.*, 2005;
43 Yu *et al.*, 2006), then there has been an increasing interest in using MLM in GWAS, because of
44 their demonstrated effectiveness in accounting for relatedness between individuals and in
45 controlling population stratification. This has stimulated the development of the MLM-based
46 GWAS methods (Kang *et al.*, 2008; Zhang *et al.*, 2010; Lippert *et al.*, 2011; Zhou and Stephens,
47 2012; Segura *et al.*, 2012; Wang *et al.*, 2016). Furthermore, these methods have been widely used
48 in GWAS; the loci identified in GWAS explain only a fraction of heritability of complex trait,
49 indicating that additional loci influencing those traits exist.

50
51 To increase the robustness of quantitative trait nucleotide (QTN) detection in GWAS,
52 non-parametric approaches have been recommended. Up to now several existing non-parametric
53 methods have been used to conduct GWAS. For example, Atwell *et al.* (2010) adopted Wilcoxon
54 rank-sum test (Wilcoxon, 1945; Mann and Whitney, 1947) to carry out GWAS for 107 phenotypes
55 in a common set of *Arabidopsis thaliana* inbred lines; the 107 phenotypes were re-analyzed by
56 Kruskal-Wallis test (Kruskal and Wallis, 1952) and more significantly associated SNPs were
57 identified as compared with those using efficient mixed model association (EMMA) (Filiault and
58 Maloof, 2012); the Kruskal-Wallis test was also generalized to group uncertainty when comparing
59 k samples, and one application to a GWAS of type 1 diabetic complications demonstrated the
60 utility of the generalized Kruskal-Wallis test for study with group uncertainty (Acar and Sun,
61 2013). Similarly, Beló *et al.* (2008) used Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov,
62 1948) to detect an allelic variant of *fad2* associated with increased oleic acid levels in maize, and
63 Terao *et al.* (2014) and Tan *et al.* (2014) adopted Jonckheere-Terpstra test (Terpstra, 1952;
64 Jonckheere, 1954) to detect a T allele of rs2395185 in human leukocyte antigen (HLA) locus and a
65 T allele of rs1260326 and rs780094 in glucokinase regulatory (GCKR) loci, respectively. None of
66 the above approaches have included population structure in their genetic model. Thus, Yang *et al.*
67 (2014) integrated Anderson-Darling test with a population structure correction. This method was

68 used to analyze 17 agronomic traits in maize, and some important loci were identified. In practice,
69 the true model for a quantitative trait is rarely known, and model misspecification can lead to a
70 loss of power. To address this issue, Kozlitina and Schucany (2015) proposed a rank-based
71 maximum test (MAX3), which has favorable properties relative to other tests, especially in the
72 case of symmetric distributions with heavy tails. We found that all the above methods have high
73 false positive rates in our simulation experiments. To overcome this problem, multi-locus model
74 methodologies should be recommended. For example, Li *et al.* (2014) proposed a two-stage
75 non-parametric approach, in which all the markers potentially associated with quantitative trait are
76 identified and their effects in one multi-locus model are estimated by shrinkage estimation for true
77 QTN detection. However, none of the above methods have controlled polygenic background in
78 single-marker association in genome scans.

79
80 In this study, we proposed a two-stage method for multi-locus GWAS. First, the model
81 transformation of Wen *et al.* (2017) was used to control polygenic background in single-marker
82 association in genome scans. Using the transformed model, Kruskal-Wallis test along with least
83 angle regression of Efron *et al.* (2004) was then used to select all the markers that were potentially
84 associated with the trait. Finally, all the selected markers were placed into multi-locus model,
85 these effects were estimated by empirical Bayes, and all the nonzero effects were further identified
86 by a likelihood ratio test. Clearly, this method integrates the Kruskal-Wallis test with empirical
87 Bayes under polygenic background control. This method, named pKWmEB, was validated by a
88 series of Monte Carlo simulation studies and real data analyses for four flowering time related
89 traits in *Arabidopsis*.

90 **Materials and Methods**

91 ***The Arabidopsis thaliana* dataset**

92 The *Arabidopsis thaliana* dataset was downloaded from <http://www.arabidopsis.usc.edu/> (Atwell
93 *et al.*, 2010) and used to conduct simulation experiments and real data analysis. This dataset
94 contained 199 accessions each with 216130 genotyped SNPs.

95 Genetic model and model transformation

96 The standard mixed linear model (MLM) for an $n \times 1$ phenotypic vector \mathbf{y} of quantitative trait is

$$97 \quad \mathbf{y} = \mathbf{1}\mu + \mathbf{Q}\mathbf{v} + \mathbf{G}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (1)$$

98 where n is the number of individuals; $\mathbf{1}$ is a $n \times 1$ vector of 1; μ is overall average; \mathbf{Q} is an
99 $n \times c$ matrix of fixed effects, including population structure (Yu *et al.*, 2006) or principle
100 component (Price *et al.*, 2010), and \mathbf{v} is a $c \times 1$ vector of fixed effects excluding the intercept μ ;
101 \mathbf{G} is an $n \times 1$ vector of putative QTN genotypes, and $\boldsymbol{\beta}$ is fixed effect of putative QTN;
102 $\mathbf{u} \sim \text{MVN}_m(\mathbf{0}, \sigma_g^2 \mathbf{K})$ is an $m \times 1$ vector of polygenic effects, \mathbf{K} is an $m \times m$ kinship matrix, σ_g^2 is
103 polygenic variance, and MVN denotes multivariate normal distribution; $\mathbf{Z} = (z_{ij})_{n \times m}$ is the
104 corresponding designed matrix for \mathbf{u} , $z_{ij} = 1$ if individual i comes from family j ($j = 1, \dots, m$) and
105 $z_{ij} = 0$ otherwise; and $\boldsymbol{\varepsilon} \sim \text{MVN}_n(\mathbf{0}, \sigma_e^2 \mathbf{I}_n)$ is an $n \times 1$ vector of residual errors, σ_e^2 is residual error
106 variance, \mathbf{I}_n is an $n \times n$ identity matrix. To simplify population structure, let $m = n$ and $\mathbf{Z} = \mathbf{I}_n$
107 in this study (Atwell *et al.*, 2010). Note that the observed data is (\mathbf{y}, \mathbf{G}) , matrices \mathbf{Q} and \mathbf{K} can be
108 calculated from \mathbf{G} , and the parameters to be estimated are $\mu, \mathbf{v}, \boldsymbol{\beta}, \sigma_g^2$ and σ_e^2 .

109
110 Based on model (1), phenotypic values \mathbf{y} were affected by population structure, QTN and
111 polygenes. In other words, a nonparametric test for k samples cannot be directly applied. Thus, we
112 must remove the effects for population structure and polygenes before using a nonparametric test.

113 114 Population structure correction

115 If we delete $\mathbf{G}\boldsymbol{\beta}$ and $\mathbf{Z}\mathbf{u}$ in model (1), its reduced model is

$$116 \quad \mathbf{y} = \mathbf{1}\mu + \mathbf{Q}\mathbf{v} + \boldsymbol{\varepsilon} \quad (2)$$

117 Using least squares method, the effect of \mathbf{v} , denoted by $\hat{\mathbf{v}}$, can be estimated from \mathbf{y} , \mathbf{Q} and $\mathbf{1}$.

118 Thus, we can correct the effect of population structure from

$$119 \quad \mathbf{y}_{\cdot Q} = \mathbf{y} - \mathbf{Q}\hat{\mathbf{v}} = \mathbf{1}\mu + \mathbf{G}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (3)$$

120 121 Polygenic background correction

122 Based on model (3), the variance of $\mathbf{y}_{\cdot Q}$ is

123
$$\begin{aligned}\text{Var}(\mathbf{y}_{\cdot Q}) &= \sigma_g^2 \mathbf{ZKZ}^T + \sigma_e^2 \mathbf{I}_n \\ &= \sigma_e^2 (\lambda_g \mathbf{ZKZ}^T + \mathbf{I}_n)\end{aligned}\quad (4)$$

124 where $\lambda_g = \sigma_g^2 / \sigma_e^2$. Using the EMMA algorithm of Kang *et al.* (2008), the estimate of λ_g , denoted
125 by $\hat{\lambda}_g$, can be easily obtained. Replacing λ_g in (4) by $\hat{\lambda}_g$, so

126
$$\text{Var}(\mathbf{y}_{\cdot Q}) = \sigma_e^2 (\hat{\lambda}_g \mathbf{ZKZ}^T + \mathbf{I}_n) = \sigma_e^2 \mathbf{B} \quad (5)$$

127 where $\mathbf{B} = \hat{\lambda}_g \mathbf{ZKZ}^T + \mathbf{I}_n$. An eigen decomposition of positive semi-definite matrix \mathbf{B} is

128
$$\begin{aligned}\mathbf{B} &= \mathbf{Q}_B \mathbf{\Lambda}_B \mathbf{Q}_B^T \\ &= (\mathbf{Q}_1 \quad \mathbf{Q}_2) \begin{pmatrix} \mathbf{\Lambda}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix} \\ &= (\mathbf{Q}_1 \quad \mathbf{Q}_2) \begin{pmatrix} \mathbf{\Lambda}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix} \\ &= (\mathbf{Q}_1 \quad \mathbf{Q}_2) \begin{pmatrix} \mathbf{\Lambda}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix} (\mathbf{Q}_1 \quad \mathbf{Q}_2) \begin{pmatrix} \mathbf{\Lambda}_r^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_1^T \\ \mathbf{Q}_2^T \end{pmatrix} \\ &= (\mathbf{Q}_1 \mathbf{\Lambda}_r^{\frac{1}{2}} \mathbf{Q}_1^T) (\mathbf{Q}_1 \mathbf{\Lambda}_r^{\frac{1}{2}} \mathbf{Q}_1^T)\end{aligned}\quad (6)$$

129 where \mathbf{Q}_B is orthogonal, $\mathbf{\Lambda}_r$ is a diagonal matrix with positive eigen values, $r = \text{Rank}(\mathbf{B})$, \mathbf{Q}_1
130 and \mathbf{Q}_2 are the $n \times r$ and $n \times (n-r)$ block matrices of \mathbf{Q}_B , and $\mathbf{0}$ is the corresponding block
131 zero matrix (Wen *et al.*, 2017).

132 Let $\mathbf{C} = \mathbf{Q}_1 \mathbf{\Lambda}_r^{-\frac{1}{2}} \mathbf{Q}_1^T$, a new model with polygenic background control is

134
$$\mathbf{y}_c = \mathbf{1}_c \mu + \mathbf{G}_c \beta + \boldsymbol{\varepsilon}_c \quad (7)$$

135 where $\mathbf{y}_c = \mathbf{C} \mathbf{y}_{\cdot Q}$, $\mathbf{1}_c = \mathbf{C} \mathbf{1}$, $\mathbf{G}_c = \mathbf{C} \mathbf{G}$ and $\boldsymbol{\varepsilon}_c = \mathbf{C} (\mathbf{Z} \boldsymbol{\mu} + \boldsymbol{\varepsilon})$. Clearly, the observed data is $(\mathbf{y}_c, \mathbf{G}_c)$,
136 and the parameter to be estimated is β . Using $\lambda_g = \hat{\lambda}_g$, equation (6) and $\mathbf{Q}_1^T \mathbf{Q}_1 = \mathbf{I}_r$, so

137
$$\begin{aligned}\text{Var}(\boldsymbol{\varepsilon}_c) &= \sigma_e^2 \mathbf{C} (\hat{\lambda}_g \mathbf{ZKZ}^T + \mathbf{I}_n) \mathbf{C}^T \\ &= \sigma_e^2 \mathbf{C} \mathbf{B} \mathbf{C}^T \\ &= \sigma_e^2 \left[\mathbf{Q}_1 \mathbf{\Lambda}_r^{-\frac{1}{2}} \mathbf{Q}_1^T (\mathbf{Q}_1 \mathbf{\Lambda}_r^{\frac{1}{2}} \mathbf{Q}_1^T) (\mathbf{Q}_1 \mathbf{\Lambda}_r^{\frac{1}{2}} \mathbf{Q}_1^T) (\mathbf{Q}_1 \mathbf{\Lambda}_r^{-\frac{1}{2}} \mathbf{Q}_1^T)^T \right] \\ &= \sigma_e^2 \mathbf{I}_n\end{aligned}$$

138 It should be noted that model (7) includes QTN variation and normal residual error (Wen *et al.*,
139 2017). Although the polygenic background has been corrected, non-parametric test cannot be
140 implemented owing to continual \mathbf{G}_c values.

141 **Kruskal-Wallis test**

142 Based on model (7), we used Kruskal-Wallis test to detect whether one SNP was associated with
 143 the trait. However, the values of \mathbf{G}_c were not binary variable. Thus, we must transfer \mathbf{G}_c into
 144 binary variable. Let $\mathbf{G}_c = (g_{ij})_{n \times p}$, $\mathbf{G}_c^* = (g_{ij}^*)_{n \times p}$, p is the number of QTNs under study and

145
$$\bar{g}_{.j} = \frac{1}{n} \sum_{i=1}^n g_{ij}, \text{ so}$$

146
$$g_{ij}^* = \begin{cases} 1, & g_{ij} \geq \bar{g}_{.j} \\ -1, & g_{ij} < \bar{g}_{.j} \end{cases} \quad (8)$$

147 Therefore, $(\mathbf{y}_c, \mathbf{G}_c^*)$ is the dataset for Kruskal-Wallis test. All the transferred phenotypes \mathbf{y}_c
 148 were grouped by the values of \mathbf{G}_c^* . In this situation, there are two groups for the transferred
 149 phenotypes \mathbf{y}_c . In the two groups, let their sizes be n_i , and their cumulative distribution
 150 functions be $F_i(y|\theta_i)$ ($i=1, 2$). The null hypothesis for Kruskal-Wallis test was

151
$$H_0 : \theta_1 = \theta_2; H_1 : \theta_1 \neq \theta_2 \quad (9)$$

152 When precise category assignment of \mathbf{G}_c^* is available, Kruskal-Wallis test for (9) is conducted by
 153 ranking all the transferred phenotypes \mathbf{y}_c together and comparing the rank sum for each group. If
 154 $H_0 : \theta_1 = \theta_2$, so the estimate for β in equation (7) equals to zero. The statistic H

155
$$H = \frac{12}{n(n+1)} \sum_{i=1}^2 \frac{R_i^2}{n_i} - 3(n+1) \quad (10)$$

156 follows an asymptotic χ^2 distribution with one degree of freedom (Kruskal, 1952), where r_j is
 157 the rank of the j th phenotype of \mathbf{y}_c in the overall sample; and $R_i = \sum_{j=1}^n I_{ij} r_j$ ($i=1, 2$), I_{ij} is an
 158 indicator variable, $I_{ij} = 1$ if the j th phenotype of \mathbf{y}_c belongs to the i th group and $I_{ij} = 0$
 159 otherwise; and $n_i = \sum_{j=1}^n I_{ij}$.

160 **Empirical Bayes estimation for QTN effects**

161 In GWAS, the number of SNPs is frequently 1000 times larger than sample size. In this situation,
 162 fitting all the genome markers in one model is not feasible. As we know, most SNPs are not
 163 associated with the trait. Once we delete these SNPs with zero effects, the reduced model is
 164 estimable. The purpose of the above Kruskal-Wallis test is to select all the potentially associated

165 SNPs. If the number of markers passing the 0.05 level of significance test is more than o_i
166 ($o_i = 50, 100$ and 150), we invoke least angle regression (LARS) of Efron *et al.* (2004) to select
167 o_i variables that are most likely associated with the trait of interest. LARS is a flexible method
168 for variable selection, which is implemented by lars package in R language
169 (<http://cran.r-project.org/web/packages/lars/>). The o_i markers are then included in a multi-locus
170 model. If the number of markers passing the initial test is less than o_i , we skip the LARS step and
171 proceed to include all the selected markers in a multi-locus model

$$172 \quad \mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^q \mathbf{G}_i \beta_i + \boldsymbol{\varepsilon} \quad (11)$$

173 where \mathbf{y} , $\mathbf{1}$, μ and $\boldsymbol{\varepsilon}$ are the same as those in model (1); q is the number of markers
174 selected in Krusal-Wallis test; β_i is the effect for marker i , and \mathbf{G}_i is the corresponding
175 designed matrix for β_i . Clearly, the observed data is $(\mathbf{y}, \mathbf{G}_1, \dots, \mathbf{G}_q)$, the parameters to be
176 estimated are β_1, \dots, β_q . In model (11), the polygenic background is not considered. In theory, this
177 is because all the potentially associated loci have been included in this model. However, we
178 should determine whether population structure is considered. To solve this issue, the linkage
179 disequilibrium score regression test of Bulik-Sullivan *et al.* (2015) is used (see Discussion). In the
180 selection of markers, a less stringent criterion is adopted.

181
182 Empirical Bayes of Xu (2010) was used to estimate the SNP effects in model (11). In this method,
183 each SNP effect β_i is viewed as random. We adopt normal prior for β_i , $P(\beta_i | \sigma_i^2) = N(0, \sigma_i^2)$, and
184 the scaled inverse χ^2 prior for σ_i^2 , $P(\sigma_i^2 | \tau, \omega) \propto (\sigma_i^2)^{-(\tau+2)} \exp\left(-\frac{\omega}{2\sigma_i^2}\right)$, where $(\tau, \omega) = (0, 0)$,
185 which represents the Jeffreys' prior (Figueiredo, 2003), $P(\sigma_i^2 | \tau, \omega) = 1/\sigma_i^2$. The procedure for
186 parameter estimation in empirical Bayes is as follows.

187 1) Initial-step: To initialize parameters with

$$188 \quad \begin{aligned} \mu &= \mathbf{1}^T \mathbf{y} / n \\ \sigma_e^2 &= \frac{1}{n} (\mathbf{y} - \mathbf{1}\mu)^T (\mathbf{y} - \mathbf{1}\mu) \\ \sigma_i^2 &= \left[(\mathbf{G}_i^T \mathbf{G}_i)^{-1} \mathbf{G}_i^T (\mathbf{y} - \mathbf{1}\mu) \right]^2 + (\mathbf{G}_i^T \mathbf{G}_i)^{-1} \sigma_e^2 \end{aligned}$$

189 2) E-step: marker effect can be predicted by

$$190 \quad E(\beta_i) = \sigma_i^2 \mathbf{G}_i^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{1}\mu) \quad (12)$$

191 where $\mathbf{V} = \sum_{i=1}^q \mathbf{G}_i \mathbf{G}_i^T \sigma_i^2 + \mathbf{I}\sigma_e^2$.

192 3) M-step: To update parameters σ_i^2 , μ and σ_e^2

$$193 \quad \begin{aligned} \sigma_i^2 &= \frac{E(\beta_i^T \beta_i) + \omega}{\tau + 3} \\ \mu &= (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{y} \\ \sigma_e^2 &= \frac{1}{n} (\mathbf{y} - \mathbf{1}\mu)^T \left(\mathbf{y} - \mathbf{1}\mu - \sum_{i=1}^q \mathbf{G}_i E(\beta_i) \right) \end{aligned} \quad (13)$$

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

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

196

197 Owing to $o_i = 50, 100$ and 150 , so three models would be established by the above procedures.

198 Their AIC values were calculated in order to pick up an optimal model.

199 Likelihood ratio test

200 Based on the estimate of marker effect β_i in the optimal model, all the markers with $|\hat{\beta}_i| \leq 10^{-4}$

201 are deemed not to be associated with the trait. The other markers with the effects $\theta = \{\beta_{(1)}, \dots, \beta_{(o)}\}$

202 are potentially associated with the trait. To test the null hypothesis $H_0: \beta_{(i)} = 0$, which is no QTN

203 linked to the i th marker, LR test was conducted by

$$204 \quad \text{LR}_i = -2[\text{L}(\theta_{-i}) - \text{L}(\theta)] \quad (14)$$

205 where $\theta_{-i} = \{\beta_{(1)}, \dots, \beta_{(i-1)}, \beta_{(i+1)}, \dots, \beta_{(o)}\}^T$, $\text{L}(\theta) = \sum_{i=1}^n \ln \phi(y_i; \mathbf{1}\mu + \sum_{o=1}^O \mathbf{G}_o \beta_o, \sigma_e^2)$ is log-likelihood function,

206 $\phi(y_i; \mathbf{1}\mu + \sum_{o=1}^O \mathbf{G}_o \beta_o, \sigma_e^2)$ is a normal density with mean $\mathbf{1}\mu + \sum_{o=1}^O \mathbf{G}_o \beta_o$ and variance σ_e^2 , and

207 $\text{LOD} = \text{LR}/4.605$. Although the general 0.05 critical value may be used for significance test, we

208 decided to set up a slightly more stringent criterion of $\text{LOD}=3.0$. The criterion is frequently

209 adopted in linkage analysis and is the equivalent of $P = \Pr(\chi_1^2 > 3.0 \times 4.605) \approx 0.0002$, in which χ_1^2
210 under H_0 , follows a χ^2 distribution with one degree of freedom.

211
212 The flow diagram of pKWmEB is shown in **Fig 1**. pKWmEB has been implemented in R and its
213 software can be downloaded from <https://cran.rproject.org/web/packages/mrMLM/index.html>.

214 **Genome-wide efficient mixed model association (GEMMA)**

215 This is an existing GWAS method (Zhou and Stephens, 2012) and used as a gold standard for
216 comparison. This method is the fixed model version of the original MLM, in which β_i was
217 treated as fixed effect with no distribution assigned. The method was implemented in the C
218 software GEMMA (Zhou and Stephens, 2012) (<http://www.xzlab.org/software.html>). The
219 threshold of P-value was set as $0.05/p$ after Bonferroni correction for multiple tests, where p is the
220 number of markers.

221 **Monte Carlo simulation experiments**

222 Five Monte Carlo simulation experiments were used to validate pKWmEB. In the first experiment,
223 all the SNP genotypes were derived from 216,130 SNPs in Atwell *et al.* (2010) and 2000 SNPs
224 were randomly sampled from each chromosome. The positions for the sampled SNPs were
225 described by Wang *et al.* (2016). The sample size was the number of accessions (199) in Atwell *et al.*
226 *et al.* (2016). Six quantitative trait nucleotides (QTNs) were simulated and placed on the SNPs with
227 allelic frequencies of 0.30; their heritabilities were set as 0.10, 0.05, 0.05, 0.15, 0.05 and 0.05,
228 respectively; and their positions and effects were listed on Table S1. Using
229 $h_r^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_e^2) = 0.05 \times 4 + 0.10 + 0.15 = 0.45$ and residual variance $\sigma_e^2 = 10.0$, total genetic
230 variance for six simulated QTNs (σ_G^2) and individual genetic variance for each simulated QTN
231 ($\sigma_r^2, r=1, \dots, 6$) could be obtained. σ_r^2 was a function of QTN effect and frequency of common
232 allele. Thus, QTN effect could be obtained. The average was set at 10.0. The new phenotypes
233 were simulated by the model: $y = \mu + \sum_{i=1}^6 x_i b_i + \varepsilon$, where $\varepsilon \sim \text{MVN}_n(0, 10 \times I_n)$. The simulation
234 was replicated 1000 times. In the Kruskal-Wallis test, the o_i most associated SNPs were selected

235 and placed into multi-locus model. A detected QTN within 1 kb of the simulated QTN was
 236 considered to be a true QTN. For each simulated QTN, we counted the samples in which the LOD
 237 statistic exceeded 3.0. The ratio of the number of such samples to the total number of replicates
 238 (1000) represented the empirical power of this QTN. False positive rate (FPR) was calculated as
 239 the ratio of the number of false positive effects to the total number of zero effects considered in
 240 the full model. To measure the variance and bias of gene effect estimate, mean squared error
 241 (MSE)

$$242 \quad \text{MSE}_k = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\beta}_{k(i)} - \beta_k)^2 \quad (15)$$

243 was calculated, where $\hat{\beta}_{k(i)}$ is the estimate of β_k in the i th sample.

244
 245 To investigate the effect of polygenic background on pKWmEB, polygenic effects were simulated
 246 in the second experiment by multivariate normal distribution $\text{MVN}_n(0, \sigma_{pg}^2 \mathbf{K})$, where σ_{pg}^2 is
 247 polygenic variance and \mathbf{K} is kinship matrix between a pair of individuals. Here $\sigma_{pg}^2 = 2$, so
 248 $h_{pg}^2 = 0.092$. The QTN size (h^2), average, residual variance, and other parameter values were the
 249 same as those in the first experiment, and all the parameters were listed on Table S2. The new

250 phenotypes were simulated by the model: $y = \mu + \sum_{i=1}^6 x_i b_i + u + \varepsilon$, where $u \sim \text{MVN}_n(0, 2 \times \mathbf{K})$

251 and $\varepsilon \sim \text{MVN}_n(0, 10 \times \mathbf{I}_n)$.

252
 253 To investigate the effect of epistatic background on pKWmEB, three epistatic QTNs were
 254 simulated in the third simulation experiment. The related parameters for the three epistatic QTNs
 255 were described in Wang *et al.* (2016). The QTN sizes (h^2), average, residual variance, and other
 256 parameter values were also the same as those in the first experiment, and all the parameters were

257 listed on Table S3. The new phenotypes were simulated by $y = \mu + \sum_{i=1}^6 x_i b_i + \sum_{j=1}^3 (A_j \# B_j) b_{jj} + \varepsilon$,

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

259
 260 All simulated data sets are available from <http://dx.doi.org/10.5061/dryad.sk652> (the Dryad
 261 Digital Repository).

262
263 To investigate the effect of skewed phenotypic distribution on pKWmEB, normal distribution for
264 residual error in the first simulation experiment was replaced by log-normal distribution in the
265 fourth simulation experiment and logistic distribution in the fifth simulation experiment, and other
266 parameter values were the same as those in the first simulation experiment. To let residual error
267 variance be 10, the standard deviation was set at 1.144 in log-normal distribution and 1.743 in
268 logistic distribution. The means for the two skewed distributions were also zero. The two
269 simulation datasets were included in Dataset S2.

270 **Results**

271 **Monte Carlo simulation studies**

272 *Statistical power for QTN detection* To validate pKWmEB, five simulation experiments were
273 conducted. In the first simulation experiment, each sample was analyzed by five methods:
274 pKWmEB, the new method without polygenic background control (KWmEB), Kruskal-Wallis test
275 with Bonferroni correction (KWsBC), genome-wide efficient mixed model association (GEMMA),
276 and multi-locus random-SNP-effect mixed linear model (mrMLM). All the power results are
277 shown in Table S1 and Fig 2a. Clearly, the average powers for the above five methods were 69.8,
278 67.3, 60.7, 46.0 and 68.6 (%), respectively, indicating the highest average power of pKWmEB
279 (Fig 2a). More importantly, the power using pKWmEB was significantly higher than those using
280 KWmEB and GEMMA (Table 1). Note that there were four QTNs with the same 5% heritability.
281 The standard deviation of powers across the four QTNs might be used to measure the robustness
282 of each method. As a result, the standard deviation was 13.01 for pKWmEB, 11.98 for KWmEB
283 and 10.57 for mrMLM, which were much less than 35.17 for KWsBC, indicating the better
284 stability of pKWmEB. On one occasion, the power for the fifth QTN using pKWmEB was 47.7%
285 less than that using KWsBC. To further confirm the effectiveness of pKWmEB, polygenic effect
286 simulated by multivariate normal distribution ($r^2=9.2\%$) was added to each phenotypic observation
287 in the second simulation experiment and the polygenic background was replaced by three epistatic
288 QTN ($r^2=15\%$) in the third simulation experiment. These results are listed in Tables S2 and S3,
289 which show that the average powers for the above five methods were 69.1, 67.7, 58.9, 42.5 and

290 67.6 (%) in the second simulation experiment (Table S2, Fig 2b), and 61.9, 59.9, 54.9, 39.1 and
291 58.9 (%), respectively, in the third simulation experiment (Table S3, Fig 2c). The standard
292 deviation of statistical powers among all the 5% QTNs was 21.31 for pKWmEB and 31.39 for
293 KWsBC in the second simulation experiment, and 15.05 for pKWmEB and 40.77 for KWsBC in
294 the third simulation experiment. Similarly, the power for the fifth QTN using pKWmEB was 47.2
295 and 68.3 (%) less than those using KWsBC in the second and third simulation experiments,
296 respectively. In addition, residual error distributions in the above three experiments were replaced
297 by log-normal (the fourth simulation experiment) and logistic (the fifth simulation experiment)
298 distributions. The average powers for the above five methods were 76.2, 74.4, 80.1, 53.9 and 78.3
299 (%) in the fourth simulation experiment (Table S4, Fig 2d), and 68.7, 66.9, 60.9, 44.1 and 68.0
300 (%), respectively, in the fifth simulation experiment (Table S5, Fig 2e). Similar phenomena were
301 observed for the fifth QTN and the standard deviation of statistical powers across all the 5% QTNs
302 in the last two experiments. In summary, pKWmEB with polygenic background control is better
303 than KWmEB without polygenic background control; pKWmEB retains the high power of
304 KWsBC, and it is better in the stability of statistical power than KWsBC.

305
306 *Accuracies of estimated QTN effects* The accuracy of QTN effect estimation was measured
307 by mean squared error (MSE) and smaller MSE indicates higher accuracy of parameter estimation.
308 All the MSE results from four approaches in the five simulation experiments are shown in Fig 3
309 and Tables S6 to S10, because KWsBC doesn't provide the estimates for QTN effects. Results
310 showed that the average MSEs using pKWmEB, KWmEB, GEMMA and mrMLM were 0.0797,
311 0.0825, 0.5467 and 0.0940 in the first simulation experiment, respectively, indicating the
312 minimum average MSE of pKWmEB (Fig 3a and Table S6). More importantly, the MSE using
313 pKWmEB was almost significantly less than that using GEMMA (Table 1). Almost similar trends
314 were found in the other simulation experiments (Tables S16 to S19, Fig 3a to 3e). Average value
315 of each QTN effect across 1000 replicates was listed in Tables S11 to S15. These results were also
316 confirmed the above trends.

317
318 *False positive rate (FPR)* The FPR is similar to the empirical Type 1 error rate. The FPRs in
319 all the five simulation experiments were 0.0356 ± 0.0085 (%) for pKWmEB, 0.0385 ± 0.0073 (%)

320 for KWmEB, 0.6130 ± 0.1644 (%) for KWsBC, 0.0290 ± 0.0094 (%) for GEMMA and $0.0214 \pm$
321 0.0043 (%) for mrMLM (Fig 4 and Tables S1 to S5). In summary, the FPRs are less than 0.05 %
322 for pKWmEB, KWmEB, mrMLM and GEMMA, and more than 0.6 % for KWsBC, indicating the
323 best FPR control of pKWmEB even if a less stringent significant criterion was adopted.

324
325 **Computational efficiency** Each sample in the first simulation experiment was analyzed by
326 pKWmEB, KWmEB, KWsBC, mrMLM and GEMMA. These analyses were implemented on the
327 computer (Intel(R) Xeon(R) CPU E5-2637 v2 @ 3.50GHz CPU). As a result, the computing times
328 using the above five methods were 35.30, 35.20, 32.63, 13.08 and 1.63 (hours), respectively (Fig
329 S1). Although pKWmEB runs slightly longer than KWsBC, pKWmEB has significantly lower
330 FPR than KWsBC.

331 **Real data analysis in *Arabidopsis thaliana***

332 Four flowering time related traits in *Arabidopsis thaliana* derived from Atwell *et al.* (2010) were
333 re-analyzed by pKWmEB, KWmEB, mrMLM and GEMMA. The four flowering time related
334 traits were FLC gene expression (FLC), FRI gene expression (FRI), days to flowering of plants
335 grown in the field (FT Field) and days to flowering growth in greenhouse (FT GH). We also
336 downloaded the results of EMMA from Atwell *et al.* (2010), with the significance criterion of
337 Bonferroni correction ($0.05/p$, p is the number of markers). All the results are listed in Table S23.
338 Results showed that the numbers of SNPs significantly associated with the four traits were 80 for
339 pKWmEB, 77 for KWmEB, 56 for mrMLM and 53 for GEMMA.

340
341 These significantly associated SNPs were used to mine candidate genes associated with the traits.
342 These candidate genes were compared with those in previous studies. All the previously reported
343 genes detected by the above four methods are listed in Table S24. As a result, 23, 16, 10 and 5
344 previously reported genes were found to be in the region of the significantly associated SNPs
345 detected by pKWmEB, KWmEB, mrMLM and GEMMA, respectively (Table S23), indicating
346 that pKWmEB identified the most previously reported genes. Among these known genes, five
347 were identified only by pKWmEB and were not included in the list of the previously reported
348 genes in Atwell *et al.* (2010) (Table 2).

349 Discussion

350 Recently, our group has developed several multi-locus GWAS methods, i.e., mrMLM (Wang *et al.*,
351 2016), FASTmrEMMA (Wen *et al.*, 2017), ISIS EM-BLASSO (Tamba *et al.*, 2017) and
352 pLARmEB (Zhang *et al.*, 2017). Actually, these are parametric methods. As we know,
353 nonparametric GWAS methods are also very useful in GWAS. However, polygenic background in
354 the nonparametric methods isn't controlled, so their FPRs are high. To overcome this issue, we
355 developed pKWmEB in this study. In addition, pKWmEB can find some previously reported
356 genes that aren't detected by parametric methods (Table 2).

357
358 No existing nonparametric methods in GWAS have considered polygenic background control.
359 This leads to the inflation of false positive rate. To overcome this issue, the model transformation
360 of Wen *et al.* (2017) is used to whiten the covariance matrix of the polygenic matrix K and
361 environmental noise. Meanwhile, genotypic incidence matrix and phenotypes are also transferred.
362 Owing to continually transferred genotypic values, it is necessary to change the transferred
363 genotypic values into binary variables (1 and -1) in order to carry out Kruskal-Wallis test. The
364 question is how to conduct this transfer. If the values are larger than their mean or median, the
365 values are transferred into 1. If the values are not larger than their mean or median, the values are
366 transferred into -1. Thus, new incidence values are obtained. These new incidence values along
367 with new phenotypes are used to conduct the Kruskal-Wallis test. Using this test, all the markers
368 potentially associated with the trait are identified. These selected markers are placed into a
369 multi-locus model, and original genotype and phenotype information is used to estimate their
370 effects using empirical Bayes. Thus, true QTNs can be identified. Our results showed that mean
371 threshold is better than median threshold in statistical power (Fig. S3 and Table S22). Although
372 the Kruskal-Wallis test is used in this study, in addition, other nonparametric tests are also
373 available, for example, the Jonckheere-Terpstra test (Terpstra, 1952; Jonckheere, 1954) and
374 Anderson–Darling test (Anderson and Darling, 1952, 1954). As compared with the methods
375 without polygenic background control, the new method demonstrates a significant improvement in
376 statistical power and robustness for QTL detection and in accuracy for QTN-effect estimation.

377

378 In real data analysis, we should consider whether it is necessary to include population structure in
379 the genetic model. Recently, Bulik-Sullivan *et al.* (2015) proposed a linkage disequilibrium score
380 regression test to solve this issue. This method is to test the significance of difference between
381 regression intercept and one. Results showed that population structure should be included in
382 multi-locus model for all the four traits in this study (Table S25). Principal component analysis is
383 also available for this purpose. We also need to consider the heterozygotes. In this case, a
384 heterozygote is coded as zero and the others are the same as those in pKWmEB. If so, there is no
385 significant power difference between the two homozygote genotypes (AA and aa) and the three
386 genotypes (AA, Aa and aa). However, the accuracy of QTN effect estimation significantly
387 decreased as compared with no heterozygotes (Table S20 and S21).

388
389 The current nonparametric GWAS methods are almost a single-locus genome scan analysis, and
390 such a single marker test often requires a Bonferroni correction. To control the experimental error
391 at a genome-wide significance level of 0.01, the significance level for each test should be adjusted
392 as $0.01/p$, which is $1e-8$ if there are one million markers (p). This criterion is too stringent to detect
393 many important loci. To avoid this issue, many multi-locus approaches have been suggested
394 (Segura *et al.*, 2012; Moser *et al.*, 2015; Wang *et al.*, 2016). In these multi-locus approaches, there
395 is no need for such a multiple test correction. At this situation, less stringent critical P-value
396 (approximately $2e-4$, which is the equivalent of LOD=3.0) can be adopted. This is because its FPR
397 is similar to that from single-locus genome scan analysis with a stringent significance criterion.

398
399 In Monte Carlo simulation studies, the estimates of powers for the four QTNs with the same effect
400 size are highly variable. This is different from the situation in quantitative trait locus mapping. To
401 dissect this phenomenon, the simulated datasets in this study were also analyzed by ADGWAS of
402 Yang *et al.* (2014) and Jonckheere-Terpstra test with Bonferroni correction (Liu, 2016). As a result,
403 similar phenomenon was observed as well. This may be due to two reasons. One is about the
404 genotypic datasets, which are derived from the 216130 SNPs in Atwell *et al.* (2010). Several
405 significant correlations of genotypes between a pair of QTNs were observed. This is not similar to
406 ideal segregation populations in linkage analysis. Another is about single-locus genome-wide
407 scanning of nonparametric tests. When KWsBC is implemented in the first simulation experiment,

408 the 85.6, 46.9, 14.2 and 70.9 (%) P-values in the detection of the 2nd, 3rd, 5th and 6th QTNs are
409 between $5e-6$ and 0.01. Owing to the stringent Bonferroni correction criterion, QTN2 and QTN6
410 were not detected in most situations.

411
412 We compared the results in this study with those in Atwell *et al.* (2010), and found that individual
413 previously reported genes are common, for example, *FLA*, *AT4G00690* (similar to *ESD4*,
414 268809/276143 bp on chromosome 4) and *ATARP4* (6371569 bp on chromosome 1) are detected
415 by all the four methods. However, most previously reported genes depend on methods (Table S24)
416 and some previously reported genes are detected only by pKWmEB (Table 2). This indicates that
417 pKWmEB is a complement to the widely-used GWAS methods (such as GEMMA). The possible
418 reason is that each method has its own distinct assumptions.

419 **References**

- 420 Acar EF, Sun L (2013). A generalized Kruskal-Wallis test incorporating group uncertainty with application to
421 genetic association studies. *Biometrics* **69**: 427–435.
- 422 Anderson TW, Darling DA (1954). A test of goodness-of-fit. *J. Am. Stat. Assoc.* **49**: 765–769.
- 423 Anderson TW, Darling DA (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic
424 processes. *Ann. Math. Stat.* **23**: 193–212.
- 425 Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y *et al.* (2010). Genome-wide association study
426 of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. *Nature* **465**: 627–631.
- 427 Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B *et al.* (2008). Whole genome scan detects an allelic variant of
428 *fad2*, associated with increased oleic acid levels in maize. *Molec. Genet. Genom.* **279**: 1–10.
- 429 Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric
430 Genomics Consortium *et al.* (2015). LD score regression distinguishes confounding from polygenicity in
431 genome-wide association studies. *Nat. Genet.* **47**: 291–295.
- 432 Efron B, Hastie T, Johnstone I, Tibshirani, R (2004). Least angle regression. *Ann Statist.* **32**: 407–451.
- 433 Figueiredo MA (2003). Adaptive sparseness for supervised learning. *IEEE. T. Pattern. Anal.* **25**: 1151–1159.
- 434 Filiault DL, Maloof JN (2012). A genome-wide association study identifies variants underlying the *Arabidopsis*
435 *thaliana* shade avoidance response. *PLoS Genet.* **8**: e1002589.
- 436 Holt BF, Boyes DC, Ellerström M, Siefers N, Wiig A, Kauffman S *et al.* (2002). An evolutionarily conserved
437 mediator of plant disease resistance gene function is required for normal *Arabidopsis* development. *Dev. Cell* **2**:
438 807-817.

439 Huang Z, Shi T, Zheng B, Yumul RE, Liu X, You C, Gao Z *et al.* (2016). APETALA2 antagonizes the
440 transcriptional activity of AGAMOUS in regulating floral stem cells in *Arabidopsis thaliana*. *New Phytol.* DOI:
441 10.1111/nph.14151.
442 Izawa T, Takahashi Y, Yano M (2003). Comparative biology comes into bloom: genomic and genetic comparison
443 of flowering pathways in rice and Arabidopsis. *Curr. Opin. Plant. Biol.* **6**: 113–120.
444 Jonckheere AR (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika* **41**:133–145.
445 Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al.* (2008) Efficient control of population
446 structure in model organism association mapping. *Genetics* **178**: 1709–1723.
447 Kolmogorov AN (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto*
448 *Italiano degli Attuari* **4**: 83–91.
449 Kozlitina J, Schucany WR (2015). A robust distribution-free test for genetic association studies of quantitative
450 traits. *Stat. Appl. Genet. Mol. Biol.* **14**: 443–464.
451 Kruskal WH (1952). A nonparametric test for the several sample problem. *Ann. Math. Stat.* **23**: 525–540.
452 Kruskal WH, Wallis WA (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **47**: 583–621.
453 Li J, Zhang J, Wang X, Chen J (2010). A membrane-tethered transcription factor ANAC089 negatively regulates
454 floral initiation in *Arabidopsis thaliana*. *Sci. China Life Sci.* **53**: 1299–1306.
455 Li JH, Dan J, Li CL, Wu RL (2014). A model-free approach for detecting interactions in genetic association
456 studies. *Brief. Bioinform.* **15**: 1057–1068.
457 Li QZ, Li ZB, Zheng G, Gao GM, Yu K (2013). Rank-based robust tests for quantitative-trait genetic association
458 studies. *Genet. Epidemiol.* **37**: 358-365.
459 Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. (2011). FaST linear mixed models for
460 genome-wide association studies. *Nat. Methods* **8**: 833–835.
461 Liu Q (2016). A multi-locus Jonckheere-Terpstra method for genome-wide association study. Master of Science,
462 Nanjing Agricultural University.
463 Mann HB, Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the
464 other. *Ann. Math. Stat.*, **18**: 50–60.
465 Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM (2015). Simultaneous discovery, estimation
466 and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* **11**: e1004969.
467 Price AL, Zaitlen NA, Reich D, Patterson N (2010). New approaches to population stratification in genome-wide
468 association studies. *Nat. Rev. Genet.* **11**: 459–463.
469 Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q *et al.* (2012). An efficient multi-locus mixed-model
470 approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**: 825–830.
471 Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D *et al.* (2007). A genome-wide association study identifies
472 novel risk loci for type 2 diabetes. *Nature* **445**: 881–885.

473 Smirnov N (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **19**:
474 279–281.

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

477 Tan HL, Zain SM, Mohamed R, Rampal S, Chin KF, Basu RC *et al.* (2014). Association of glucokinase regulatory
478 gene polymorphisms with risk and severity of non-alcoholic fatty liver disease: an interaction study with
479 adiponutrin gene. *J. Gastroenterol.* **49**: 1056–1064.

480 Terao C, Ohmura K, Yamada R, Kawaguchi T, Shimizu M, Tabara Y *et al.* (2014). Association between
481 antinuclear antibodies and the HLA class II locus and heterogeneous characteristics of staining patterns.
482 *Arthritis Rheumatol.* **66**: 3395–3403.

483 Terpstra TJ (1952). The asymptotic normality and consistency of Kendalls test against trend, when ties are present
484 in one ranking. *Indagat. Math.* **14**: 327–333.

485 The Wellcome Trust Case Control Consortium (WTCCC) (2007). Genome-wide association study of 14,000 cases
486 of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678.

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

489 Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY *et al.* (2017). Methodological implementation of mixed
490 linear models in multi-locus genome-wide association studies. *Briefings in Bioinformatics*, DOI:
491 10.1093/bib/bbw145.

492 Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics Bull.* **1**: 80–83.

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

495 Yang N, Lu Y, Yang X, Huang J, Zhou Y, Ali F *et al.* (2014). Genome wide association studies using a new
496 nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association
497 panel. *PLoS Genet.* **10**: 821–833.

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

500 Zhang J, Feng JY, Ni YL, Wen YJ, Niu Y, Tamba CL *et al.* (2017). pLARmEB: integration of least angle
501 regression with empirical Bayes for multi-locus genome-wide association studies. *Heredity* **118**: 517–524.

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

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

506 Zhao XY, Wang Q, Li S, Ge FR, Zhou LZ, McCormick S *et al.* (2013). The juxtamembrane and carboxy-terminal
507 domains of *Arabidopsis* PRK2 are critical for ROP-induced growth in pollen tubes. *J. Exp. Bot.* **64**: 5599–5610.
508 Zhou X, Stephens M (2012). Genome-wide efficient mixed model analysis for association studies. *Nat. Genet.* **44**:
509 821–824.

510 DATA ARCHIVING

511 All simulated data sets are available from the Dryad Digital Repository:
512 <http://dx.doi.org/10.5061/dryad.sk652> and supplementary file (Simulated phenotypes [Data Sets](#)).
513 The real data set can be retrieved from: <http://www.arabidopsis.org/>.

514 Acknowledgements

515 This work was supported by the National Natural Science Foundation of China (31571268,
516 U1602261), Huazhong Agricultural University Scientific & Technological Self-innovation
517 Foundation (Program No. 2014RC020) and State Key Laboratory of Cotton Biology Open Fund
518 (CB2017B01).

519 Author Contributions

520 Y.-M.Z. conceived and supervised the study, and improved the manuscript. W.-L.R. and Y.-J.W.
521 performed the experiments, analyzed the data, and wrote the draft. W.-L.R. wrote the R software.
522 J.M.D. improved the language within the manuscript. All authors reviewed the manuscript.

523 Figure Legends

524 **Figure 1. A flow chart of pKWmEB method.**

525

526 **Figure 2. Comparison of statistical powers of six simulated QTNs using five GWAS methods**
527 **(pKWmEB, KWmEB, KWsBC, GEMMA and mrMLM).** (a) no polygenic background; (b) an
528 additive polygenic variance (explaining 0.092 of the phenotypic variance); (c) three epistatic
529 QTNs each explaining 0.05 of the phenotypic variance. Residual error is normal distribution with
530 mean zero and variance 10 in (a) to (c), log-normal distribution with mean zero and standard
531 deviation 1.144 (d), and logistic distribution with mean zero and standard deviation 1.743 (e).

532
533 **Figure 3. Comparison of mean squared errors of each simulated QTN effect using four**
534 **GWAS methods (pKWmEB, KWmEB, GEMMA and mrMLM).** The descriptions in (a) to (e)
535 are the same as those in Fig 2.

536
537 **Figure 4. Comparison of false positive rates using five GWAS methods (pKWmEB, KWmEB,**
538 **KWsBC, GEMMA and mrMLM).** The descriptions in (a) to (e) are the same as those in Fig 2.

539 **Additional information**

540 **Competing financial interests:** The authors declare no competing financial interests.

541 Supplementary information accompanies this manuscript in the file entitled with “Additional
542 information”.

543 **Table 1. Paired t tests and their P-values for power and mean squared error (MSE) between pKWmEB and each of the other four methods in the first**
 544 **simulation experiment**

Case		KWmEB	KWsBC	GEMMA	mrMLM
Power	t -value	2.58	0.60	3.65	1.16
	P-value	0.0495*	0.5760	0.0148*	0.2972
MSE	t -value	-3.76	-	-3.94	-0.96
	P-value	0.0132*	-	0.0110*	0.3824

545 * and **: significances at the 0.05 and 0.01 levels, respectively.

546 **Table 2. Previously reported genes that were identified only by pKWmEB**

Chr	Position (bp)	LOD	Effect	r ² (%)	Gene	Trait	Allele with code 1	Reference
2	2916675	4.90	0.062	0.92	<i>PRK2</i>	FT GH	A	Zhao <i>et al.</i> (2013)
2	10574932	3.23	0.098	1.38	<i>ATCOL3</i>	FT Field	T	Izawa <i>et al.</i> (2003)
4	17392527	3.05	-0.183	2.03	<i>APETALA2</i>	FLC	C	Huang <i>et al.</i> (2006)
5	7372523	3.96	0.122	1.86	<i>ANAC089</i>	FT Field	G	Li <i>et al.</i> (2010)
5	7372523	3.96	0.122	1.86	<i>ATTIP49A</i>	FT Field	G	Holt <i>et al.</i> (2002)

547 The genes in this table were not detected by Atwell *et al.* (2010).