UC Riverside

UC Riverside Previously Published Works

Title

A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations

Permalink

https://escholarship.org/uc/item/4r86g9jx

Journal

Genetics, 202(2)

ISSN

0016-6731

Authors

Wei, Julong

Xu, Shizhong

Publication Date

2016-02-01

DOI

10.1534/genetics.115.179945

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations

Julong Wei*,† and Shizhong Xu*,[1]
*Department of Botany and Plant Sciences, University of California, Riverside, California 92521, and †College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

ABSTRACT Most standard QTL mapping procedures apply to populations derived from the cross of two parents. QTL detected from such biparental populations are rarely relevant to breeding programs because of the narrow genetic basis: only two alleles are involved per locus. To improve the generality and applicability of mapping results, QTL should be detected using populations initiated from multiple parents, such as the multiparent advanced generation intercross (MAGIC) populations. The greatest challenges of QTL mapping in MAGIC populations come from multiple founder alleles and control of the genetic background information. We developed a random-model methodology by treating the founder effects of each locus as random effects following a normal distribution with a locus-specific variance. We also fit a polygenic effect to the model to control the genetic background. To improve the statistical power for a scanned marker, we release the marker effect absorbed by the polygene back to the model. In contrast to the fixed-model approach, we estimate and test the variance of each locus and scan the entire genome one locus at a time using likelihood-ratio test statistics. Simulation studies showed that this method can increase statistical power and reduce type I error compared with composite interval mapping (CIM) and multiparent whole-genome average interval mapping (MPWGAIM). We demonstrated the method using a public *Arabidopsis thaliana* MAGIC population and a mouse MAGIC population.

KEYWORDS best linear unbiased prediction; empirical Bayes; mixed model; polygene; restricted maximum likelihood; multiparental populations; Multiparent Advanced Generation Inter-Cross (MAGIC); MPP

THERE is an urgent need to develop and study multiparent advanced generation intercross (MAGIC) populations (Rakshit *et al.* 2012). Along with nested association mapping populations (Yu *et al.* 2008), the MAGIC population is called a *second-generation mapping resource* (Rakshit *et al.* 2012). Using MAGIC populations to perform QTL mapping was first proposed for mice by Threadgill *et al.* (2002). Such a population is called the *Collaborative Cross* (CC) population (Churchill *et al.* 2004; Collaborative Cross Consortium 2012). Simulation studies showed that an eight-parent CC population with 1000 progenies is capable of increasing mapping resolution to the sub-centimorgan range (Valdar *et al.* 2006). MAGIC populations in *Drosophila melanogaster* are called *Drosophila Synthetic Population Resources* (DSPR) (MacDonald and Long 2007; King *et al.* 2012a, *et al.*b). A review of MAGIC populations in crops can be found in Huang *et al.* (2015). The first plant MAGIC population was developed in *Arabidopsis thaliana* by Kover *et al.* (2009). The population will be described later. Subsequently, MAGIC populations have been developed in wheat (Huang *et al.* 2012; Mackay *et al.* 2014), rice (Bandillo *et al.* 2013), and other crop species (Gaur *et al.* 2012; Pascual *et al.* 2015; Sannemann *et al.* 2015). One key difference between MAGIC populations and other multiparent populations is that all MAGIC lines have experienced multiple generations of inbreeding and thus all are inbred lines. As a result, they are also considered genetic reference populations whose particular genome arrangement can be replicated indefinitely. MAGIC populations in plants undoubtedly will become more popular in the future of plant genetics and breeding (Varshney and Dubey 2009; Rakshit

Copyright © 2016 by the Genetics Society of America
doi: 10.1534/genetics.115.179945
Manuscript received June 26, 2015; accepted for publication December 15, 2015; published Early Online December 29, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179945/-/DC1
[1]Corresponding author: Department of Botany and Plant Sciences, University of California, Riverside, CA 92521. E-mail: shizhong.xu@ucr.edu

et al. 2012; Huang *et al.* 2015), which calls attention to the need for improvements in statistical methods to analyze and interpret data derived from these populations. A recent call for papers on QTL mapping in MAGIC populations by *GENETICS* and *G3* (http://www.genetics.org/) further indicates the urgent need for new technologies in MAGIC population QTL mapping.

Current methods of QTL mapping for MAGIC populations are adopted primarily from methods used in biparental populations. For example, composite interval mapping (CIM) (Zeng 1994), originally developed for biparental populations, has been used in QTL mapping for MAGIC populations to control genomic background. Other methods and programs of QTL mapping in MAGIC populations include MCQTL (Jourjon *et al.* 2005), R/qtl (Broman *et al.* 2003), R happy (Mott *et al.* 2000), and R/mpMap (Huang and George 2011), most of which have an option to perform CIM. However, there is an intrinsic limitation in cofactor selection, which is more problematic in MAGIC populations than in biparental populations. In an eight-parent-initiated MAGIC population, each marker has $8 - 1 = 7$ founder effects to estimate. The total number of effects will soon saturate the linear model as the number of cofactors increases. For example, a MAGIC population of size 500 will allow only fewer than $500/7 \approx 71$ cofactors to be included in the model. When the number of cofactors is small, the CIM procedure is sensitive to the selection of cofactors. Ideally, a model should include all markers in a single model. However, when the marker density is high, genome scanning (a single-QTL model) provides a better alternative method for QTL mapping, but the cofactors should be replaced by a polygenic effect, as done in genome-wide association studies (GWAS) (Yu *et al.* 2006). We recently developed a QTL mapping procedure by fitting a polygene using a marker-inferred relationship matrix (replacing cofactors) and demonstrated the robustness of the method (Xu 2013b).

Recently, Gatti *et al.* (2014) developed a mixed model for QTL mapping in Diversity Outbred (DO) mice by treating the effects of scanned markers as fixed and a polygenic effect as random. The polygenic effect essentially replaced cofactors to control the genetic background. The method tends to have a low power because part of the effect of the marker currently scanned is absorbed by the polygene. Our simulation studies showed that dramatic improvement can be achieved in terms of resolution and statistical power of mapped QTL if the effect of the current QTL captured by the polygene is taken into account. Verbyla *et al.* (2014) developed a multiple-QTL model for QTL mapping in MAGIC populations. The method is called *multiparent whole-genome QTL analysis* (MPWGAIM), and several steps are involved in selecting markers for inclusion in the model. First, a polygenic base model is implemented to detect the whole-genome effect on the traits of interest. If the polygenic variance is significantly larger than zero, then markers are subject to selection under a random-model approach; *i.e.*, the founder allelic effects of a marker are treated as random effects, and the variance of those founder effects is estimated and the marker is then selected if the variance is sufficiently large. The final model will include all markers selected (forward selection). This is a variable-selection approach and may be costly if the number of markers and the number of QTL found are large. We will treat this model as the "gold standard" for simulation and comparison. Another recent study of QTL mapping in MAGIC populations is the Bayesian modeling of haplotype effects (Zhang *et al.* 2014), where the founder haplotype effects are estimated via Markov chain Monte Carlo (MCMC) sampling or importance sampling (IS). One important feature of the Bayesian method is the ability to handle uncertainty of the founder allelic inheritance. The only concern with the Bayesian method is the high computational cost when the sample size and the number of markers are very large because Monte Carlo sampling is involved. It is recommended to use the Bayesian method to fine-tune the model after markers are selected using some simple methods such as interval mapping (IM) and CIM.

Table 1 Information for the seven simulated QTL using genotypes of the first MAGIC population of mice

QTL	Chromosome	Position (cM)	Bin	Variance[a]	Proportion[b]
QTL-1	1	41.35	209	0.10	0.046
QTL-2	2	21.16	602	0.20	0.092
QTL-3	3	58.79	1313	0.30	0.138
QTL-4	3	65.18	1348	0.30	0.138
QTL-5	4	27.42	1564	0.40	0.185
QTL-6	4	41.19	1641	0.40	0.185
QTL-7	5	28.65	1994	0.10	0.046

[a] Variance of a QTL, which is defined as $\mathrm{var}(\mathbf{Z}_k\boldsymbol{\gamma}_k)$, and the variance is taken across all individuals in the MAGIC population.
[b] Proportion of the total phenotypic variance explained by the QTL.

In this study, we extended the mixed-model methodology of QTL mapping in MAGIC populations by fitting a polygenic effect as random and a scanned marker effect either as fixed or random. Furthermore, we released the polygenic counterpart of a scanned marker effect back to the model to avoid competition between the marker effect and its polygenic counterpart. This improved mixed-model methodology has significantly improved the statistical power of QTL detection. We used a CC mouse population (Collaborative Cross Consortium 2012) to perform simulations to examine the properties of the new methods (there are no phenotypic values available for the CC mouse population). The *Arabidopsis* MAGIC population of Kover *et al.* (2009) and the pre-CC mouse population of Rutledge *et al.* (2014) were reanalyzed using the new methods to demonstrate the differences between the new and existing methods.

Materials and Methods

MAGIC populations

Three MAGIC populations were used in this study to demonstrate the new methods of QTL mapping, two populations in mice and one in *A. thaliana*. The first MAGIC population in mice does not have phenotypes available on the website (http://www.csbio.unc.edu/CCstatus) and was used only

Table 2 Founder effects for the seven simulated QTL using genotypes of the first MAGIC population of mice

			Founder name[a]							
QTL	Chr.	Position	A/J	C57BL	129S1	NOD	NZO	CAST	PWK	WSB
QTL-1	1	41.35	−0.174	−0.015	0.145	−0.409	0.046	−0.281	−0.058	−0.073
QTL-2	2	21.16	−0.473	−0.095	−0.063	0.352	0.052	0.161	−0.074	0.303
QTL-3	3	58.79	0.21	−0.181	−0.398	−0.414	0.391	−0.422	−0.089	−0.174
QTL-4	3	65.18	−0.294	0.116	0.58	0.067	−0.111	0.172	−0.443	0.267
QTL-5	4	27.42	0.549	0.113	0.595	0.266	−0.13	0.161	−0.43	−0.265
QTL-6	4	41.19	−0.287	0.225	−0.027	0.104	−0.123	−0.227	0.809	0.061
QTL-7	5	28.65	−0.252	−0.042	0.042	−0.083	0.028	0.346	−0.202	0.132

[a] Strain names of the 8 founder strains initiating the CC population.

for simulation studies. The second MAGIC population of mice has both genotype and phenotype information and was used as a real application example. The MAGIC population in *A. thaliana* also has both genotype and phenotype information and was reanalyzed to compare the results of the different methods.

First MAGIC population of mice: This MAGIC population is called the *CC population* (Churchill *et al.* 2004). The genotype data were published by the Collaborative Cross Consortium (2012). No phenotype information is available in the 458 CC mice, and thus the data were used only for simulation study. The CC population is an eight-parent MAGIC population derived from a funnel mating design. We downloaded the recombination breakpoint data of 19 autosomes from 458 CC mice posted on the University of North Carolina (UNC) System Genetics website (http://www.csbio.unc.edu/CCstatus). Using the breakpoint information, we inferred 6683 bins (intact chromosome segments). A bin is defined as a segment that contains no breakpoints across all lines within the segment. Within a bin, all markers segregate in exactly the same pattern across lines (perfect LD). Therefore, a single marker can represent the whole bin. For detailed information on bin data analysis, see Xu (2013a). The bin data are available in Supporting Information, File S1.

Second MAGIC population of mice: The second MAGIC population was derived from the same eight parents as the first CC population, but the CC mice were not fully inbred, and therefore, the population is called the *pre-CC population*. The data were obtained from Rutledge *et al.* (2014) and consist of 151 individuals. This data set includes 27,039 SNPs evenly distributed among the 20 chromosomes (including the *X* chromosome). Probabilities of the parental origins of the SNPs were calculated using the HAPPY program based on the hidden Markov model (Mott *et al.* 2000). In the original study of this population, the authors focused on two traits associated with severe asthma and decrements in lung function, including airway polymorphonuclear neutrophil (PMN) recruitment and the concentration of CXCL1 in lung lavage fluid. Here we reanalyzed the first trait, PMN.

MAGIC population of Arabidopsis: The MAGIC population of *A. thaliana* (Kover *et al.* 2009) consists of 527 lines descended from a heterogeneous stock of 19 intermated parents. These lines and the 19 founders were genotyped with 1260 SNP markers [minor allele frequency (MAF) $> 5\%$] and phenotyped for two development-related traits, the number of days between bolting and flowering (DBF) and growth rate (GR), where GR was measured as the residual of regression by fitting the number of leaves to the number of days to germination. The 527 lines were derived from the 19 founder accessions of *A. thaliana*, intermating for four generations, and then inbreeding for six additional generations, forming nearly homozygous lines. The authors further updated the database after the initial publication. We downloaded the updated genotypes and phenotypes from http://mus.well.ox.ac.uk/magic/. There were only 426 lines having both the genotype and phenotype information. In this analysis, we included the 426 lines and 1254 markers distributed among five chromosomes (total length of the genome is 118 Mb). The founder strain probabilities for all loci were calculated using the HAPPY program. We analyzed both DBF and GR.

Statistical methods

Polygenic model: The polygenic model is the null model used to scan the entire genome for QTL identification. We now use an eight-parent MAGIC population as an example to demonstrate the model. The method holds for any *p*-parent MAGIC populations. Let $\boldsymbol{y}$ be an $n \times 1$ vector of phenotypic values for n individuals. Define $\mathbf{Z}_k$ as an $n \times 8$ matrix of founder allele inheritance indicators for locus k. The jth row of matrix $\mathbf{Z}_k$ is defined as a 1×8 vector. If this individual is a heterozygote carrying the first and second founder alleles, then we define

$$\mathbf{Z}_{jk} = [1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

If the individual is a homozygote inheriting both alleles from the fifth founder, then $\mathbf{Z}_{jk}$ is defined as

$$\mathbf{Z}_{jk} = [0 \quad 0 \quad 0 \quad 0 \quad 2 \quad 0 \quad 0 \quad 0]$$

The general rule for defining $\mathbf{Z}_{jk}$ is that there are at most two nonzero elements, and the sum of all eight elements equals 2. We then define the following polygenic model, which is the null model used to test significance of an individual marker:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{1}$$

Table 3 Statistical powers for the seven simulated QTL and FDR drawn from 1000 replicated simulation experiments

Method	QTL-1	QTL-2	QTL-3	QTL-4	QTL-5	QTL-6	QTL-7	FDR
FIXED-A	0.001	0.699	0.827	0.911	0.985	0.628	0.003	0.0015
FIXED-B	0.003	0.868	0.919	0.957	0.993	0.750	0.014	0.0034
RANDOM-A	0.002	0.704	0.849	0.916	0.985	0.636	0.004	0.0014
RANDOM-B	0.003	0.868	0.923	0.959	0.993	0.754	0.014	0.0036
MPWGAIM	0.150	0.500	0.690	0.700	0.670	0.770	0.340	0.0179
IM	0.003	0.352	0.326	0.355	0.494	0.389	0.001	0.0608
CIM-30	0.002	0.476	0.620	0.761	0.986	0.754	0.015	0.0113
CIM-50	0.001	0.136	0.043	0.015	0.512	0.009	0.003	0.0317
CIM-65	0.000	0.000	0.001	0.000	0.000	0.002	0.000	0.940

where $\mathbf{X}$ is an $n \times r$ design matrix for r fixed effects; $\boldsymbol{\beta}$ is an $r \times 1$ vector for the r fixed effects; $\boldsymbol{\xi}$ is an $n \times 1$ vector of polygenic effects with an assumed multivariate normal distribution $\boldsymbol{\xi} \sim N(0, \mathbf{K}\phi^2)$, where $\mathbf{K}$ is a marker-derived kinship matrix and ϕ^2 is a polygenic variance; and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2)$ is a vector of residual errors with an unknown error variance σ^2. The marker inferred kinship matrix is defined as

$$\mathbf{K} = \frac{1}{d}\sum_{k=1}^{m} \mathbf{Z}_k \mathbf{Z}_k^{T} \tag{2}$$

where $d = (1/n)\mathrm{tr}(\sum_{k=1}^{m} \mathbf{Z}_k \mathbf{Z}_k^{T})$ is a normalization factor. The expectation of y is $E(y) = \mathbf{X}\boldsymbol{\beta}$, and the variance-covariance matrix is

$$\mathrm{var}(\mathbf{y}) = \mathbf{K}\phi^2 + \mathbf{I}\sigma^2 = (\mathbf{K}\lambda + \mathbf{I})\sigma^2 = \mathbf{H}\sigma^2 \tag{3}$$

where $\lambda = \phi^2/\sigma^2$ is the variance ratio, and $\mathbf{H} = \mathbf{K}\lambda + \mathbf{I}$. After absorbing $\boldsymbol{\beta}$ and σ^2, the restricted maximum likelihood is only a function of λ, which is

$$L(\lambda) = -\frac{1}{2}\ln|\mathbf{H}| - \frac{1}{2}\ln|\mathbf{X}^{T}\mathbf{H}^{-1}\mathbf{X}| - \frac{n-r}{2}\ln(\mathbf{y}^{T}\mathbf{P}\mathbf{y}) \tag{4}$$

where

$$\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{H}^{-1} \tag{5}$$

The restricted maximum likelihood solution of λ was obtained by maximizing the preceding likelihood function using the Newton iteration algorithm. The eigen-decomposition algorithm proposed by Kang *et al.* (2008) was used to evaluate the likelihood function for fast computation. The estimated variance ratio is denoted by $\widehat{\lambda}$ and will be used as a known constant in the genomic scanning model that follows. File S2 describes the method of estimating λ along with the effect of the marker scanned, the so-called exact method (Zhou and Stephens 2012).

Fixed model: To test the significance of the kth marker, we first used the fixed-model approach proposed by Gatti *et al.* (2014). The model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{Z}_k$ is the allelic inheritance matrix for marker k, as defined earlier, and

$$\boldsymbol{\gamma}_k = [\,\gamma_{1k} \quad \gamma_{2k} \quad \gamma_{3k} \quad \gamma_{4k} \quad \gamma_{5k} \quad \gamma_{6k} \quad \gamma_{7k} \quad \gamma_{8k}\,]^{T} \tag{7}$$

is an 8×1 vector for the eight founder allelic effects. Under this model, the $\boldsymbol{\gamma}_k$ vector is assumed to be fixed effects. The model is in fact a mixed model because it contains both fixed and random effects. We call it the *fixed model* because later on we will treat $\boldsymbol{\gamma}_k$ as random effects. Under the fixed model, we can only estimate and test seven (8 – 1 = 7) effects by deleting the last founder allele from the model. The maximum-likelihood method was used to estimate $\boldsymbol{\gamma}_k$, and the result turned out to be identical to the weighted-least-squares estimate after premultiplying all variables ($\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{Z}_k$) by the eigenvectors of the $\mathbf{K}$ matrix and the weight for the jth individual being $W_j = 1/(\delta_j\widehat{\lambda} + 1)$, where δ_j is the jth eigenvalue of the $\mathbf{K}$ matrix (Xu 2013b). Note that $\widehat{\lambda}$ is the estimated variance ratio under the polygenic model, as described earlier. This method is called the *approximate method* (Zhou and Stephens 2012). The likelihood-ratio test was used as the test statistic and is defined as

$$\Gamma_k = -2\left[L_0\left(\widetilde{\beta}, \widetilde{\sigma}^2\right) - L_1\left(\widehat{\beta}, \widehat{\gamma}_k, \widehat{\sigma}^2\right)\right] \tag{8}$$

where L_0 is the log likelihood function under the null model ($\boldsymbol{\gamma}_k = 0$), and L_1 is the log likelihood function under the alternative model. Note that the estimated β and σ^2 under the two models are different. The *P*-value was calculated from the chi-square distribution with seven degrees of freedom. This method is called *FIXED-A* when compared with other methods.

Recall that $\mathbf{Z}_k$ contributes to the calculation of the kinship matrix $\mathbf{K}$, as shown in equation 2. Although not explicitly estimated in the polygene, the effect of marker k has a polygenic counterpart that may compete with the estimated $\boldsymbol{\gamma}_k$ when marker k is scanned. Let $\widehat{\boldsymbol{\xi}}_k = \boldsymbol{Z}_k\widehat{\boldsymbol{a}}_k$ be the estimated polygenic effect contributed by marker k, and $\widehat{\boldsymbol{a}}_k$ is the estimated effect for this marker under the polygenic model. We can release this effect from the polygene back to the model to avoid this competition. The revised model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_k + \boldsymbol{\varepsilon} \tag{9}$$

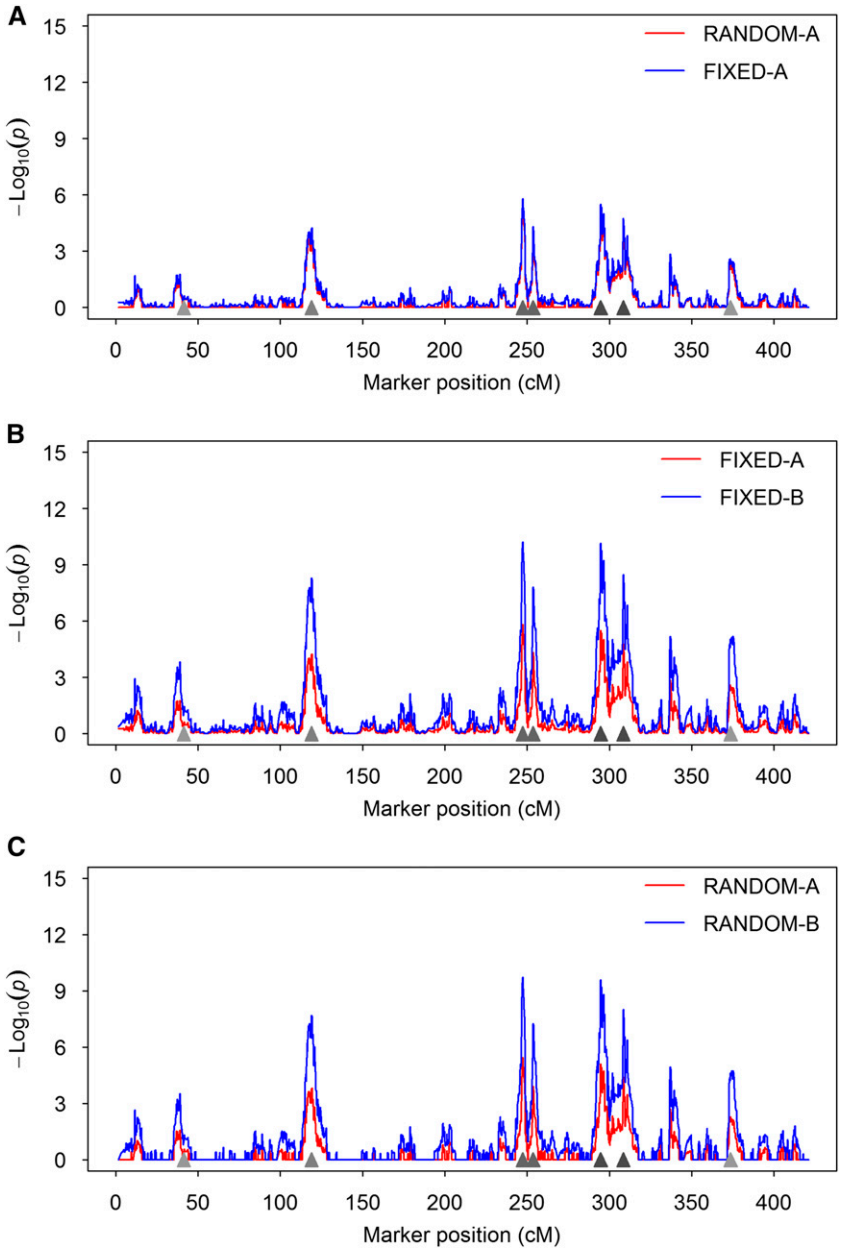

Figure 1 Test-statistic profiles of different methods from a simulated data set. The test statistics are presented as $-\log_{10}(P)$. Locations of the simulated QTL are represented by the filled triangles on the *x*-axis. This figure demonstrates the common behaviors of the different methods that are expected in a real data analysis. (A) Comparison between RANDOM-A and FIXED-A. (B) Comparison between FIXED-A and FIXED-B. (C) Comparison between RANDOM-A and RANDOM-B.

Rearranging this equation leads to

$$\mathbf{y} + \widehat{\boldsymbol{\xi}}_k = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{10}$$

Further defining $\mathbf{y}_k = \mathbf{y} + \widehat{\boldsymbol{\xi}}_k$, we now have a new model

$$\mathbf{y}_k = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_k\boldsymbol{\gamma}_k + \boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{11}$$

which is the same as equation 6 except that the $\mathbf{y}$ vector changes every time a marker is scanned. Note that $\widehat{\boldsymbol{\xi}}_k$, the polygenic component from marker k, is calculated only once under the null model. Therefore, this revised method does not present much additional computational burden. The method to obtain $\widehat{\boldsymbol{\xi}}_k$ is called the *best linear unbiased prediction* (BLUP) and is described in File S2. This revised method is called *FIXED-B* when compared with other methods.

Random model: The fixed-model approach may not be stable when the number of founders is large (Gatti *et al.* 2014), and the design matrix $\mathbf{Z}_k$ may have variable ranks across different markers. Under the null model, the likelihood-ratio test statistic follows a chi-square distribution with degrees of freedom depending on the number of founders. We propose to treat the eight founder effects as random variables following a normal distribution with mean zero and a common variance. Although it is still a mixed model, we call it a *random model* to distinguish it from the fixed model described earlier. The linear model remains the same as equation 6, but $\boldsymbol{\gamma}_k \sim N(0, \mathbf{I}_8\phi_k^2)$ is assumed, where ϕ_k^2 is a locus-specific variance. The expectation of $\mathbf{y}$ remains $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, and the variance-covariance matrix is

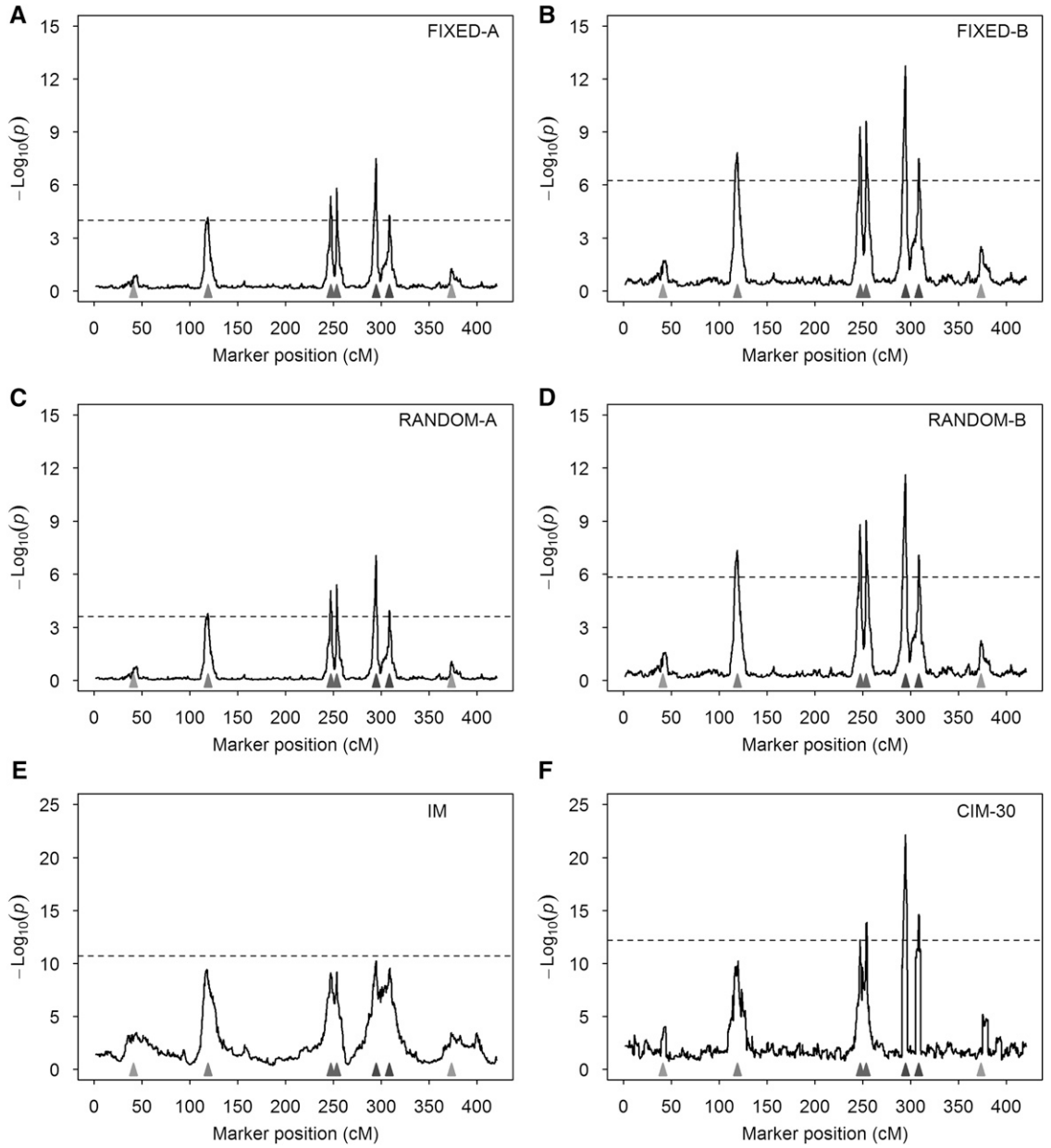

Figure 2 Average test-statistic profiles [$-\log_{10}(P)$] of six methods from 1000 replicated simulation experiments. The horizontal dashed lines represent the 95% thresholds drawn from 1000 simulated samples under the null model. The true locations of the seven simulated QTL are represented by the filled triangles on the *x*-axis. (A) Result of FIXED-A from the simulated data. (B) Result of FIXED-B from the simulated data. (C) Result of RANDOM-A from the simulated data. (D) Result of RANDOM-B from the simulated data. (E) Result of IM from the simulated data. (F) Result of CIM-30 from the simulated data.

$$\begin{aligned}\mathrm{var}(\mathbf{y}) &= \mathbf{Z}_k\mathbf{Z}_k^{\mathrm{T}}\phi_k^2 + \mathbf{K}\phi^2 + \mathbf{I}\sigma^2 = \mathbf{Z}_k\mathbf{Z}_k^{\mathrm{T}}\phi_k^2 + (\mathbf{K}\lambda + \mathbf{I})\sigma^2 \\ &= \mathbf{Z}_k\mathbf{Z}_k^{\mathrm{T}}\phi_k^2 + \mathbf{H}\sigma^2\end{aligned} \tag{12}$$

where λ in $\mathbf{H}$ is replaced by the estimated value under the polygenic model. A restricted-maximum-likelihood (REML) estimate of ϕ_k^2 is obtained by maximizing the restricted likelihood function. Woodbury matrix identities (Golub and Van Loan 1996) are applied along with the eigen-decomposition to ease the computational burden (File S2). The null hypothesis for marker k is $\phi_k^2 = 0$, which is tested using the likelihood-ratio test

$$\Gamma_k = -2\left[L_0\left(\widetilde{\beta}, \widetilde{\sigma}^2\right) - L_1\left(\widehat{\beta}, \widehat{\phi}_k^2, \widehat{\sigma}^2\right)\right] \tag{13}$$

Under the null model, this test statistic follows approximately a mixture of χ_0^2 and χ_1^2 distributions with an equal weight (Chernoff 1954; Visscher 2006). This method is called *RANDOM-A* when compared with other methods.

We also developed a revised version of the random model by avoiding competition between the current marker scanned and its polygenic counterpart using model 11 as we did for the fixed model. This revised random model is called *RANDOM-B* to distinguish it from other methods.

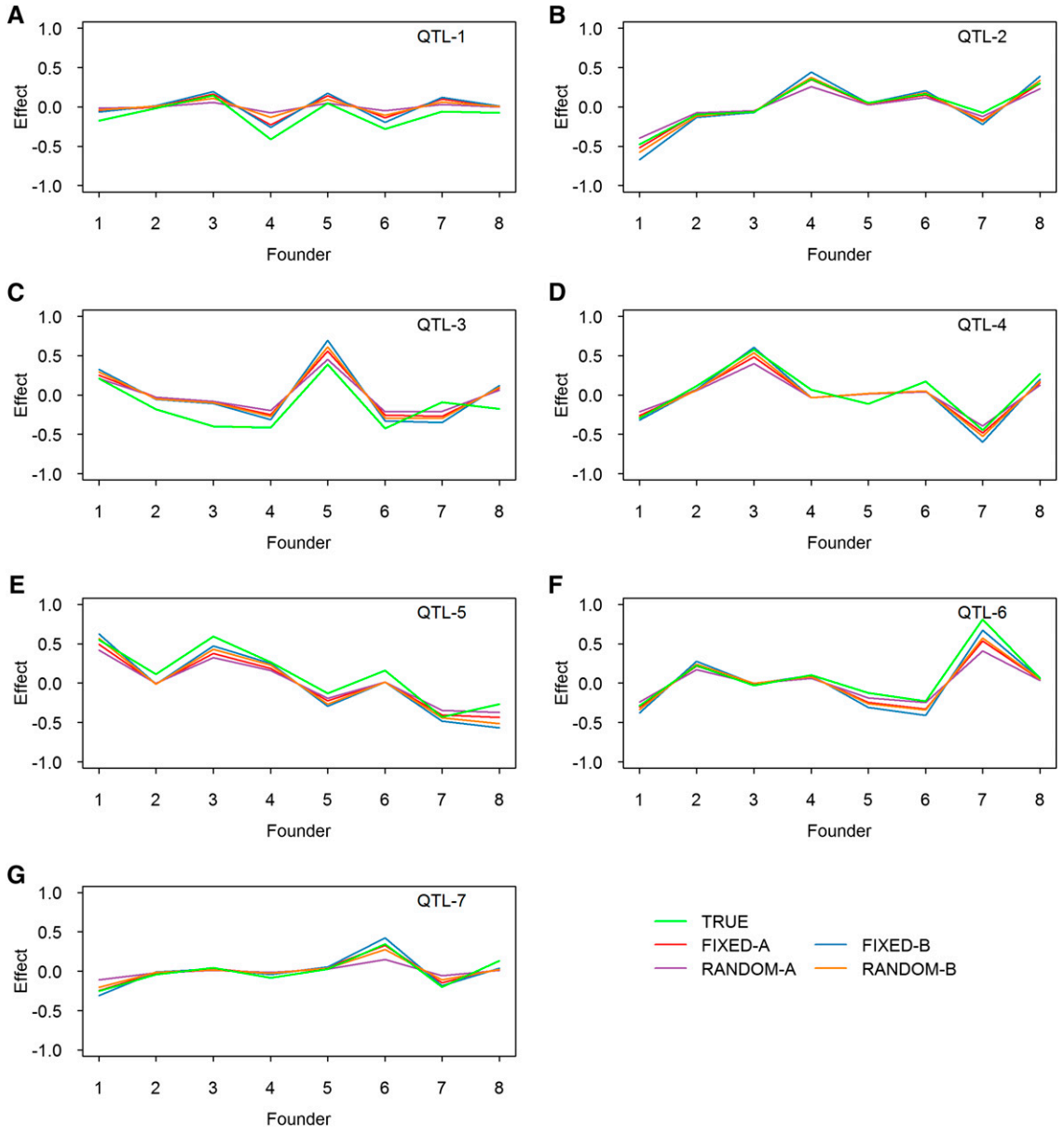

Figure 3 True and estimated allelic effects of eight founders for seven simulated QTL in the simulation experiment. The estimated effects are the average effects of 1000 replicated experiments. Results from four methods are presented: FIXED-A, FIXED-B, RANDOM-A, and RANDOM-B. (A) Effect of QTL-1. (B) Effect of QTL-2. (C) Effect of QTL-3. (D) Effect of QTL-4. (E) Effect of QTL-5. (F) Effect of QTL-6. (G) Effect of QTL-7.

Multiparent whole-genome average interval mapping (MPWGAIM): Here we also performed the analysis using the MPWGAIM approach proposed by Verbyla *et al.* (2014) for comparison using their R package mpwgaim. In the mpwgaim package, only detected markers are reported without test statistics attached. For comparison with our methods, we calculated the Wald test statistics of detected markers based on their estimated effects and variances and then obtained the *P*-value from the chi-square distribution with $8 - 1 = 7$ degrees of freedom. For the simulated data analysis, we also applied the MPWGAIM method. The empirical critical value for hypothesis test was inferred from multiple (1000) simulations under the null model. The 95th percentile of the highest Wald test from each of the multiple simulations was chosen as the empirical critical value. The *P*-value was transformed by $-\log_{10}$ and used to determine whether or not a marker exceeds the empirical critical value.

IM and CIM: IM (Lander and Botstein 1989) and CIM (Zeng 1994) also were used to analyze the data to compare the results with the new methods. These two methods are called *IM* and *CIM-x*, respectively, where *x* indicates the number of cofactors included in the model for background control. The statistical model for IM differs from model 6 by ignoring the polygenic effect. The model for CIM differs from model 6 by replacing the polygenic effect with selected cofactors. The IM

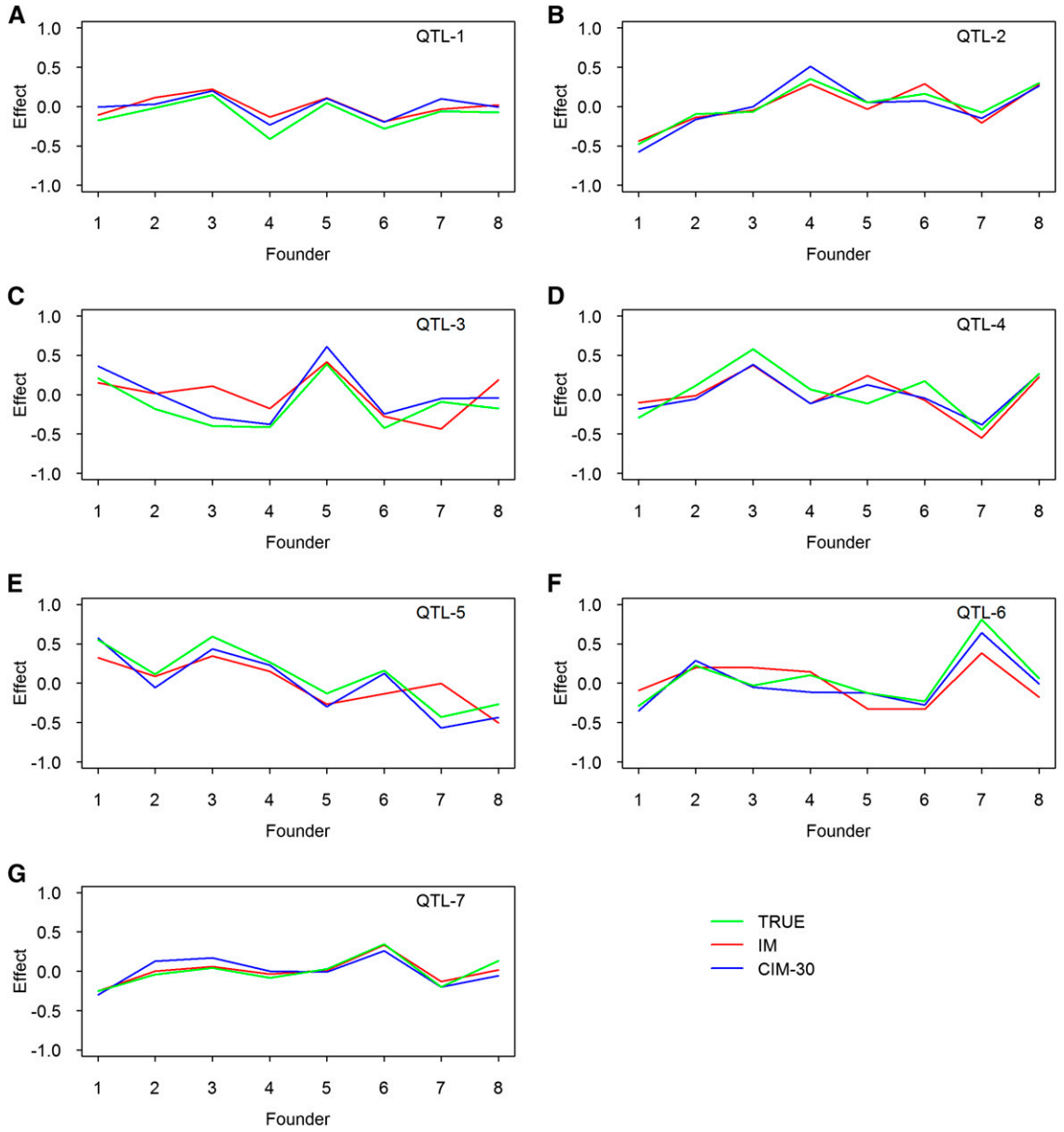

Figure 4 True and estimated allelic effects of eight founders for seven simulated QTL in the simulation experiment. The estimated effects are the average effects of 1000 replicated experiments. Results from two methods are presented: IM and CIM-30, where -30 means that 30 markers are used as cofactors. (A) Effect of QTL-1. (B) Effect of QTL-2. (C) Effect of QTL-3. (D) Effect of QTL-4. (E) Effect of QTL-5. (F) Effect of QTL-6. (G) Effect of QTL-7.

method was implemented in the HAPPY program (Mott *et al.* 2000). The CIM method was implemented using our own R program. For the CIM-x method, the number of cofactors x was set at the following levels for the first MAGIC population of mice: 65, 50, and 30. For a sample size of 458, the maximum number of cofactors cannot be higher than $458/7 \approx 65$; otherwise, there will not be any degrees of freedom left to estimate the residual error variance. For the second MAGIC population of mice (the pre-CC population), the number of cofactors was set at 20, 10, and 5. The population size is 151, and thus the number of cofactors cannot be higher than $151/7 \approx 20$. For the *Arabidopsis* population, the number of cofactors was set at 20, 15, and 10. The maximum number of possible cofactors cannot be greater than $428/18 \approx 23$. The likelihood-ratio test statistic also was used for the IM and CIM methods.

P-value and permutation: We now have a total of seven methods to compare: FIXED-A, MPWGAIM, IM, and CIM are existing methods, and FIXED-B, RANDOM-A, and RANDOM-B are new methods proposed in this study. The P-value of a marker was calculated from the central chi-square distribution with $8 - 1 = 7$ degrees of freedom for the two mouse populations and $19 - 1 = 18$ degrees of freedom for the *Arabidopsis* population under the FIXED-A, FIXED-B, MPWGAIM, IM, and CIM methods. For the RANDOM-A and

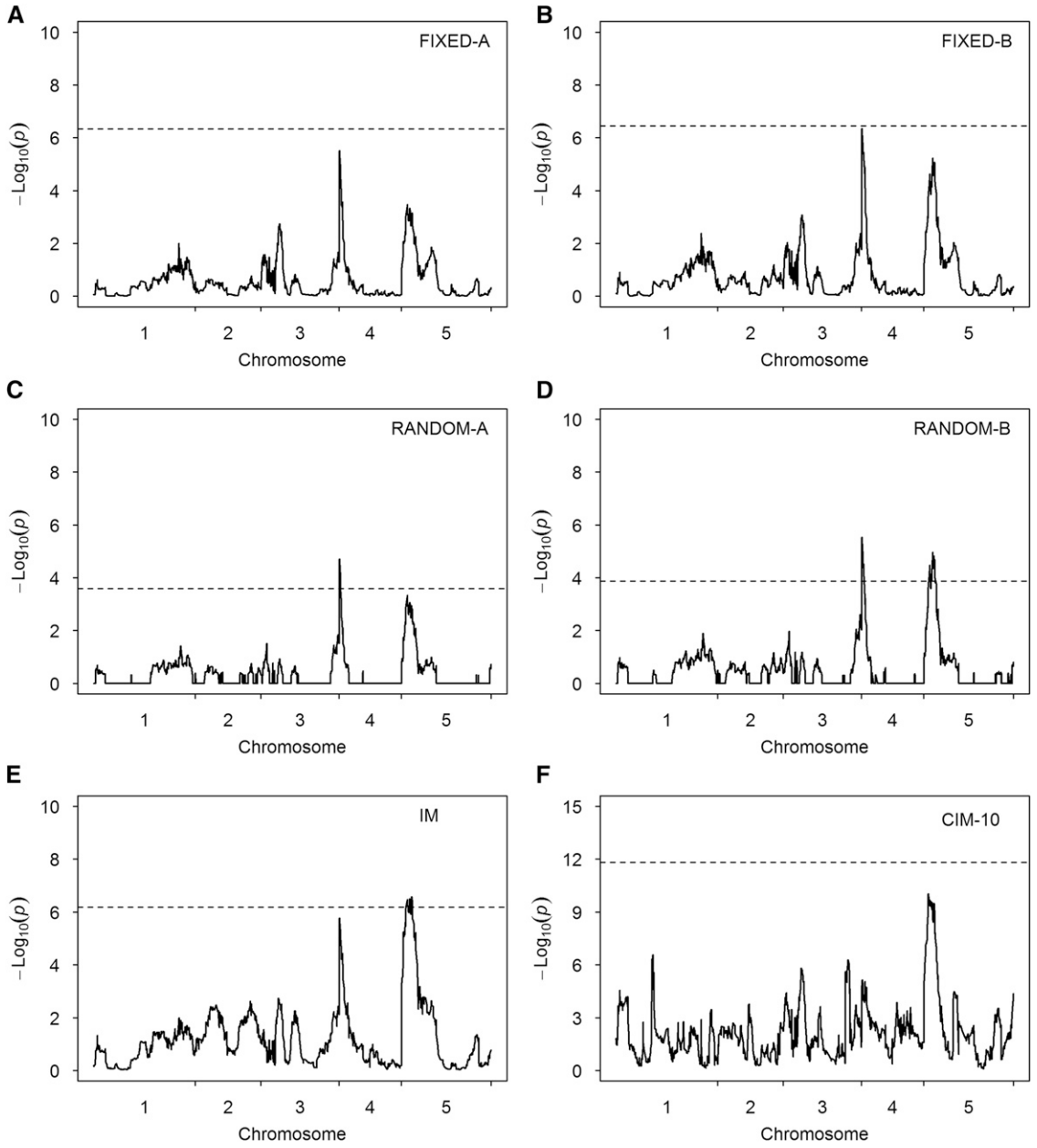

Figure 5 Test-statistic profiles $[-\log_{10}(P)]$ for DBF of the *A. thaliana* MAGIC population obtained from six methods. The horizontal dotted lines represent the 95% thresholds generated from 1000 permuted samples. (A) Result of the FIXED-A method. (B) Result of the FIXED-B method. (C) Result of the RANDOM-A method. (D) Result of the RANDOM-B method. (E) Result of the IM method. (F) Result of the CIM-10 method.

RANDOM-B methods, the *P*-value for each marker was calculated from a mixture of two chi-square distributions, denoted by $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, where χ_0^2 is just a fixed number of 0 (Chernoff 1954; Visscher 2006). Let P_k be the *P*-value for marker *k*, it was calculated using

$$P_k = \begin{cases} 1 & \Gamma_k = 0 \\ \frac{1}{2}\Pr(\chi_1^2 > \Gamma_k) & \Gamma_k > 0 \end{cases} \quad (14)$$

where Γ_k is the likelihood-ratio test statistic calculated using equation 13, and χ_1^2 is a chi-square variable with one degree of freedom. In the real data analysis, we permuted the data 1000 times to generate a null distribution of the test statistics $[-\log_{10}(P)]$. From this null distribution, we determined the 95% quantile and used it as an empirical critical value of a test statistic. A marker with the test statistic $[-\log_{10}(P)]$ greater than this critical value is claimed to be significant at the 0.05 genome-wide type I error rate. For the IM and CIM methods, a permuted sample was generated by randomly shuffling the phenotypes and keeping the genotypes intact. For the four methods with polygenic background control, the labels of the kinship matrix go with the reshuffled phenotypes

Table 4 Estimated variance components and heritability of two traits in the *A. thaliana* population and one trait in the mouse population

Population	Trait	Polygenic variance	Residual variance	Heritability
A. thaliana	Bolt to flower	2.187	4.203	0.342
	Growth rate	1.989	2.215	0.473
Mouse	PMN	1.373	1.127	0.549

so that the polygenic covariance structure remains the same as that in the original data set. This kind of permutation will not destroy the polygenic variance (Cheng and Palmer 2013). Note that permutation was used only in real data analysis to generate empirical critical values for significance tests. In power calculation of the simulated data analysis, empirical critical values were generated from multiple simulations under the null model.

Simulation experiment

The simulation experiment was conducted based on the genotypic data of the first MAGIC population of mice (the CC population). As a result, the sample size was fixed at 458. We used genotypes of the first five chromosomes as the true genotypes to conduct the simulation experiment. The five chromosomes contain 490, 503, 428,423, and 406 bins, respectively, leading to a total of 2250 bins. The design of the simulated QTL mimicked closely that of Verbyla *et al.* (2014). We simulated a total of seven QTL distributed on the five chromosomes. Information about the seven simulated QTL is shown in Table 1. The simulated allelic effects of the eight founders are given in Table 2. The polygenic and residual error variances were set at $\phi^2 = 0.5$ and $\sigma^2 = 0.5$, respectively. The seven QTL collectively have a total variance of 1.1752, which is partitioned into the sum of variances for all seven QTL (1.80) plus twice the sum of all covariances -0.6248 ($1.80 - 0.6248 = 1.1752$). The total phenotypic variance is $1.1752 + 0.5 + 0.5 = 2.1752$. Therefore, the proportion contributed to the phenotypic variance by all seven QTL is $1.1752/2.1752 = 0.5403$. The proportion of the polygenic variance contributed to the phenotypic variance is $0.5/2.1752 = 0.2298$. The total genetic contribution (QTL + polygene) is $0.5403 + 0.2298 = 0.7701$. QTL-1 and -7 are small in terms of the proportions contributed to the trait phenotypic variance. The remaining four QTL are relatively large.

Under the preceding parameter setups, we generated 1000 independent data sets to evaluate the empirical powers under a 0.05 type I error. We also generated 1000 additional data sets under the null model (no QTL were simulated but the polygene). Results of the data analysis from the null model were used to generate the empirical distribution of the test statistics $[-\log_{10}(P)]$ and draw the empirical thresholds of the test statistics for hypothesis tests. The statistical powers from the 1000 replicated simulation experiments were reported by comparing the results with the empirically drawn thresholds of the test statistics. For each simulated QTL, a $\pm$5-cM window around the true position was reserved for power calculation, as done by Verbyla *et al.* (2014). If any bin within this window was detected, the QTL covered by this window was claimed to be detected. Any detected bins beyond this window were counted as false positives. All seven methods mentioned earlier were used to analyze the simulated data. The empirical powers were compared for the seven methods.

Data availability

The new methods of QTL mapping for MAGIC populations were implemented in an R package called *MagicQTL*, which is provided in the Supporting Information and downloadable from the journal article website (see File S3 for the R package and File S4 for the user instruction of the R package). R codes for data simulation, data preparation, and data analysis are downloadable from https://github.com/JulongWei/MagicQTL. This website also provides the R code for calling the MPWGAIM package.

Results

Simulation studies

Statistical powers and false discovery rate (FDR): The empirical statistical powers drawn from 1000 replicated simulations are given in Table 3. In general, the RANDOM-A and RANDOM-B methods have slightly higher powers than the FIXED-A and FIXED-B methods for the five large simulated QTL. The FIXED-B and RANDOM-B methods have substantially higher powers than the FIXED-A and RANDOM-A methods. The MPWGAIM method has lower power for the first four large QTL (QTL-2 to QTL-5) than that of the FIXED-A, FIXED-B, RANDOM-A, and RANDOM-B methods. The MPWGAIM method has an advantage over the other methods for detecting the following three QTL: QTL-1, QTL-6, and QTL-7. Except for the MPWGAIM method, no methods have sufficient power to detect the two small QTL (QTL-1 and QTL-7). Overall, the new methods (*i.e.*, FIXED-B, RANDOM-A, and RANDOM-B) are more powerful than the existing methods (*i.e.*, FIXED-A, MPWGAIM, IM, and CIM) for large QTL.

We also compared the FDR for the seven methods (see the last column of Table 3). Here we define the FDR as the proportion of detected QTL that are not true ($\pm$5.00 cM away from a simulated QTL). Clearly, the FIXED-A, FIXED-B, RANDOM-A, and RANDOM-B methods achieve better control of the FDR than the MPWGAIM method, which, in general, is better than the IM and CIM methods.

Behaviors of the methods: We first demonstrate the difference between the random model and the fixed model in terms of the test statistic expressed as $-\log_{10}(P)$ of scanned markers using a single simulated data set (Figure 1). Figure 1A shows the difference between the RANDOM-A and FIXED-A methods. Clearly, the test statistic of the FIXED-A method is

Table 5 Significant SNPs associated with two traits in the *A. thaliana* population and one trait in the mouse population

Population	Trait	Method	SNP	Chr.	Position (kb)	*P*-value[a]	Variance[b]	Candidate gene (kb)
A. thaliana	Bolt to flower	RANDOM-A	MN4_142943	4	143	0.011	0.437	*FRIGIDA*
		RANDOM-B	MN4_142943	4	143	0.005	0.517	Chr. *4*: 269–272
			MN5_2707605	5	2,708	0.023	0.598	
		IM	MASC02783	5	2,522	0.041	0.874	
		MPWGAIM	MN4_142943	4	143	—[c]	0.527	
			MN5_1931248	5	1,931	—[c]	0.629	
	Growth rate	FIXED-B	GA1_3232	4	1,243	0.013	0.626	*FRIGIDA*
		RANDOM-B	GA1_8429	4	1,238	0.010	0.299	Chr. *4*: 269–272
		MPWGAIM	GA1_8429	4	1,238	—[c]	0.253	AT4G02990 (*RUG2*)
		IM	FRI_1888	4	270	0.005	0.493	Chr. *4*: 1322–1324
		CIM-10	GA1_7762	4	1,239	0.011	0.853	AT4G02780 (*GA1*)
								Chr. *4*: 1238–1245
Mouse	PMN	FIXED-B	M2.887	2	87,583	0.040	0.471	
		RANDOM-B	M2.887	2	87,583	0.030	0.292	
		MPWGAIM	M2.887	2	87,583	—[c]	0.263	
		IM	M2.887	2	87,583	0.013	0.450	
		CIM-5	M2.824	2	80,663	0.028	0.496	

[a] *P*-value was obtained from a permutation test.
[b] Denotes the variance of the effect of the detected marker combining the founder allele inheritance indicators.
[c] Owing to the high computational cost, no permutation analysis was conducted.

slightly higher than that of the RANDOM-A method. We also noticed that the $-\log_{10}(P)$ statistic for the RANDOM-A method is very close to zero in regions where no QTL was simulated. This demonstrates the shrinkage property of the random method. In either case, the test-statistic profiles show clear peaks at positions where simulated QTL reside, and the heights of the peaks are proportional to the sizes of the simulated QTL. Figure 1B compares the FIXED-A and FIXED-B methods, where the $-\log_{10}(P)$ profile of the FIXED-B method shows higher peaks than the FIXED-A method. This implies that the FIXED-B method may have a higher power than the FIXED-A method. Figure 1C compares the RANDOM-A and RANDOM-B methods. This also implies that releasing the polygenic counterpart of a marker back to the model may help to increase the power of detecting this marker. These types of behaviors are expected to be observed in data analyses of real experiments.

Average test-statistic profiles: We replicated the simulation experiment 1000 times under both the null model (without QTL effects) and the alternative model (with simulated QTL). The average test-statistic profiles $[-\log_{10}(P)]$ over the 1000 replicates and the 95% threshold values are illustrated in Figure 2. Comparing the fixed models (Figure 2, A and B) with the random models (Figure 2, C and D), we found that the test statistics are slightly higher for the fixed models than for the random models, but the former are also associated with higher threshold values in the test statistics. Comparing -A models (Figure 2, A and C) with -B models (Figure 2, B and D), the latter have higher peaks at positions where simulated QTL reside. For the four models, peaks corresponding to the five large QTL are higher than the threshold values, but peaks corresponding to the two small QTL are below the thresholds. The peaks for the second QTL barely touch the thresholds for -A models (Figure 2, A and C), indicating that the modified models (releasing the polygenic effect back to the model) help to boost the power. None of the peaks in IM reaches the threshold value (Figure 2E). CIM with 30 cofactors only detected four of the five large QTL (Figure 2F). When we increased the number of cofactors to 50 and 65, the CIM method behaved very badly (Figure S1). For the MPWGAIM method, owing to the lack of the test statistics in the package, we only reported the power and FDR.

Average estimated founder effects: We also estimated the founder effects for the seven simulated QTL based on all simulations, and they are illustrated in Figure 3 for the fixed and random models and in Figure 4 for the IM and CIM procedures. The true effects also were plotted along with the estimated effects. All methods provided good estimates of the founder effects. The random models tend to shrink the estimated effects toward zero when the simulated QTL sizes are small (Figure 3, A and G). Although the IM and CIM methods are not as good as the other methods in terms of statistical power, both gave very good estimated founder effects. Figure S2 shows the average estimated effects of the founders when 50 and 65 markers were used as cofactors for the CIM method.

Results of experimental data analyses: *MAGIC population in A. thaliana:* Under the polygenic model, we estimated the variance and heritability for each of the two traits, the days between bolting and flowering (DBF), and the growth rate (GR). The results are shown in Table 4. The heritability of the two traits is 0.342 and 0.473, respectively. The variance ratios for DBF and GR are $\widehat{\lambda}_{\mathrm{DBF}} = \widehat{\phi}^2/\widehat{\sigma}^2 = 0.5203$ and $\widehat{\lambda}_{\mathrm{GR}} = \widehat{\phi}^2/\widehat{\sigma}^2 = 0.8980$, respectively, which were used as known values and incorporated into the covariance structures

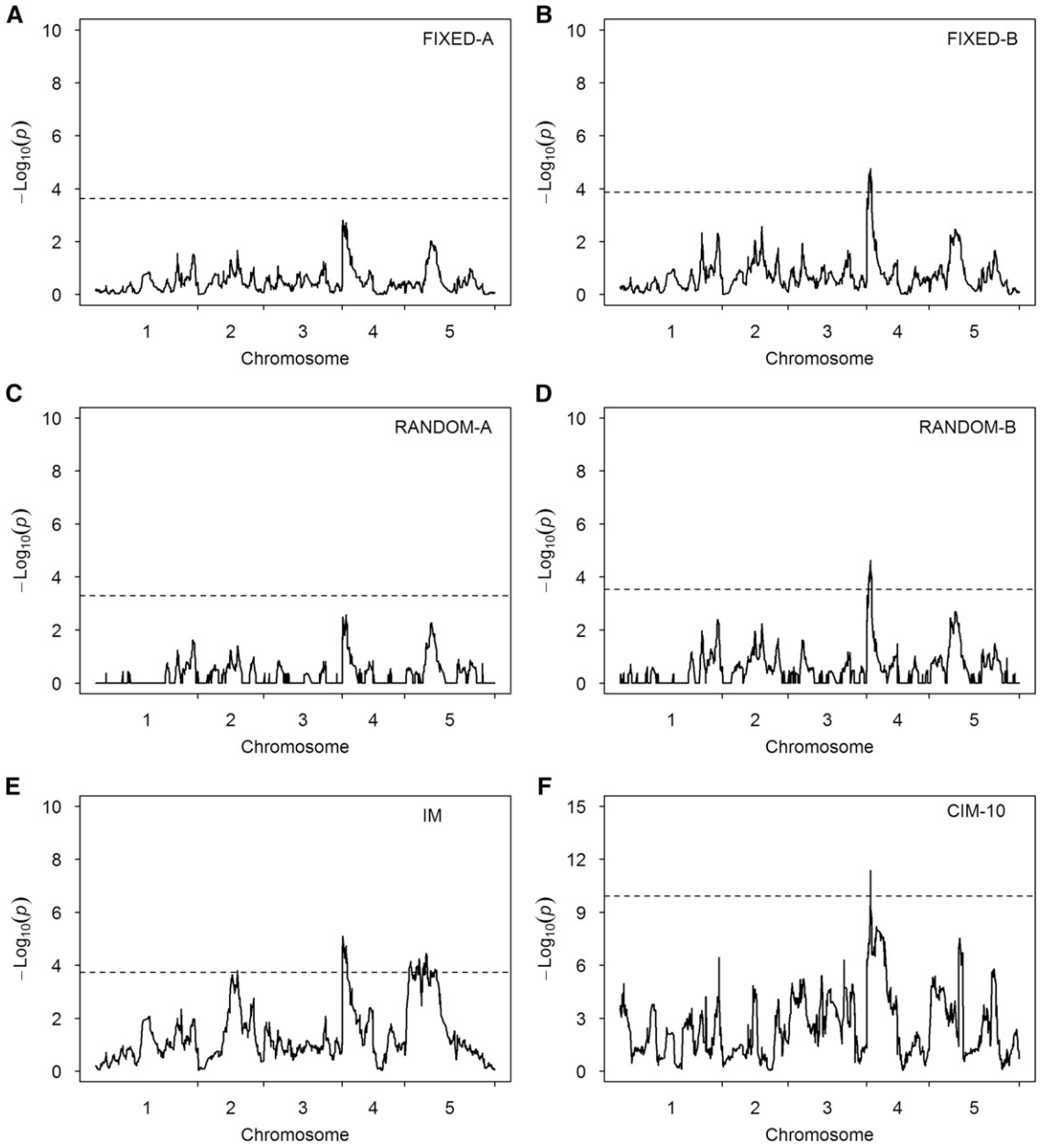

Figure 6 Test-statistic profiles [$-\log_{10}(P)$] for GR of the *A. thaliana* MAGIC population obtained from six methods. The horizontal dotted lines represent the 95% thresholds generated from 1000 permuted samples. (A) Result of the FIXED-A method. (B) Result of the FIXED-B method. (C) Result of the RANDOM-A method. (D) Result of the RANDOM-B method. (E) Result of the IM method. (F) Result of the CIM-10 method.

for genomic scanning of all markers. Figure 5 illustrates the test-statistic profiles [$-\log_{10}(P)$] along with the 95% thresholds generated from 1000 permuted samples for DBF. Markers with test-statistic values greater than the thresholds were claimed to be statistically significant. There are two peaks standing out on chromosomes *4* and *5*, respectively, for all methods except CIM-10. These two regions also showed up in the original analysis of Kover *et al.* (2009). However, the only method that detected both peaks is RANDOM-B, implying that this method may be the most powerful method. The detected QTL on chromosome *4* is located near a known gene called *FRIGIDA*. No related genes were found near the detected QTL on chromosome *5*. The MPWGAIM method also detected the two QTL in the same regions (Table 5). In addition, the MPWGAIM method detected three more QTL, one on chromosome *1* (PERL0236029) and two on chromosome *3* (MASC00175 and MN3_22843506). Figure S3 (A and B) shows the results of this data analysis using the CIM method when 15 and 20 markers were used as cofactors.

The test-statistic profiles along with permutation-generated thresholds are illustrated in Figure 6 for GR. There are many bumps in the test-statistic profiles below the thresholds, indicating that this trait is mostly polygenic. One peak in the beginning of chromosome *4* appears to be common to

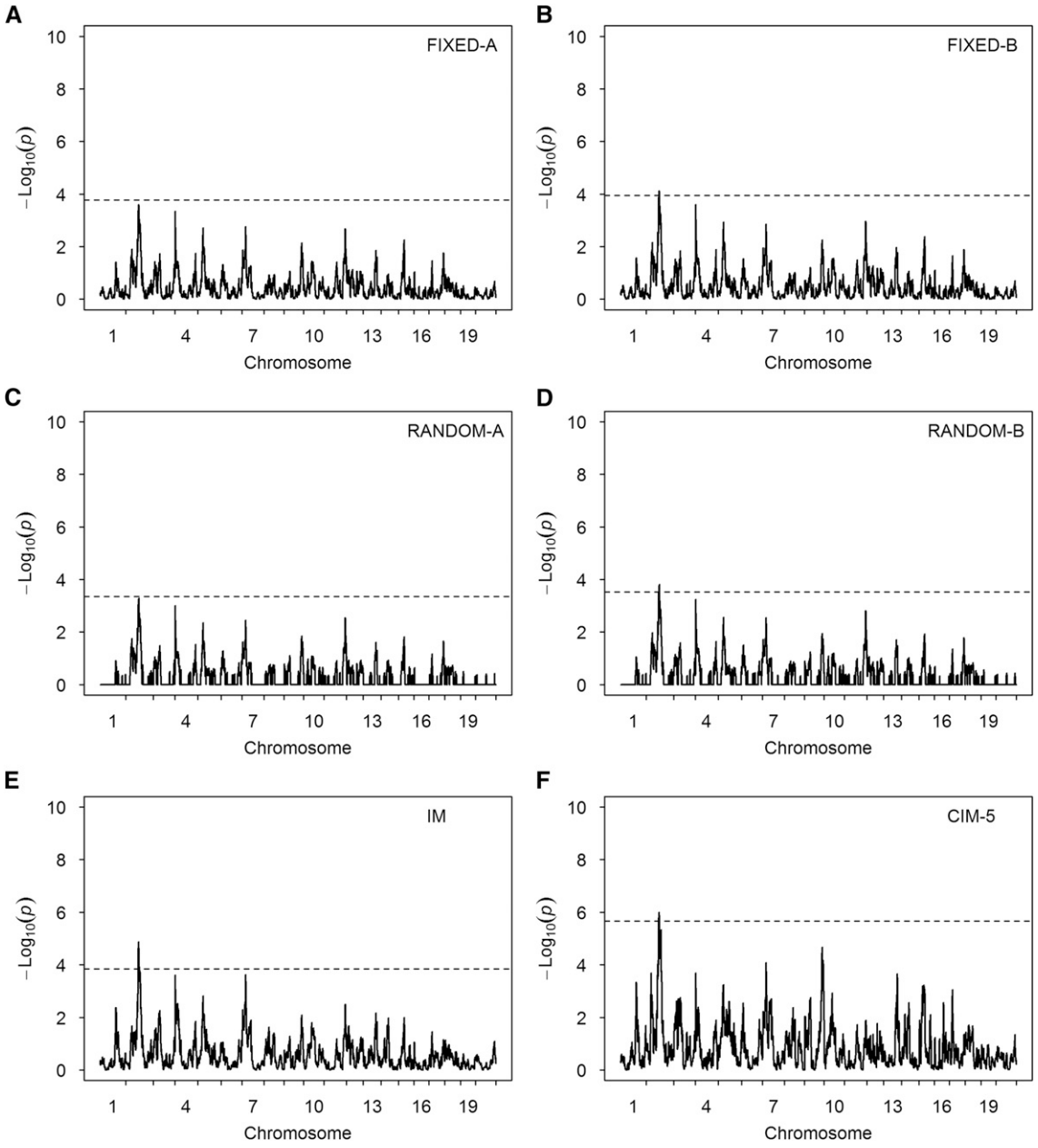

Figure 7 Test-statistic profiles [$-\log_{10}(P)$] for PMN of the pre-CC mouse population obtained from six methods. The horizontal dotted lines represent the 95% thresholds generated from 1000 permuted samples. (A) Result of the FIXED-A method. (B) Result of the FIXED-B method. (C) Result of the RANDOM-A method. (D) Result of the RANDOM-B method. (E) Result of the IM method. (F) Result of the CIM-5 method.

all six methods. Except the FIXED-A and RANDOM-A methods, all other methods have detected the peak as statistically significant. We list the SNPs exceeding the threshold values in Table 5. Three candidate genes are found in this area (about ±200 kb around the detected marker), *FRIGIDA*, AT4G02990 (*RUG2*), and AT4G02780 (*GA1*). The first candidate gene (*FRIGIDA*) is known to affect flowering time. This gene is also related to growth rate in the original study (Kover *et al.* 2009), where it was pointed out that this gene not only plays an important role in plan reproduction but also is a major determinant of the plant developmental process. The second candidate gene (*RUG2*) is important for leaf development in *A. thaliana*, and its loss of function leads to a pleiotropic phenotype, including leaf variegation, reduced growth, and perturbed mitochondrial and chloroplastic gene expression and development (Quesada *et al.* 2011). The third candidate gene (*GA1*) codes for the enzyme *ent*-kaurene synthase A. In *GA1* mutants, the gibberellin biosynthesis pathway is inactivated. As a result, these mutants are deficient in bioactive Gas (Sun and Kamiya 1994). Some additional markers are detected by the IM and MPWGAIM methods, and they are listed in Table S1. The markers on chromosomes *2* and *5* detected by the IM method overlap with the additional markers detected by the MPWGAIM method. The other

Table 6 Computational performances of different methods with different sample sizes and different numbers of markers

Method	Mouse-458-2250[a]	Mouse-458-6683[b]	Mouse-151-27309	*Arabidopsis*-426-1254
FIXED-A	22 sec	53 sec	1 min 43 sec	23 sec
FIXED-B	41 sec	1 min 44 sec	4 min 25 sec	34 sec
RANDOM-A	36 sec	1 min 42 sec	5 min 21 sec	27 sec
RANDOM-B	55 sec	2 min 31 sec	8 min 10 sec	39 sec
MPWGAIM	32 min 29 sec	2 h 33 min	1h 33 min	11 min 51 sec

[a] The first number after the species name is the sample size, and the second number is the number of markers.
[b] The number of bins is 6683, which is the total number of bins of the entire 19 chromosomes of the mouse genome.

methods also show some bumps in regions near the additional markers detected by the IM and MPWGAIM methods. These regions (about ± 200 kb around the peaked markers) harbored several candidate genes, which are not related to GR in terms of gene function. Figure S3 (C and D) shows the results when 15 and 20 markers were used as cofactors for the CIM method.

Pre-CC population of mice: We analyzed a trait named PMN from this population. The phenotypic values were log transformed prior to the analysis, as done in the original study. We estimated the genetic variance and heritability of the trait, which are presented in Table 3. The trait is highly heritable, with a heritability of 0.55. The variance ratio is $\widehat{\lambda} = \widehat{\phi}^2/\widehat{\sigma}^2 = 1.373/1.127 = 1.2183$, which was used along with the kinship matrix to control the polygenic effect in QTL mapping. We scanned the entire genome using all seven methods. The test-statistic profiles are illustrated in Figure 7. Except for the FIXED-A and RANDOM-A methods, all other methods detected a marker on chromosome *2* (Table 5). This marker also was detected by Rutledge *et al.* (2014) in the original study. They found a candidate gene (*Dpn1*) near this marker. No other candidate genes were found in the neighborhood of this marker. Figure S3 (E and F) shows the results when 10 and 20 markers were used as cofactors for the CIM method.

Discussion

A key difference between QTL mapping in MAGIC and biparental populations is the difference in the number of effects to be estimated and tested per locus. Under the fixed-model framework, for an eight-parent MAGIC population, the number of effects per locus is $8 - 1 = 7$, while it is always $2 - 1 = 1$ for a biparental population. Under the null model, the likelihood-ratio text follows a chi-square distribution with 7 degrees of freedom. In a p-parent MAGIC population, $p - 1$ is the degrees of freedom. When p is large, this test is not convenient and sometimes can be unstable (Gatti *et al.* 2014). For example, if some founder alleles fail to appear in the progeny for some loci, the **Z** matrices for these loci will not have the same rank as those loci with full representation of all founders. This variable-rank situation will cause some difficulty in programming. More important, the degree of freedom will vary across loci, so the likelihood-ratio text statistic will not be comparable across loci. We developed a random-model approach to estimate and test the variance among all founder effects per locus. As a result, we only need to estimate and test a single parameter (the variance) regardless how large the number of founders is in a MAGIC population. Simulation studies showed that the random-model approach is slightly more powerful than the fixed-model approach.

Some investigators also considered founder allelic effects as random in MAGIC population QTL mapping (Verbyla *et al.* 2014; Zhang *et al.* 2014). The MPWGAIM procedure of Verbyla *et al.* (2014) assumes that all founder allelic effects of the same locus share a common variance and that this variance varies across loci. A forward variable selection approach was adopted by adding one locus at a time to the model until no further improvement was achieved. For consistency of comparison, we adopted the critical value generated from the null model, similar to the other methods, to evaluate the power of QTL detection by the MPWGAIM method using the same test criterion. We demonstrated lower powers (for large QTL) and higher FDR for the MPWGAIM method. The MPWGAIM method can be time consuming if the numbers of markers and QTL included in the model are large. Table 6 compares the computational times of our methods with that of the MPWGAIM method under several different scenarios. Clearly, the new genome-scanning approaches proposed in this study are substantially faster than the MPWGAIM method. The Bayesian method of Zhang *et al.* (2014) also treats founder allelic effects as random, and it is a multiple-QTL model. Because the method is implemented via an MCMC sampling scheme, it is also computationally expensive. The authors suggested that the method is better used to fine-tune the results after an initial genome scan of all markers.

When cofactors are replaced by the polygene for background control, there is a potential competition between a currently scanned marker and its counterpart in the polygene, which is detrimental to the power. The competition can be very serious when the number of markers used to calculate the kinship matrix is small, although it may be negligible when a very large number of markers are used to calculate the kinship matrix. To prevent such a competition, we proposed releasing the polygenic component corresponding to the scanned marker back to the model. This has dramatically increased the statistical power of QTL detection. The BLUP estimate of a marker effect in the polygene is calculated only once prior to the marker scanning step, and thus little additional computational cost is present. We could have removed the currently

scanned marker from the kinship matrix to avoid the competition. However, this would substantially increase the computational burden because a new kinship matrix would have to be provided for each marker scanned. Special algorithms, such as the spectrally transformed linear mixed model (FaST-LMM) proposed by Lippert *et al.* (2011), may be used to ease the computational intensity. However, the fast speed is not achieved without a cost. One has to use markers with a number substantially smaller than the sample size to gain the fast speed. When the number of markers used to construct the kinship matrix is too small, optimal control of the polygene may not be guaranteed (Zhou and Stephens 2012).

The genotype coding system of QTL mapping in MAGIC populations is different from that in biparental populations. We used the $\mathbf{Z}_k$ variable (an $n \times 8$ matrix) to indicate the founder allelic inheritances for the kth marker. This variable also was used to calculate the marker-inferred kinship matrix **K**. The kinship matrix was eventually rescaled by a normalization factor, which is the average of the diagonal elements of the original unnormalized kinship matrix. After normalization, the diagonal elements of the kinship matrix are all around unity. Such normalization will bring the estimated polygenic variance into the same scale as the residual error variance. Our normalization factor is different from that proposed by VanRaden (2008), which is the sum of heterozygosity across all loci. The normalization factor only changes the scale of the estimated polygenic variance; it affects neither the hypothesis tests nor the results of QTL mapping. In GWAS, where the $\mathbf{Z}_k$ variable is simply a vector, Kang *et al.* (2008) placed a weight variable for each marker in calculating the kinship matrix to take into account variable information contents (allele frequencies) across different marker loci. It is not obvious how to evaluate information contents when the genotype indicator variable $\mathbf{Z}_k$ for each marker is a matrix. In CC and pre-CC mice, all founders contributed equally to the mapping population, and thus, the weight variable can be safely ignored (*e.g.*, taking the default value of 1 from all markers). In the 19-parent MAGIC population of *Arabidopsis*, where the parental contribution varies across founders, a weighted kinship matrix may be more appropriate. Further study is needed to develop an appropriate weight matrix. Alternatively, the method of Gatti *et al.* (2014) for calculating the kinship matrix may be adopted here. The relationship between each pair of individuals is a kind of average "scaled similarity" over all loci. In our notation, the relationship between individuals i and j (the ith row and the jth column of the kinship matrix) is expressed as

$$\mathbf{K}_{ij} = \frac{1}{m}\sum_{k=1}^{m}\frac{\mathbf{Z}_{ik}^{T}\mathbf{Z}_{jk}}{\sqrt{\mathbf{Z}_{ik}^{T}\mathbf{Z}_{ik}}\sqrt{\mathbf{Z}_{jk}^{T}\mathbf{Z}_{jk}}} \tag{15}$$

We did not use this kinship matrix because the polygenic counterpart of marker k (used in the FIXED-B and RANDOM-B methods) would be difficult to interpret when this **K** matrix is used. Furthermore, whether or not such a kinship matrix can adjust unbalanced contributions from different founders is still questionable.

The random-model approach is a kind of Bayesian analysis if the founder effects are considered as parameters and the variance of the founder effects is considered as a prior variance. Because the prior variance is estimated from the data, it is called *empirical Bayes* (Xu 2007).The random model developed for QTL mapping in MAGIC populations can be used in a number of other situations. The method can be extended to QTL mapping in DO populations, such as the DO population of mice developed from the same eight parents as the CC mice (Gatti *et al.* 2014).

The random-model approach is computationally more intensive than the fixed-model approach, where the QTL effects are treated as fixed effects because it requires estimation of a variance component for each marker scanned. We adopted the eigen-decomposition algorithms for the polygenic (null) model and combined them with the Woodbury matrix identity for estimation of QTL variance. It would not be realistic to perform such a random-model QTL mapping without resort to these special algorithms. There may be room for further improvement in the computational speed. However, we emphasize the concept and the novelty of the method, which are far more important than technical improvement in computational speed. Finally, all analyses were performed using an R program written by the authors. We developed an R package named *MagicQTL*, which is provided on the journal website.

Acknowledgments

We thank Arunas P. Verbyla for sharing the mpwgaim program and for the tremendous help in running this program. We also thank Samir N. P. Kelada for providing the pre-CC mouse data. This project was supported by National Science Foundation grant 005400 to S.X. and a China Scholarship Council Award to J.W.

Literature Cited

Bandillo, N., C. Raghavan, P. A. Muyco, M. A. L. Sevilla, I. T. Lobina *et al.*, 2013 Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6: 1–15.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890.

Cheng, R., and A. A. Palmer, 2013 A simulation study of permutation, bootstrap, and gene dropping for assessing statistical significance in the case of unequal relatedness. Genetics 193: 1015–1018.

Chernoff, H., 1954 On the distribution of the likelihood ratio. Ann. Math. Statist. 25: 573–578.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36: 1133–1137.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401.

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar *et al.*, 2014 Quantitative trait locus mapping methods for Diversity Outbred mice. G3 4: 1623–1633.

Gaur, P. M., A. K. Jukanti, and R. K. Varshney, 2012 Impact of genomic technologies on chickpea breeding strategies. Agronomy 2: 199–221.

Golub, G. H., and C. F. Van Loan, 1996 *Matrix Computations*, Ed. 3. Johns Hopkins University Press, Baltimore.

Huang, B. E., and A. W. George, 2011 R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. Bioinformatics 27: 727–729.

Huang, B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol. J. 10: 826–839.

Huang, B. E., K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh *et al.*, 2015 MAGIC populations in crops: current status and future prospects. Theor. Appl. Genet. 128: 999–1017.

Jourjon, M.-F., S. Jasson, J. Marcel, B. Ngom, and B. Mangin, 2005 MCQTL: multi-allelic QTL mapping in multi-cross design. Bioinformatics 21: 128–130.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

King, E. G., S. J. MacDonald, and A. D. Long, 2012a Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. Genetics 191: 935–949.

King, E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen *et al.*, 2012b Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. Genome Res. 22: 1558–1566.

Kover, P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich *et al.*, 2009 A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet. 5: e1000551.

Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson *et al.*, 2011 FaST linear mixed models for genome-wide association studies. Nat. Methods 8: 833–835.

MacDonald, S. J., and A. D. Long, 2007 Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of Drosophila melanogaster. Genetics 176: 1261–1281.

Mackay, I. J., P. Bansept-Basler, T. Barber, A. R. Bentley, J. Cockram *et al.*, 2014 An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3 4: 1603–1610.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA 97: 12649–12654.

Pascual, L., N. Desplat, B. E. Huang, A. Desgroux, L. Bruguier *et al.*, 2015 Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant Biotechnol. J. 13: 565–577.

Quesada, V., R. Sarmiento-Mañús, R. González-Bayón, A. Hricová, R. Pérez-Marcos *et al.*, 2011 Arabidopsis RUGOSA2 encodes an mTERF family member required for mitochondrion, chloroplast and leaf development. Plant J. 68: 738–753.

Rakshit, S., A. Rakshit, and J. Patil, 2012 Multiparent intercross populations in analysis of quantitative traits. J. Genet. 91: 111–117.

Rutledge, H., D. L. Aylor, D. E. Carpenter, B. C. Peck, P. Chines *et al.*, 2014 Genetic regulation of Zfp30, CXCL1, and neutrophilic inflammation in murine lung. Genetics 198: 735–745.

Sannemann, W., B. E. Huang, B. Mathew, and J. Léon, 2015 Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. Mol. Breed. 35: 1–16.

Sun, T. P., and Y. Kamiya, 1994 The Arabidopsis GA1 locus encodes the cyclase ent-kaurene synthetase A of gibberellin biosynthesis. Plant Cell 6: 1509–1518.

Threadgill, D. W., K. W. Hunter, and R. W. Williams, 2002 Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. Mamm. Genome 13: 175–178.

Valdar, W., J. Flint, and R. Mott, 2006 Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics 172: 1783–1797.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Varshney, R. K., and A. Dubey, 2009 Novel genomic tools and modern genetic and breeding approaches for crop improvement. J. Plant Biochem. Biotechnol. 18: 127–138.

Verbyla, A. P., A. W. George, C. R. Cavanagh, and K. L. Verbyla, 2014 Whole-genome QTL analysis for MAGIC. Theor. Appl. Genet. 127: 1753–1770.

Visscher, P. M., 2006 A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res. Hum. Genet. 9: 490–495.

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

Xu, S., 2013a Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195: 1103–1115.

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

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

Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539–551.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.

Zhang, Z., W. Wang, and W. Valdar, 2014 Bayesian modeling of haplotype effects in multiparent populations. Genetics 198: 139–156.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44: 821–824.

Communicating editor: I. Hoeschele

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179945/-/DC1

A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations

Julong Wei and Shizhong Xu

Copyright © 2016 by the Genetics Society of America
DOI: 10.1534/genetics.115.179945

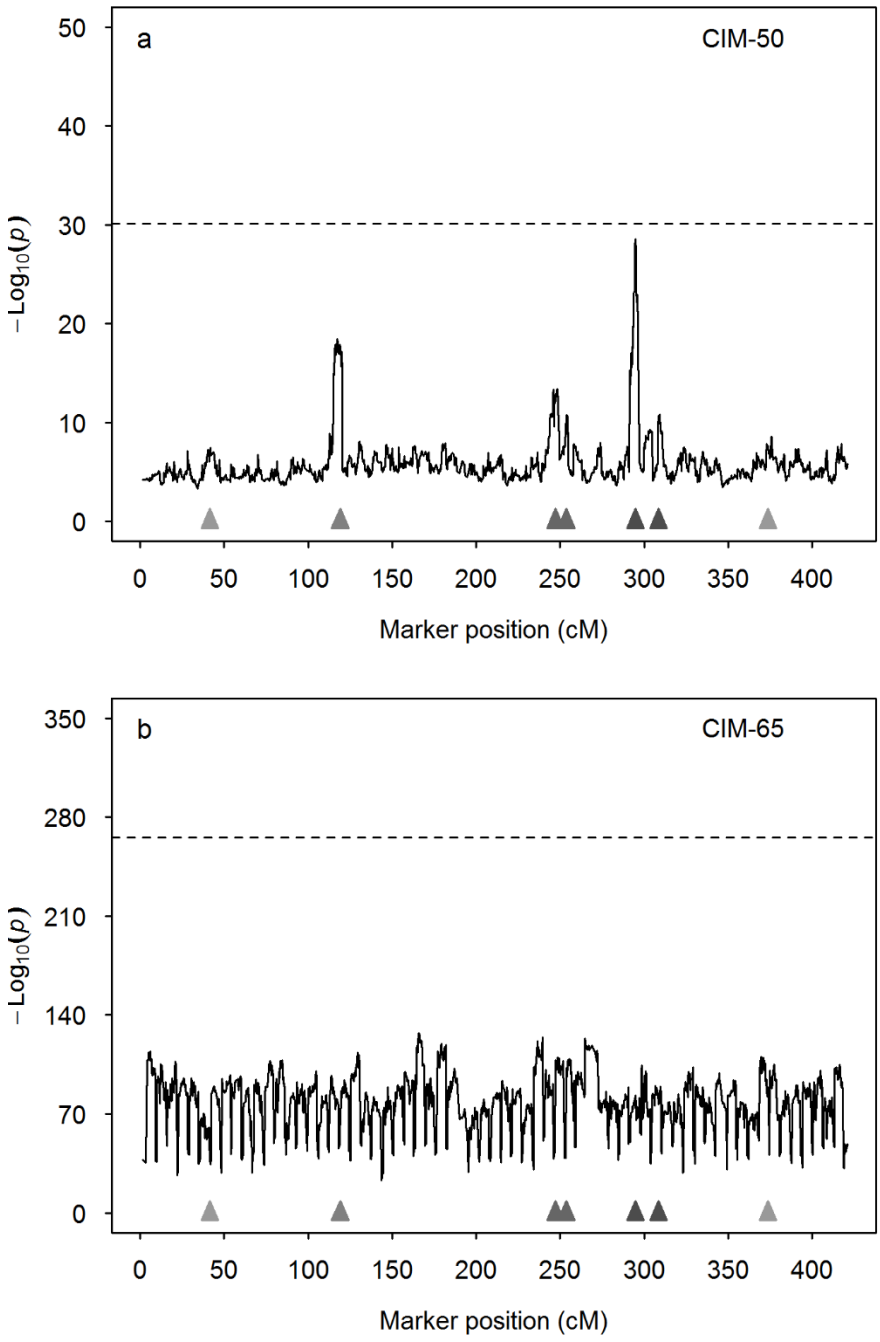

Figure S1. Average test statistic profiles ($-\log_{10}(p)$) of the CIM method using different numbers of co-factors (CIM-50 and CIM-65) from 1000 replicated simulation experiments. The horizontal dashed lines represent the 95% thresholds drawn from 1000 simulated samples under the null model. The true locations of the seven simulated QTL are represented by the filled triangles on the *x*-axis.

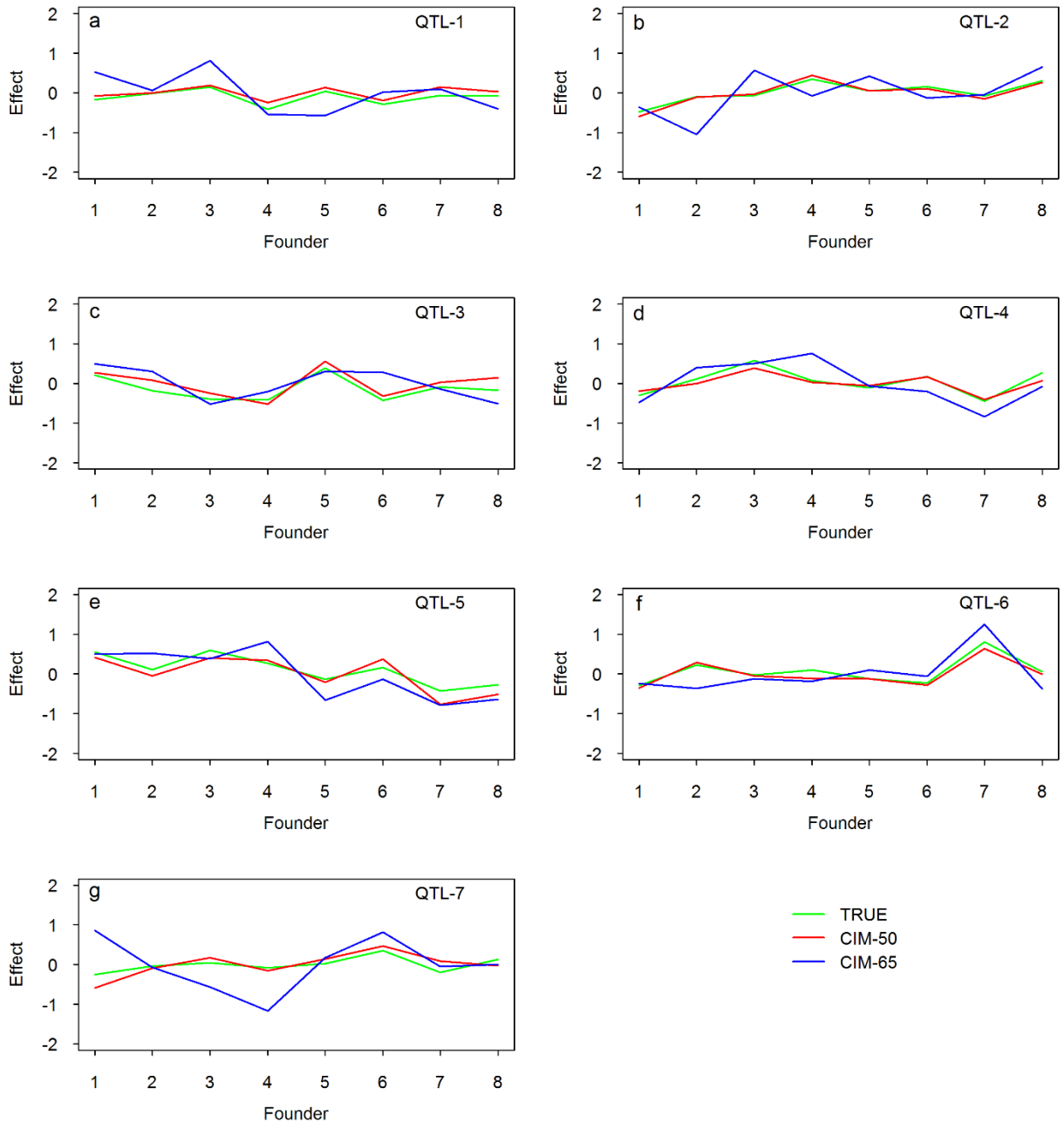

Figure S2. True and estimated allelic effects of eight founders for seven simulated QTL in the simulation experiment. The estimated effects are the average effects of 1000 replicated experiments. Results from two methods are presented in this figure: CIM-50 and CIM-65, where the numbers after CIM represent the numbers of co-factors.

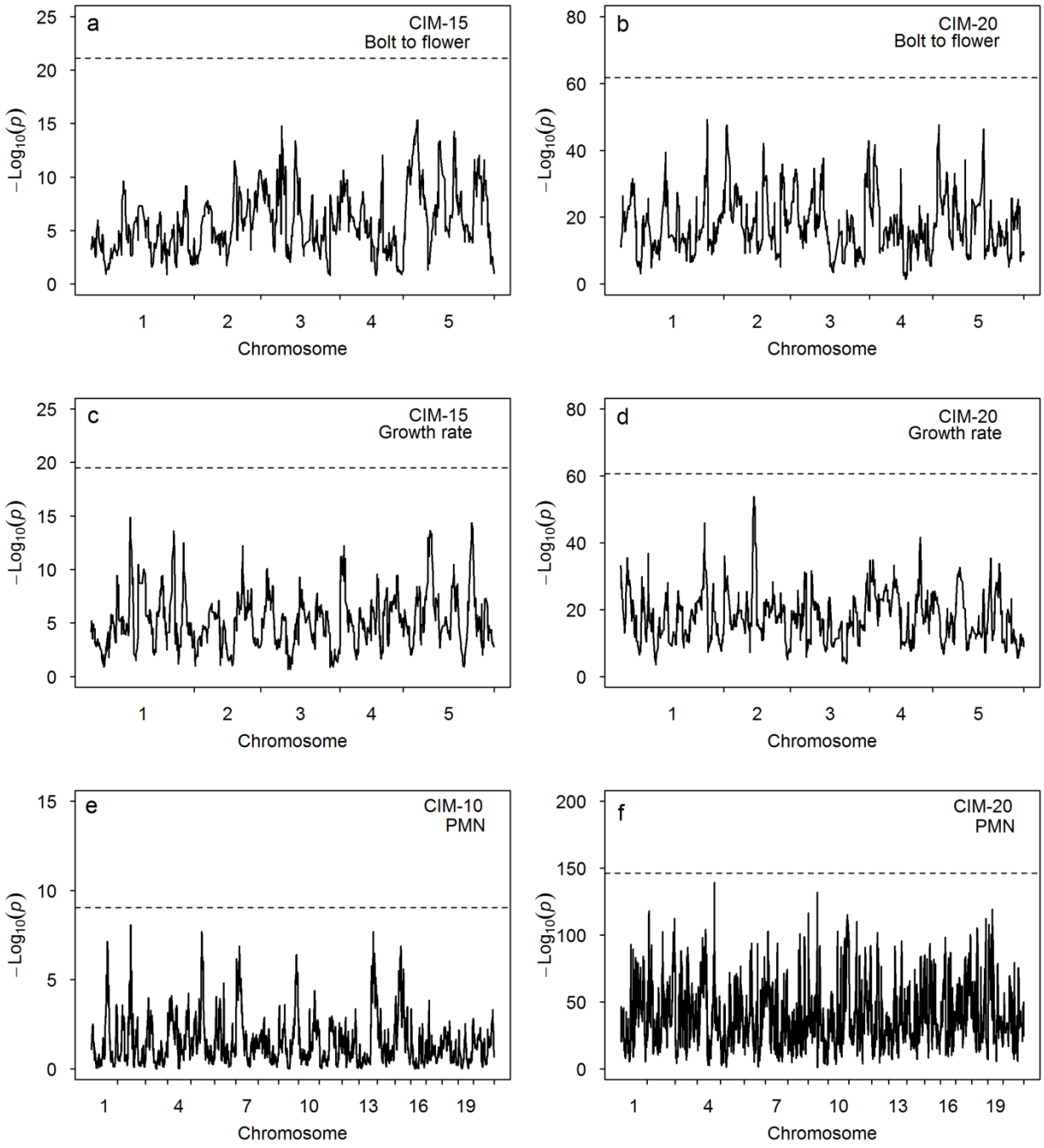

Figure S3. Test statistic profiles ($-\log_{10}(p)$) for three traits in two populations using the CIM methods with alternative numbers of co-factors. The horizontal dotted lines represent the 95% thresholds generated from 1000 permuted samples.

Table S1. More SNPs related to growth rate detected by the IM and MPWGAIM methods in *Arabidopsis thaliana*.

Method	SNP	Chr	Position (kb)	p-value[a]	Variance[b]
IM	ATC_828	2	11,773	0.043	0.422
	ATMYB33_119	5	1,837	0.024	0.442
	MN5_4344025	5	4,344	0.023	0.447
	NMSNP5_652310	5	6,523	0.01	0.459
MPWGAIM	SGCSNP10779	1	28,831	—[c]	0.072
	MASC05360	2	5,179	—	0.057
	MASC02928	2	9,753	—	0.091
	HOS1_1176	2	16,614	—	0.111
	MN3_4470311	3	4,470	—	0.079
	PHYD_2806	4	9,197	—	0.066
	MN5_1399959	5	1,400	—	0.048
	MASC07384	5	8,001	—	0.156
	VIN3_300	5	23,249	—	0.033

[a]-p-value obtained from 1000 permutation analysis.
[b] variance of effects of the detected marker combining the founder allele inheritance indicators.
[c] "—" due to extensive computing time for the MPWGAIM method, no permutation was conducted.

File S1: Bin data of the Collaborative Cross (CC) mouse population of 458 individuals. (.RData, 494 KB)

Available for download as a .RData file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179945/-/DC1/FileS1.RData

Supporting Data and R program

File S1 Bin data of the Collaborative Cross (CC) mouse population of 458 individuals.

File S2 Supplementary notes: derivation of various formulas.

File S3 MagicQTL_1.0.tar.gz the R package (MagicQTL).

File S4 Documents for the MagicQTL R package.

File S2: Derivation of various formulas

Restricted maximum likelihood estimation of variance component via eigen-decomposition:

Under the polygenic model, the restricted log likelihood function is,

$$L(\theta) = -\frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2}\ln|H| - \frac{1}{2\sigma^2}(y - X\beta)^T H^{-1}(y - X\beta) - \frac{1}{2}\ln|X^T H^{-1} X| \quad (1)$$

where $\theta = \{\beta, \lambda, \sigma^2\}$ is the parameter vector, β is a vector of fixed effects, $\lambda = \phi^2 / \sigma^2$ the variance ratio, ϕ^2 is the polygenic variance, σ^2 is the residual variance, *n* is the sample size, *r* is the rank of matrix *X*, $H = K\lambda + I$ is the covariance structure and *K* is a marker inferred kinship matrix. Given λ, the maximum likelihood estimates of β and σ^2 are

$$\begin{aligned} \hat{\beta} &= (X^T H^{-1} X)^{-1} X^T H^{-1} y \\ \hat{\sigma}^2 &= \frac{1}{n-r}(y - X\hat{\beta})^T H^{-1}(y - X\hat{\beta}) \end{aligned} \quad (2)$$

These two estimated parameters are expressed as functions of λ. Substituting β and σ^2 in equation (1) by $\hat{\beta}$ and $\hat{\sigma}^2$ in equation (2) yields a profiled likelihood function that is only a function of λ, as shown below,

$$L(\lambda) = -\frac{1}{2}\ln|H| - \frac{1}{2}\ln|X^T H^{-1} X| - \frac{n-r}{2}\ln(y^T P y) \quad (3)$$

where

$$P = H^{-1} - H^{-1} X (X^T H^{-1} X)^{-1} X^T H^{-1} \quad (4)$$

A numeric solution of λ can be found iteratively using the Newton algorithm,

$$\lambda^{(t+1)} = \lambda^{(t)} - \left[\frac{\partial^2 L(\lambda^{(t)})}{\partial \lambda^2}\right]^{-1} \left[\frac{\partial L(\lambda^{(t)})}{\partial \lambda}\right] \quad (5)$$

The likelihood function requires inverse and determinant of matrix *H*, an $n \times n$ matrix, and the computation can be demanding for large sample sizes. We used the eigen-decomposition approach to deal with the *K* matrix (KANG *et al.* 2008; ZHOU and STEPHENS 2012). Further investigation of equation (3) shows that the profiled restricted log likelihood function only requires the log determinant of matrix *H* and various quadratic forms involving H^{-1}. Let us perform eigen-decomposition for *K* so that $K = UDU^T$, where $D = \text{diag}\{\delta_1, ..., \delta_n\}$ is a diagonal

matrix for the eigenvalues and U is the eigenvectors, an $n \times n$ matrix. The eigenvectors have the property of $U^T = U^{-1}$ so that $UU^T = I$. Now, let us rewrite matrix *H* by

$$H = K\lambda + I = UDU^T\lambda + I = U(D\lambda + I)U^T \tag{6}$$

The determinant of *H* is

$$|H| = |U(D\lambda + I)U^T| = |D\lambda + I||UU^T| = |D\lambda + I| \tag{7}$$

where $D\lambda + I$ is a diagonal matrix. Therefore, the log determinant of matrix *H* is

$$\ln|H| = \sum_{j=1}^{n} \ln(\delta_j\lambda + 1) \tag{8}$$

The restricted log likelihood function also involves various quadratic terms in the form of $a^T H^{-1} b$, for example, $X^T H^{-1} X$, $X^T H^{-1} y$ and $y^T H^{-1} y$. Using eigenvalue decomposition, we can rewrite the quadratic form by

$$a^T H^{-1} b = a^T U(D\lambda + I)^{-1} U^T b = a^{*T}(D\lambda + I)^{-1} b^* = \sum_{j=1}^{n} a_j^{*T} b_j^* (\delta_j\lambda + 1)^{-1} \tag{9}$$

where $a^* = U^T a$ and $b^* = U^T b$. Note that a_j^* is the *j*th element (row) of vector (matrix) a^* and b_j^* is the *j*th element (row) of vector (matrix) b^*. Using eigenvalue decomposition, matrix inversion and determinant calculation have been simplified into simple summations, and thus, the computational speed can be substantially improved.

Best linear unbiased prediction (BLUP) of a marker effect under the polygenic model: Under the polygenic model, all marker effects share the same variance, i.e., $a_k \sim N(0, I\phi^2 / m))$ for $k = 1, ..., m$, where $\phi^2 = \lambda\sigma^2$ is estimated from the data under the polygenic model. The BLUP estimate of a_k is

$$\hat{a}_k = \mathrm{E}(a_k \mid y) = Z_k^T(\hat{\phi}^2 / m)(K\hat{\phi}^2 + I\hat{\sigma}^2)^{-1}(y - X\hat{\beta}) \tag{10}$$

We have a total of *m* markers and thus *m* effects to estimate under the polygenic model (prior to the marker scanning step). The polygenic effect associated with marker *k* is $\hat{\xi}_k = Z_k\hat{a}_k$. Here, eigen-decomposition is also required to avoid direct calculation of $(K\hat{\phi}^2 + I\hat{\sigma}^2)^{-1}$.

Estimating variance components via Woodbury matrix identity and eigen-decomposition: The genomic scanning model for the *k*th locus is

$$y = X\beta + Z_k\gamma_k + \xi + \varepsilon \tag{11}$$

where ξ is the polygene and the general error term $\xi + \varepsilon$ has $\mathrm{E}(\xi + \varepsilon) = 0$ and $\mathrm{var}(\xi + \varepsilon) = (K\hat{\lambda} + I)\sigma^2$. We assume $\gamma_k \sim N(0, I_8\phi_k^2)$ and perform a significance test for $H_0 : \phi_k^2 = 0$. Under the null hypothesis, the *k*th locus is not linked to QTL. The expectation of *y* remains $\mathrm{E}(y) = X\beta$, but the variance-covariance matrix is

$$\mathrm{var}(y) = Z_k Z_k^T \phi_k^2 + K\phi^2 + I\sigma^2 = (Z_k Z_k^T \lambda_k + K\hat{\lambda} + I)\sigma^2 \tag{12}$$

where $\lambda_k = \phi_k^2 / \sigma^2$ is the variance ratio. Let $y^* = U^T y$, $X^* = U^T X$ and $Z_k^* = U^T Z_k$ be transformed variables so that

$$y^* = X^*\beta + Z_k^*\gamma_k + U^T(\xi + \varepsilon) \tag{13}$$

The variance-covariance matrix of y^* is

$$\begin{aligned}\mathrm{var}(y^*) &= Z_k^* Z_k^{*T} \phi_k^2 + (D\hat{\lambda} + I)\sigma^2 \\ &= (Z_k^* Z_k^{*T} \lambda_k + R)\sigma^2\end{aligned} \tag{14}$$

where $R = D\hat{\lambda} + I$ is a known diagonal matrix for the general covariance structure. Let $H_k = Z_k^* Z_k^{*T} \lambda_k + R$ and define the restricted log likelihood function for parameter vector $\theta = \{\beta, \lambda_k, \sigma^2\}$ by

$$L(\theta) = -\frac{n-r}{2}\ln(\sigma^2) - \frac{1}{2}\ln|H_k| - \frac{1}{2\sigma^2}(y^* - X^*\beta)^T H_k^{-1}(y^* - X^*\beta) - \frac{1}{2}\ln|X^{*T} H_k^{-1} X^*| \tag{15}$$

Given λ_k, the maximum likelihood estimates of β and σ^2 are

$$\begin{aligned}\hat{\beta} &= (X^{*T} H_k^{-1} X^*)^{-1} X^{*T} H_k^{-1} y^* \\ \hat{\sigma}^2 &= \frac{1}{n-r}(y^* - X^*\hat{\beta})^T H_k^{-1}(y^* - X^*\hat{\beta})\end{aligned} \tag{16}$$

The above estimated parameters are expressed as functions of λ_k. Substituting β and σ^2 in equation (15) by $\hat{\beta}$ and $\hat{\sigma}^2$ in equation (16) yields a profiled likelihood function that is only a function of λ_k, as shown below,

$$L(\lambda_k) = -\frac{1}{2}\ln|H_k| - \frac{1}{2}\ln|X^{*T} H_k^{-1} X^*| - \frac{n-r}{2}\ln(y^{*T} P_k y^*) \tag{17}$$

where

$$P_k = H_k^{-1} - H_k^{-1} X^* (X^{*T} H_k^{-1} X^*)^{-1} X^{*T} H_k^{-1} \tag{18}$$

The Newton algorithm for the numeric solution of λ_k is

$$\lambda_k^{(t+1)} = \lambda_k^{(t)} - \left[\frac{\partial^2 L(\lambda_k^{(t)})}{\partial \lambda_k^2}\right]^{-1} \left[\frac{\partial L(\lambda_k^{(t)})}{\partial \lambda_k}\right] \tag{19}$$

Once the iteration process has converged, the solution is the MLE of λ_k, denoted by $\hat{\lambda}_k$.

Efficient matrix inversion and determinant calculation are required to evaluate the log likelihood function shown in equation (17). We used the Woodbury matrix identities to improve the computational speed (GOLUB and VAN LOAN 1996). The Woodbury matrix identities are

$$\begin{aligned}H_k^{-1} &= (Z_k^* Z_k^{*T} \lambda_k + R)^{-1} \\ &= R^{-1} - \lambda_k R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8)^{-1} Z_k^{*T} R^{-1}\end{aligned} \tag{20}$$

and

$$\begin{aligned}|H_k| &= |Z_k^* Z_k^{*T} \lambda_k + R| \\ &= |R|\,|\lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8|\end{aligned} \tag{21}$$

Because $R = D\hat{\lambda} + I$ is a diagonal matrix, the Woodbury identities convert the above calculations into inversion and determinant of matrices with dimension 8×8. The restricted likelihood function also involves various quadratic terms in the form of $a^T H_k^{-1} b$, which can be expressed as

$$a^T H_k^{-1} b = a^T R^{-1} b - \lambda_k a^T R^{-1} Z_k^* (\lambda_k Z_k^{*T} R^{-1} Z_k^* + I_8)^{-1} Z_k^{*T} R^{-1} b \tag{22}$$

Note that the quadratic term involving H_k^{-1} has been expressed as a function of various simplified $a^T R^{-1} b$ terms. The simplified quadratic term is calculated using

$$a^T R^{-1} b = \sum_{j=1}^{n} a_j^T b_j (\delta_j \hat{\lambda} + 1)^{-1} \quad (23)$$

where a_j and b_j are the *j*th rows of matrices a and b, respectively, for $j = 1,...,n$.

LITERATURE CITED

Golub, G. H., and C. F. Van Loan, 1996 Matrix computations. 1996. Johns Hopkins University, Press, Baltimore, MD, USA**:** 374-426.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178**:** 1709-1723.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nature genetics 44**:** 821-824.

File S3: MagicQTL_1.0.tar.gz the R package (MagicQTL). (.gz, 590 KB)

Available for download as a .gz file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179945/-/DC1/FileS3.gz

Documents for MagicQTL R package

MagicQTL is an R package to perform QTL mapping in Multi-parent Advanced Generation Inter-cross (MAGIC) populations under both the fixed model and the random model methodology. The program also include two conventional QTL mapping methods, interval mapping (IM) and composite interval mapping (CIM). Users only need to call one function, **magicScan**. This user instruction has two parts: (1) how to install **MagicQTL** package in your computer; (2) an example to show the workflow using the **MagicQTL** package.

1. Install magicQTL package

In the Unix or Linux platform,

Just type the following command, R CMD INSTALL MagicQTL_1.0.tar.gz

Then complete installing the MagicQTL package!

In the windows platform,

The first step, download the Rtools from R CRAN (https://www.r-project.org/), then install the Rtools. Notes that you should add the "c:\program files\Rtools\bin", "c:\program files\Rtools\gcc-4.6.3\bin", "c:\program files\R\R.3.x.x\bin\i386" and "c:\program files\R\R.3.x.x\bin\x64" into the Path Variable on the Environment Variables panel.

The second step, in the search box, type "command prompt", then click.

In the command prompt, type the following command R CMD INSTALL MagicQTL_1.0.tar.gz.

Then install!

To use this package, Just type library(MagicQTL) and call the function magicScan()

2. Introduction of implementing the MagicQTL

Here we provide a test example to briefly introduce how to implement the MagicQTL package. Details can be obtained via help(magicScan) or ?magicScan.

The original data is *Arabidopsis thaliana* MAGIC population inherited from 19 founders obtained from the website (http://mus.well.ox.ac.uk/magic/). In consideration of file size, the test data is a subset, which is comprised of 65 markers distributed in the five chromosomes, 60 individuals with five traits. We can offer the original data applying to our program format if requested.

Demo code

First step-load data

```
library(MagicQTL)
> data(Ara)
> names(Ara)
[1] "gen"      "map"      "Ara.phe"  "kk.eigen"
> gen<-Ara[[1]]
> map<-Ara[[2]]
> Ara.phe<-Ara[[3]]
> kk.eigen<-Ara[[4]]
> chrnum<-length(gen)
```

Data format Information

```
#gen, probability matrix
> dim(gen[[1]])
[1] 266  60
> class(gen[[1]])
[1] "matrix"
#map, marker information
> dim(map[[1]])
[1] 14  4
> class(map[[1]])
[1] "data.frame"
#Ara.phe, phenotype
> dim(Ara.phe)
[1] 60   6
#kk.eigen,including the kinship matrix, its eigendecomposition and
# the numeric
> names(kk.eigen)
[1] "kk" "qq" "cc"
```

#The probability matrix, like following

```
> gen[[1]][1:19,1:4]
             [,1]         [,2]         [,3]         [,4]
Bur  8.012807e-08 1.715535e-01 1.697059e-06 7.217872e-02
Can  1.718630e+00 3.708686e-08 6.840062e-02 1.108586e-07
Col  3.390267e-02 3.709196e-08 6.841290e-02 2.458504e-07
Ct   4.620957e-07 4.520535e-02 8.073149e-08 1.639100e+00
Edi  2.541389e-03 2.702748e-07 1.753971e-01 3.714008e-08
Hi   2.531023e-03 7.467149e-07 2.011564e-01 3.714148e-08
Kn   8.027694e-08 8.673264e-02 1.382932e-06 7.217847e-02
Ler  2.531022e-03 3.353259e-07 3.319853e-01 3.715066e-08
Mt   3.390267e-02 3.709195e-08 6.841290e-02 2.458504e-07
No   3.390267e-02 3.709196e-08 6.841290e-02 2.458504e-07
Oy   8.012806e-08 9.832535e-02 3.289830e-06 7.218036e-02
Po   8.012806e-08 9.832535e-02 3.289830e-06 7.218036e-02
Rsch 3.390267e-02 3.709195e-08 6.841290e-02 2.458504e-07
Sf   8.012807e-08 1.499856e+00 1.697417e-06 7.217872e-02
Tsu  2.531022e-03 3.353140e-07 6.757747e-01 3.715066e-08
Wil  3.390267e-02 3.709195e-08 6.841290e-02 2.458504e-07
Ws   3.390267e-02 3.709195e-08 6.840473e-02 1.197786e-06
Wu   3.391363e-02 3.712419e-08 6.840265e-02 1.287837e-07
Zu   3.390516e-02 3.716816e-08 6.840265e-02 1.286525e-07
>
```

##The map format, like

```
> map[[1]][1:10,]
           markers chr       cm       bp
1        MN1_29291   1  0.12205    29291
2        MASC07424   1  6.26250  1502999
3      MN1_3229670   1 13.45770  3229846
4      MN1_4947324   1 20.61385  4947328
5          MFT_113   1 25.94783  6227484
6          GI_2186   1 33.59517  8062852
7  NMSNP1_10720273   1 44.66781 10720291
8      SGCSNP10165   1 55.00472 13201153
9      PERL0147872   1 72.81940 17474215
10     PERL0173191   1 85.47685 20510777
>
```

##Ara.phe, the phenotype, including the five traits.

```
> Ara.phe[1:5,]
        ID bolt.to.flower days.to.bolt days.to.germ total.cm growth rate
1 MAGIC.10           10.0         40.0          6.0   37.060  -2.8677181
2 MAGIC.94           10.0         24.6          6.0   41.520   3.1500000
3 MAGIC.95            7.0         20.2          6.0   31.900  -2.8500000
4 MAGIC.96           10.0         36.0          6.0   48.200  -0.2500000
5 MAGIC.97           10.2         25.2          9.4   51.225  -0.8633877
>
```

#Second step-scan the markers

```
> indi<-nrow(Ara.phe)

> x<-rep(1,indi)

> y<-Ara.phe[,5] # Phenotype,total length, that is height of the Arabdopsis

> d<-data.frame(y=y,x=x)

> scans<-
magicScan(dataframe=d,gen=gen,map=map,kk.eigen=kk.eigen,nfounders=19,
model="Random-A")

lambda: 1.125352e-07 Residual error: 70.79299 Model: Random-A

Data of chr have been completed 0

Data of chr have been completed 0

Data of chr have been completed 0

Data of chr have been completed 0

Data of chr have been completed 0
```

#output the result after scanning

```
#Output

> parms<-lapply(1:chrnum, function(i){ return(scans[[i]][[1]]) })

> parms<-do.call(rbind,parms)

> write.csv(parms,file="Ara.parm.csv",row.names=FALSE)

> #

> blupp<-lapply(1:chrnum, function(i){ return(scans[[i]][[2]]) })

> blupp<-do.call(rbind,blupp)

> write.csv(blupp,file="Ara.blupp.csv",row.names=FALSE)
```

#Output information

#parms format, like following

```
> parms[1:5,]
  Num chr      ccM         lrt     lrt.p  lrt.logp      wald    wald.p
1   1   1  0.12205 0.212660545 0.3223450 0.4916790 2.5934390 0.9999910
2   2   1  6.26250 0.734221490 0.1957591 0.7082780 5.5863602 0.9976058
3   3   1 13.45770 0.038718467 0.4220037 0.3746837 1.2147610 1.0000000
4   4   1 20.61385 0.002749201 0.4790919 0.3195812 0.3154737 1.0000000
5   5   1 25.94783 0.685325117 0.2038795 0.6906264 5.3563576 0.9981869
     wald.logp     tau_k   sigma2       lam_k conv
1 3.890128e-06 1.6091917 67.88415 0.023704971    0
2 1.041022e-03 2.0943954 64.58632 0.032427845    0
3 7.806182e-09 0.4470102 69.36368 0.006444441    0
4 6.277772e-14 0.1225779 70.41699 0.001740743    0
5 7.881258e-04 1.7812334 64.92988 0.027433185    0
```

#blupp format, like following

```
> blupp[1:2,]
        Gamma1     Gamma2       Gamma3     Gamma4      Gamma5    Gamma6     Gamma7     Gamma8
blup 0.8591198 -0.4682053 -0.022309143 0.05509254 -0.02871033 0.7041129 -0.8938757  0.1415811
blup 1.0667156 -1.5576318  0.007362878 0.24340898 -0.23955166 1.1250004 -1.1860768 -0.1792806
          Gamma9   Gamma10   Gamma11   Gamma12   Gamma13     Gamma14    Gamma15    Gamma16
blup -0.24231020 0.4035547 0.1130185 0.1132037 0.1534979  0.01137074 -0.4144860 -0.1187667
blup  0.04603282 0.7335408 0.1129925 0.1133913 0.4658266 -0.15829037 -0.2452165 -0.2201461
        Gamma17    Gamma18   Gamma19  stderr1  stderr2  stderr3  stderr4  stderr5  stderr6
blup 0.06213689 -0.6817075 0.2536821 1.213400 1.187375 1.202695 1.194544 1.265652 1.132944
blup 0.09724851 -0.7794576 0.5541311 1.367922 1.166058 1.207533 1.236783 1.397190 1.185177
       stderr7  stderr8  stderr9 stderr10 stderr11 stderr12 stderr13 stderr14 stderr15 stderr16
blup 0.9927385 1.127535 1.237929 1.233045 1.232292 1.232277 1.160970 1.122480 1.243798 1.238761
blup 1.1331874 1.101253 1.270078 1.330291 1.385856 1.385814 1.196191 1.138972 1.442503 1.416296
     stderr17 stderr18 stderr19
blup 1.189109 1.157115 1.211248
blup 1.287618 1.190256 1.305421
>
```