Theor Appl Genet (2015) 128:1163–1174
DOI 10.1007/s00122-015-2497-x

ORIGINAL PAPER

Genome-wide association mapping in plants

Andrew W. George[1,2] · Colin Cavanagh[2]

Received: 4 July 2014 / Accepted: 10 March 2015 / Published online: 24 March 2015

Abstract
Key message **We present new association mapping methods which address the unique challenges of analyzing genome-wide data from multi-environment plant studies.**
Abstract Association studies on a genome-wide scale are being performed in plants. Unlike human studies, plant studies contain replicates whose data may be recorded across different environments. Plant studies also often employ elaborate experimental designs for controlling extraneous phenotypic variation. As a result, the genome-wide analysis of data from plant studies can be challenging. In this paper, we present QK-based association mapping for the analysis of data from plant association studies. In doing so, we have developed: (a) a general multivariate QK framework for association mapping in plant studies of arbitrary complexity; (b) a new weighted two-stage analysis approach for QK-based association mapping; (c) a heuristic procedure for determining when two-stage analysis is appropriate; and (d) a Monte Carlo sampling procedure for controlling the genome-wide type I error rate. We conduct a simulation study to evaluate the performance of our genome-wide mapping technique. We also analyze data from a multi-environment association study in wheat.

Communicated by M. J. Sillanpaa.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2497-x) contains supplementary material, which is available to authorized users.

✉ Andrew W. George
andrew.george@csiro.au

[1] Digital Productivity Flagship, CSIRO, Brisbane, Qld, Australia

[2] Agriculture Flagship, CSIRO, Canberra, ACT, Australia

Introduction

Genome-wide association studies in plants are a promising new development in the race to unlock the genetic secrets of heritable traits. Genome-wide studies have been made possible through new high-throughput genotyping technologies and decreasing genotyping costs. Unlike more traditional studies that map trait loci via linkage, association studies harness linkage disequilibrium for mapping. In doing so, trait loci are positioned with high precision. Association studies are also better suited for the genetic exploration of complex traits because of their broad genetic base. Historically, association studies have been used to refine or confirm known results. Now, it is maturing into a new resource for gene discovery on a genome-wide scale.

Association mapping methods are statistical techniques for measuring the strength of association between a marker locus and trait. Marker-trait association occurs when a quantitative trait locus (QTL), that is influencing a trait, is in linkage disequilibrium with the marker locus. A number of different statistical techniques have been developed for association mapping (Spielman et al. 1993; Abecasis et al. 2000; Pritchard et al. 2000b; Dudbridge 2003). The focus has been on human studies of unrelated individuals (case–control designs) or unrelated families (family-based designs). However, association studies in plants often involve data collected from highly structured populations (Flint-Garcia et al. 2003). Population structure can cause spurious marker-trait associations (Pritchard et al. 2000b). For the analysis of data from these populations, QK-based

association mapping has been found to be well suited (Yu et al. 2006; Zhao et al. 2007).

In QK-based association mapping, a separate linear mixed model is constructed for each marker locus and fitted to the data. The mixed model includes an intercept, a fixed marker effect, a fixed effect with design matrix $\boldsymbol{Q}$ to account for population structure, a random effect with variance matrix $\boldsymbol{K}$ to handle familial relatedness, and an error. The random effect for familial relatedness and the error are assumed to have normal distributions with zero mean and variances $\sigma_g^2\boldsymbol{K}$ and $\sigma_e^2\boldsymbol{I}$, respectively. The design matrix $\boldsymbol{Q}$ is formed from marker information (Pritchard et al. 2000a; Price et al. 2006). The variance matrix $\boldsymbol{K}$ is constructed either from the marker data or with pedigree information (Stich et al. 2008). The statistical significance of the fixed marker effect in the model is a measure of the strength of association between the marker locus and trait.

Analyzing data from genome-wide studies in plants with QK-based association mapping is not without its challenges. First, there is the computational challenge. QK-based association mapping requires the fitting of a large number of linear mixed models since each marker locus is analyzed separately. In humans, this problem has been solved. Extremely efficient computer programs have been developed that include EMMA (Kang et al. 2008), FaST-LMM (Lippert et al. 2011), and GEMMA (Zhou and Stephens 2012). However, these implementations are based on a model that contains fixed effects and only two random effects, one of which is an uncorrelated error. Modification is required before these computer programs could accommodate the more complicated models used in plants. Instead, in plants, it is common to reduce the computational complexity of fitting a linear mixed model to data by implementing it as a two-stage analysis. Estimates, which are usually predicted means, are calculated from the first-stage model and used as trait data in the second stage. Results from two-stage analyses are approximate (Smith et al. 2001b) but Piepho et al. (2012) show how a two-stage model can be formulated that is highly efficient, as long as the first-stage variance components can be estimated accurately. Unfortunately, when the number of genotypes per environment is large, fitting the second-stage model can become challenging, computationally. Instead, we present a less computationally demanding weighted two-stage analysis for genome-wide association mapping, Weighted two-stage analyses have been found to perform well across a range of experimental designs (Mohring and Piepho 2009; Welham et al. 2010).

Second, there is the challenge of how best to control the type I error rate (or false positive rate) within the QK framework. With QK-based association mapping, a separate hypothesis test is performed for each marker locus. It is from this test that the statistical significance of a fixed marker effect is obtained. However, the resulting p values do not take into account that multiple tests are being performed. If the p values are not adjusted correctly for multiple testing, the type I error rate becomes inflated. Bonferroni and Sidak (Sidak 1967) are two well know approaches for correcting for multiple testing. They are easy to implement and offer strong control of the genome-wide (or family-wise) error rate (Dudoit et al. 2003). The difficulty is that both corrections are conservative when tests are dependent, as is the case in genome-wide studies. An alternate approach for controlling type I error rates in genome-wide studies is via permutation (Churchill and Doerge 1994). Empirical thresholds calculated with permutation are the gold standard in QTL mapping studies. However, the high computational cost of permutation and its inability to account for familial relatedness appropriately (Muller et al. 2011) renders it unsuitable for managing false positives in genome-wide association studies.

In this paper, we present QK-based association mapping for the analysis of data from plant studies that are being performed on a genome-wide scale. Specifically, we have: (a) generalized the standard QK model, enabling a wide variety of experimental design to be modeled; (b) developed a highly accurate procedure for determining when two-stage analysis, which is computationally efficient but the results are approximate, can be implemented without compromising the findings; and (c) developed a Monte Carlo sampling procedure for controlling type I error rates. We conduct a simulation study, with replicates generated with data perturbation, to evaluate the performance of the association mapping technique. Also, we apply QK-based association mapping to the analysis of multiple phase multi-environment data.

Materials and methods

Wheat association study

Phenotypic and genotypic data were collected from a large multi-environment trial in wheat. The trial was run as a multiple-phase experiment. The aim of the study was to map, through association, QTL underlying important quality traits for the sponge-and-dough bread making process. The study consisted of 287 different wheat lines which were partially replicated across sites and over years. These lines were chosen for their phenotypic and genotypic diversity from Australian and International collections. The first phase of the study was a field trial. The trial was performed across three sites in Australia (Biloela, Culara, and Duaringa) and over 3 years (2006, 2007, and 2008) but the trial was not performed at each site for all 3 years (see Table 1). Lines were grown in plots. Plots were randomized in the

Table 1 Number of lines (plots) in the field trial for each site over years

Site	Year	Field Block	
		1	2
Biloela	2006	66 (99)	72 (109)
	2007	106 (136)	107 (136)
	2008	123 (153)	92 (102)
Culara	2007	96 (124)	103 (140)
	2008	104 (132)	103 (124)
Daringa	2006	76 (118)	72 (106)

field as a two-dimensional array with 22 to 72 rows and 6 to 12 columns. The study contained a total of 1479 plots. The lines were assigned to plots according to a randomized design where, on average, 25 % of the lines were replicated within a field block. Each site contained two field blocks. See Table 1 for the number of lines at each site over the 3 years.

The second phase of the study was a milling phase. Grain samples from each field plot were split into two duplicate samples. Samples were milled at Bri Australia Limited according to an optimized Bri method. Samples were allocated to the milling process according to a two-dimensional array design where the milling order are the rows and the milling days are the columns. Samples were assigned to positions in this rectangular schedule according to an incomplete block design with milling days as the blocks.

The third phase of the study was the sponge-and-dough bread making process. It is a two-step process (Lever et al. 2005). In the first step, the sponge was made by mixing 200 g of the total flour with water, yeast, and yeast food at 60 rpm. The sponge was then left to ferment for 4 h. In the second step, the sponge was incorporated with the rest of the flour, and other ingredients to make dough. Each dough sample produced two high pup loaves and one square pup loaf. Loaf volume measurements (cm^3) were made on two high top loaves on the baking day and their mean recorded for analysis. As in the second phase, flour samples were allocated to the baking process according to an incomplete block design with baking days as the blocks. A range of wheat and bread quality traits along with agronomic traits were collected. In this paper, our focus is on loaf volume, an important bread quality trait.

SNP data

Marker genotypes were collected with a 9K SNP array. The SNPs were scored as -1 and 1 as for inbred diploids, even though the wheat lines were allohexaploids ($2n = 7x = 42$). Loci that were monomorphic, had an excessive amount of missing data (>5 %), or were not mapped (Huang et al. 2012), were not considered for analysis. This resulted in genotypes from 3129 SNP across 21 homologous chromosome pairs being retained. The proportion of missing genotypes in these data is 1.7 %.

The QK model

The standard QK model (Yu et al. 2006; Zhao et al. 2007) has a relatively simple form. This is owing to the focus being on human studies. In human studies, there is limited opportunity for experimental control over subjects. Also, human studies do not include replication across environments. However, in plants, experimental designs have been developed that harness randomization, (partial) replication, and blocking for controlling multiple sources of phenotypic variation and correlation. A more general QK-model is needed to accommodate this complexity.

Our linear mixed model for QK-based association mapping, given trait data $\boldsymbol{y}^{(n\times 1)}$ which are collected across s environments and v different plant lines, is

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{Z}_g\boldsymbol{g}_j + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{X}^{(n\times x)}$, $\boldsymbol{Z}^{(n\times a)}$, and $\boldsymbol{Z}_g^{(n\times sv)}$ are design matrices with $\boldsymbol{X}$ being of full rank, $\boldsymbol{\tau}^{(x\times 1)}$ and $\boldsymbol{u}^{(a\times 1)}$ are vectors of fixed and random effects, respectively, that model the experimental design and non-genetic effects, and $\boldsymbol{e}^{(n\times 1)}$ is a vector of residuals. The vector $\boldsymbol{g}_j^{(sv\times 1)}$ contains the genetic effects and is modeled as described below. It is assumed that $\boldsymbol{u}$ and $\boldsymbol{e}$ are uncorrelated and follow a normal distribution with zero mean and variance matrices $\sigma^2\boldsymbol{G}_u$ and $\sigma^2\boldsymbol{R}$, respectively, where σ^2 is an unknown scale variance.

Genetic model The genetic model for standard QK-based association mapping, though rarely stated, is

$$\boldsymbol{g}_j = \boldsymbol{Q\tau}_Q + \boldsymbol{M}_j\boldsymbol{\tau}_j + \boldsymbol{g}_p.$$

Here, it is assumed, often implicitly, that data are collected from a single environment ($s = 1$). The terms in the model are as follows. The first term, $\boldsymbol{Q\tau}_Q$, models the effect of population structure, where $\boldsymbol{\tau}_Q^{(q\times 1)}$ is a vector of fixed population effects and $\boldsymbol{Q}^{(n\times q)}$ is a design matrix containing the probability of a line belonging to a particular subpopulation. The term $\boldsymbol{M}_j\boldsymbol{\tau}_j$ models the effect of the jth marker locus, where $\boldsymbol{\tau}_j^{(m_j\times 1)}$ is a vector of fixed marker locus effects and $\boldsymbol{M}^{(v\times m_j)}$ is a design matrix containing the marker genotypes for the jth locus. The third term, $\boldsymbol{g}_p$, is a random polygenic effect (also known as a variety effect or plant line effect) that models the genetic background. It is assumed that $\boldsymbol{g}_p$ follows a normal distribution with zero mean and variance matrix $\sigma^2\gamma_p\boldsymbol{K}$, where σ^2 and γ_p are unknown variance parameters, and $\sigma^2\gamma_p$ is the genetic variance.

For plant studies performed across multiple (or s) environments, our genetic model is

$$\boldsymbol{g}_j = (\boldsymbol{I}_s \otimes \boldsymbol{Q})\boldsymbol{\tau}_Q + (\boldsymbol{1}_s \otimes \boldsymbol{M}_j)\boldsymbol{\tau}_j + (\boldsymbol{I}_s \otimes \boldsymbol{M}_j)\boldsymbol{g} + \boldsymbol{g}_p, \quad (2)$$

where $(\boldsymbol{I}_s \otimes \boldsymbol{Q})$ and $(\boldsymbol{I}_s \otimes \boldsymbol{M}_j)$ are $sv \times sq$ and $sv \times sm_j$, respectively, design matrices of block diagonal structure, $(\boldsymbol{1}_s \otimes \boldsymbol{M}_j)$ is a $sv \times m_j$ design matrix with the s matrices $\boldsymbol{M}_j$ stacked on top of each other, $\boldsymbol{\tau}_Q^{(sq\times 1)}$ is a vector of fixed population structure by environment effects, $\boldsymbol{\tau}_j^{(m_j\times 1)}$ is a vector of fixed marker locus (main) effects for the j th locus, $\boldsymbol{g}^{(sm_j\times 1)}$ is a random vector of residual marker locus by environment effects, and $\boldsymbol{g}_p^{(sv\times 1)}$ is a vector of polygenic effects. It is assumed that $\boldsymbol{g}$ and $\boldsymbol{g}_p$ are uncorrelated and follow a normal distribution with zero mean and variance matrices $\sigma^2\,\boldsymbol{G}_g \otimes \boldsymbol{I}_v$ and $\sigma^2\,\boldsymbol{G}_p \otimes \boldsymbol{K}$, respectively. The matrices $\boldsymbol{G}_g^{(s\times s)}$ and $\boldsymbol{G}_p^{(s\times s)}$ are genetic variance matrices. The diagonal elements are the genetic variances for the individual environments. The off-diagonal elements are the genetic covariances between pairs of environments. A range of forms are possible (Smith et al. 2001b). An example of how this model might be formulated for a multi-environment experiment is described in Supplementary materials.

Our genetic model deserves some explanation. The first term on the left hand side of model (2) is for the fixed effect of population structure by environment. Here, the effect of population structure is allowed to vary across environments. The second term is a main effect. It measures the overall effect of the jth marker locus on the trait across the s environments. The third term is an error term. It measures the marker locus variation within an environment that is not being explained by the main marker locus effect. The last term is a vector of polygenic by environment effects. Here, we are assuming that the variation explained by the genetic background may vary with environment.

As an explicit example, we now present our QK model for the wheat study described previously. In forming this model, we followed standard model building practices. For the vector of fixed effects $\boldsymbol{\tau}$, it contains an overall mean, and a linear effect for bake order. For the vector of random effects $\boldsymbol{u}$, it contains effects for environments and for environments by columns by rows to model field variation within environments. Typically, $\boldsymbol{u}$ would also include effects for columns and rows to account for additional sources of field variation. However, we found little change in the likelihood through the inclusion of these effects in the model. Thus, they are not include in $\boldsymbol{u}$. The vector $\boldsymbol{u}$ also contains random effects for the milling order, the bake day, and the milling day by milling order. The residuals $\boldsymbol{e}$ are the random effects for environments by bake day by bake order. In this way, we model the variability contained in the three phases of the wheat association study. For variance structures for the random effects, we assume simple diagonal variance matrices. Except, for the environments by columns by rows, the environments by milling day by milling order, and the environments by bake day by bake order (or residual), we instead assume a variance separable structure of the form $\boldsymbol{I}_s \otimes AR1 \otimes AR1$, where $AR1$ denotes the variance matrix for an autoregressive process of order one. We found these simple variance separable forms to suffice, even though other structures are also potentially valid. For the genetic effects $\boldsymbol{g}_i$, they are as previously defined (model 2).

QK-based association mapping

In QK-based association mapping, model (1) is fitted to the data on a marker-by-marker basis. The statistical significance of $\boldsymbol{\tau}_j$ in the model is a measure of the strength of association between the trait and putative QTL that is in linkage disequilibrium with the jth marker locus. That is, we perform the hypothesis test $H_0 : \boldsymbol{L}_j\boldsymbol{\beta}_j = \boldsymbol{0}$, where $\boldsymbol{\beta}_j^{(b_j\times 1)} = \left(\boldsymbol{\tau}^T, \boldsymbol{\tau}_Q^T, \boldsymbol{\tau}_j^T\right)^T$, and $\boldsymbol{L}_j^{(m_j\times b_j)}$ is a matrix of ones and zeros that picks out $\boldsymbol{\tau}_j$ from $\boldsymbol{\beta}_j$.

The Wald statistic W_j to test H_0 is calculated as

$$W_j = \hat{\boldsymbol{\beta}}_j^T \boldsymbol{L}_j^T \left(\boldsymbol{L}_j \left(\boldsymbol{X}_j^T \boldsymbol{H}_j^{-1} \boldsymbol{X}_j\right)^{-1} \boldsymbol{L}_j^T\right)^{-1} \boldsymbol{L}_j \hat{\boldsymbol{\beta}}_j. \quad (3)$$

Here, $\hat{\boldsymbol{\beta}}_j$ is a vector of best linear unbiased estimates of $\boldsymbol{\beta}_j$. The matrix $\boldsymbol{X}_j^{(n\times b_j)}$ is a design matrix partitioned as $\boldsymbol{X}_j = \left[\boldsymbol{X}\ \boldsymbol{Z}_1\ \boldsymbol{Z}_{2,j}\right]$, where $\boldsymbol{Z}_1 = \boldsymbol{Z}_g(\boldsymbol{I}_s \otimes \boldsymbol{Q})$, $\boldsymbol{Z}_{2,j} = \boldsymbol{Z}_g(\boldsymbol{1}_s \otimes \boldsymbol{M}_j)$, and $b_j = x + sq + sm_j$. The matrix $\boldsymbol{H}_j^{(n\times n)}$ is a variance matrix calculated as

$$\boldsymbol{H}_j = \sigma^2\left(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}_u\boldsymbol{Z}^T + \boldsymbol{Z}_{3,j}(\boldsymbol{G}_g \otimes \boldsymbol{I}_v)\boldsymbol{Z}_{3,j}^T + \boldsymbol{Z}_g(\boldsymbol{G}_p \otimes \boldsymbol{K})\boldsymbol{Z}_g^T\right), \quad (4)$$

where $\boldsymbol{Z}_{3,j} = \boldsymbol{Z}_g(\boldsymbol{I}_s \otimes \boldsymbol{M}_j)$. In practice the scale variance σ^2 and the variance parameters of $\boldsymbol{R}$, $\boldsymbol{G}_u$, $\boldsymbol{G}_g$, and $\boldsymbol{G}_p$ are unknown. They are replaced by their estimates from fitting model (1) to the data. Asymptotically, when the null hypothesis is true, W_j follows a Chi-squared distribution with m_j degrees of freedom. From this distribution, an (unadjusted) p value for the significance of $\boldsymbol{\tau}_j$ is obtained.

A Monte Carlo sampling approach for controlling the type I error rate

In this section, we present a Monte Carlo sampling approach for controlling the genome-wide type 1 error rate in genome-wide association studies. Our approach, unlike permutation and other resampling techniques, does not require repeated analysis of realized datasets. Instead, samples are drawn from the joint distribution of the Wald statistic (Eq. 3) under the null hypothesis of no marker-trait

association. These samples then allow adjusted p values and/or genome-wide significance thresholds to be calculated, empirically. Our approach is based on Muller et al. (2011). An early version of our approach is discussed in George (2013) and Muller et al. (2013).

Deriving the joint distribution of the W_js under the, null hypothesis directly, is difficult because the Wald statistic follows a Chi-square distribution. However, we can express W_j as a quadratic form

$$W_j = \hat{\boldsymbol{\omega}}_j^T \boldsymbol{V}_{jj}^{-1} \hat{\boldsymbol{\omega}}_j, \tag{5}$$

where $\hat{\boldsymbol{\omega}}_j^{(m_j \times 1)} = \boldsymbol{L}_j \hat{\boldsymbol{\beta}}_j$, and $\boldsymbol{V}_{jj}^{(m_j \times m_j)} = \boldsymbol{L}_j \left(\boldsymbol{X}_j^T \boldsymbol{H}^{-1} \boldsymbol{X}_j \right)^{-1} \boldsymbol{L}_j^T$. Here, $\boldsymbol{H}$ (Eq. 4) is unknown. An estimate though, $\hat{\boldsymbol{H}}$, can be obtained by fitting model (1) to the data, but with the terms $(\mathbf{1}_s \otimes \boldsymbol{M}_j)\boldsymbol{\tau}_j$ and $(\boldsymbol{I}_s \otimes \boldsymbol{M}_j)\boldsymbol{g}$ removed. This is equivalent to fitting the model under the assumption of no marker-trait association (Muller et al. 2011).

We express W_j as a quadratic form because $\boldsymbol{\omega}_j$ follows a multivariate normal distribution of dimension m_j with mean $\mathbf{0}$ and variance $\boldsymbol{V}_{jj}$. It then follows that the joint distribution of the $\boldsymbol{\omega}$'s also follows a multivariate normal distribution of dimension m with vector mean $\mathbf{0}$ and variance matrix $\boldsymbol{V}^{(m \times m)}$, where $m = \sum_j m_j$. The covariance between $\hat{\boldsymbol{\omega}}_j$ and $\hat{\boldsymbol{\omega}}_{j'}$ is

$$\boldsymbol{V}_{jj'}^{\left(m_j \times m_{j'}\right)} = \boldsymbol{L}_j \left(\boldsymbol{X}_j^T \hat{\boldsymbol{H}}^{-1} \boldsymbol{X}_j \right)^{-1} \boldsymbol{X}_j^T \hat{\boldsymbol{H}}^{-1} \boldsymbol{X}_{j'} \left(\boldsymbol{X}_{j'}^T \hat{\boldsymbol{H}}^{-1} \boldsymbol{X}_{j'} \right)^{-1} \boldsymbol{L}_{j'}^T.$$

In practice, calculating the covariances between all possible pairs of $\hat{\boldsymbol{\omega}}$ is demanding, computationally, especially if the number of marker loci is large. Instead, we assume $\hat{\boldsymbol{\omega}}$ for unlinked marker loci are uncorrelated. The variance matrix $\boldsymbol{V}$ then becomes a block diagonal matrix where the i th block $\boldsymbol{V}_i$ only contains the covariances between $\hat{\boldsymbol{\omega}}$ for the marker loci in the ith linkage group.

Our Monte Carlo sampling approach for calculating adjusted-p values is as follows. First, we calculate $\hat{\boldsymbol{H}}$ from the variance component estimates obtained from fitting model (1) to the data but without the $(\mathbf{1}_s \otimes \boldsymbol{M}_j)\boldsymbol{\tau}_j$ and $(\boldsymbol{I}_s \otimes \boldsymbol{M}_j)\boldsymbol{g}$ terms. Second, we calculate the covariance matrix $\boldsymbol{V}$ assuming a block diagonal structure as described above. Third, we draw N vector samples $\boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \ldots, \boldsymbol{\omega}^N$ from the multivariate normal $N(\mathbf{0}, \boldsymbol{V})$, where the jth element in $\boldsymbol{\omega}^i$ corresponds to the ith realization of $\boldsymbol{\omega}_j$. The vector $\boldsymbol{\omega}^i$ contains a genome-wide set of realizations. Fourth, via Eq. (5), $\boldsymbol{\omega}^1, \boldsymbol{\omega}^2, \ldots, \boldsymbol{\omega}^N$ are converted into Wald statistics. Fifth, (unadjusted) p values $\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^N$ are obtained from the Wald statistics. Each vector $\boldsymbol{p}^i$ contains the p values for all the marker loci in the genome-wide study. Sixth, we obtain the minimum p value from each vector $\boldsymbol{p}^i$. Seventh, we calculate the adjusted-p value for the genome-wide significance of $\boldsymbol{\tau}_j$ by calculating $(a+1)/(N+1)$, where a is the number of minimum p values smaller than the unadjusted $p$value for $\boldsymbol{\tau}_j$ (North et al. 2002).

Weighted two-stage analysis

We present a weighted two-stage analysis for QK-based association mapping. Our analysis approach is based on the two-stage analysis of Smith et al. (2001a). Two-stage analyzes are employed when single-stage analysis is computationally demanding. For our two-stage analysis, model (1) is broken into a first-stage and a second-stage model. Much of the computational complexity in fitting model (1) is captured in the first-stage model which is fitted only once to the data. It is the second-stage model that is fitted on a marker-by-marker basis but this model is of a simplified form. Results from two-stage analyzes are approximate.

Our weighted two-stage analysis approach is as follows.

Stage 1: *fit first-stage QK models* For the first-stage of our weighted two-stage analysis, we analyze each environment's data separately. That is, we fit s single-environment models to the data. From these models, line effect estimates are obtained. These estimates become the (adjusted) trait data for the second stage of analysis.

Suppose for environment i, trait data $\boldsymbol{y}_i^{(n_i \times 1)}$ are collected from v_i different plant lines which have been assigned to n_i plots. The first-stage QK model for these data is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{X}_g\boldsymbol{\tau}_g + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{6}$$

where the i subscript has been dropped from the model for notational convenience, the terms $\boldsymbol{X}\boldsymbol{\tau}$, $\boldsymbol{Z}\boldsymbol{u}$, and $\boldsymbol{e}$ are defined as for model (1), and $\boldsymbol{\tau}_g^{(v \times 1)}$ is a vector of fixed plant line effects with design matrix $\boldsymbol{X}_g^{(n \times v)}$.

By fitting this model to the data, we obtained best linear unbiased estimates of $\boldsymbol{\tau}_g$. These estimates $\hat{\boldsymbol{\tau}}_g$ could be used as (adjusted) trait data for the second-stage of analysis. Smith et al. (2001a) though present a two-step approach for estimating $\boldsymbol{\tau}_g$, that we have adapted, that increases the statistical efficiency of two-stage analyzes. First, we fit

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{g}_p + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{7}$$

where $\boldsymbol{g}_p$ is a vector of random line effects with design matrix $\boldsymbol{Z}_g$ as defined in the section The QK model. Second, model (6) is fitted as before except now its variance parameters are set to the variance estimates obtained from fitting model (7) to the data. The trait data for the second stage of the weighted two-stage analysis is, as before, $\hat{\boldsymbol{\tau}}_g$. However, these estimates are now based upon a model that has its variance parameters set.

Weights for the second-stage analysis are formed from

$$\text{var}\left(\hat{\boldsymbol{\tau}}_g\right) = \boldsymbol{L}\left(\boldsymbol{X}_*^T\hat{\boldsymbol{H}}^{-1}\boldsymbol{X}_*\right)^{-1}\boldsymbol{L}^T, \tag{8}$$

where $\hat{\boldsymbol{\tau}}_g = \boldsymbol{L}\hat{\boldsymbol{\beta}}$, the matrix $\boldsymbol{L}^{(v\times b)}$ contains ones and zeros, $\hat{\boldsymbol{\beta}}^{(b\times 1)}$ is the vector of fixed effects with $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\tau}}^T, \hat{\boldsymbol{\tau}}_g^T)^T$, and $\boldsymbol{X}_*^{(n\times b)}$ is a design matrix partitioned as $\boldsymbol{X}_* = [\boldsymbol{X}\ \boldsymbol{X}_g]$. The estimated variance matrix $\hat{\boldsymbol{H}}$ is $\hat{\boldsymbol{H}} = \hat{\sigma}^2(\boldsymbol{Z}\hat{\boldsymbol{G}}_u\boldsymbol{Z}^T + \hat{\boldsymbol{R}})$ with variance parameter estimates obtained from model (7).

Stage 2: *fit weighted second-stage QK model on a marker-by-marker basis* For the second stage of our two-stage analysis, we fit a multi-environment model to the data. The trait data $\hat{\boldsymbol{\tau}}_g^{(n\times 1)} = \left(\hat{\boldsymbol{\tau}}_{g_1}^T, \hat{\boldsymbol{\tau}}_{g_2}^T, \ldots, \hat{\boldsymbol{\tau}}_{g_s}^T\right)^T$ are estimates obtained from the s first-stage analyzes. The second-stage QK model for these data is

$$\hat{\boldsymbol{\tau}}_g = \mathbf{1}\mu + \boldsymbol{Z}_g\boldsymbol{g}_j + \boldsymbol{\epsilon}_* \tag{9}$$

where μ is an overall mean, $\boldsymbol{g}_j$ is a vector of genetic effects as specified for model (2), and $\boldsymbol{\epsilon}_*$ is a vector of random effects with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \hat{\sigma}^2\,\hat{\boldsymbol{\Omega}})$. The vector term $\boldsymbol{\epsilon}_*$ is included in the model to account for uncertainty in the estimates $\hat{\boldsymbol{\tau}}_\vartheta$. The scale variance $\hat{\sigma}^2$ and variance matrix $\hat{\boldsymbol{\Omega}}$ are set to estimates obtained from the first-stage analysis.

Ideally, $\hat{\boldsymbol{\Omega}}^{(n\times n)}$ is estimated as a block diagonal matrix with the ith block being var $(\hat{\boldsymbol{\tau}}_{g_i})$ (Eq. 8). However, computationally, this is impractical. Instead, $\hat{\boldsymbol{\Omega}}^{-1}$ is treated as a diagonal matrix whose diagonal elements are weights formed from the corresponding diagonal elements of var $(\hat{\boldsymbol{\tau}}_{g_i})^{-1}$. We use the pooled estimator of Smith et al. (2001a) to estimate $\hat{\sigma}^2$. The statistical significance of $\boldsymbol{\tau}_j$ is determined as described in the Sect. "QK-Based Association Mapping".

A heuristic procedure for choosing between single-stage and two-stage analysis

Single-stage analysis should always be favored over two-stage analysis. Single-stage analyzes yield exact results. Two-stage analyzes yield approximate results. In some circumstances though, it may be desirable to sacrifice accuracy for computational expediency. The challenge is how to determine the genome-wide accuracy of two-stage analysis without benchmarking against genome-wide results from single-stage analysis. There is no published solution to this problem.

To solve this problem, we have developed a heuristic procedure that we describe below and that we have found performs exceedingly well. Our procedure does require two-stage analysis to be performed on a genome-wide scale. However, these results would be needed anyway if two-stage analysis was found to be valid. With our procedure, the more computationally demanding single-stage analysis, in practice, is only needed to be performed on a small number of select marker loci. Yet, we can still use these single-stage results to determine the overall validity of the genome-wide two-stage results.

Our heuristic procedure is as follows. First, raw p values p_j^{unadj} are calculated via single-stage analysis for each marker locus j. Adjusted p values, (p_j^{adj}), are calculated via our Monte Carlo sampling approach. Second, the distance of $-\log_(p_j^{\text{adj}})$ from its significance threshold x is calculated as $|-\log p_j^{\text{adj}} - x|$. For a genome-wide significance of 5 %(1 %), x is 1.3(2.0). The distances are placed in ascending order. Third, q_k^{unadj}, is calculated via single-stage analysis where k is the kth ordered distance in the previous step. Fourth, the test $|-\log(p_k^{\text{unadj}}) - \log(q_k^{\text{unadj}})| > |-\log p_k^{\text{adj}} - x|$ is performed. If true, a single-stage analysis is required as wrong results will be inferred if a two-stage analysis is implemented. If false, the second and third steps are repeated for the next ordered p value. Our iterative procedure is complete when a single-stage analysis is found to be required or $|-\log p_k^{\text{adj}} - x| > z$ which means a two-stage analysis can be implemented. Here, larger values of z give greater protection against wrongly implementing QK-based association mapping as a two-stage analysis. However, larger values of z may also mean a larger number of single-stage analyzes need to be performed. We have found $z = 1$ to work well in practice.

Dealing with missing marker genotypes

In genome-wide studies, it is rarely possible to observe all marker loci fully. We need an approach for dealing with missing marker genotypes. One approach is to ignore those plants with missing genotypes for the locus being tested. Alternately, missing genotypes could be replaced with expected values if allele frequencies are known at the population level. The approach of choice though is genotype imputation. Genotype imputation methods are based, typically, on a hidden Markov process that is built from a population genetics model. They work best when data are available on a dense screen of marker loci that are collected from a reference sample. The haplotype structure learnt from the reference sample guides imputation of the missing genotypes in the study sample. However, densely mapped reference populations are not typically available in plants.

Our approach is based on how marker loci are modelled in QK-based association mapping. In QK-based association mapping, the strength of association between a marker locus and trait is measured by fitting the marker locus as a fixed effect in the linear mixed model. Whether the fixed marker effect is a covariate or factor is dependent upon the ploidy level, the informativeness of the marker locus, whether the plant is inbred or outbred, and whether

the plant is autopolyploid or allopolyploid. For dominant markers or SNP markers collected on inbred diploids or inbred allopolyploids, the marker locus is treated as a covariate. The covariate measures the additive effect of a QTL in linkage disequilibrium with the marker locus. In all other cases, the marker locus is treated as a factor where the number of levels corresponds to the number of unique genotypes observed. As a factor, we are measuring the genotypic effect of a QTL in linkage disequilibrium with the marker locus.

Our goal is to impute missing data in a way that is simple, does not require the marker loci to be mapped, and has minimal impact on the statistical significance of the fixed marker effect. In deriving our approach, we make the assumption that each possible genotype at a marker locus has equal probability of being missing. That is, we assume marker genotypes are missing at random. If the marker locus is being modelled as a covariate, then we normalize its marker data and replace missing genotypes with zero values. Normalizing the data of a covariate does not affect its statistical significance. If the marker locus is being modelled as a factor, then we create a new factor level and assign it to those plants with missing marker genotypes. In this way, we are able to impute missing genotypes with minimal impact on the statistical significance of the fixed marker effect.

Simulation study

Replicates are generated with data perturbation (Zhao et al. 2007). Data perturbation makes use of the phenotypic and genotypic data observed in a real study to create a simulated (quantitative) trait. It affords us the opportunity to generate replicates whose analysis more closely mirrors the complexities of analyzing real data. For computational expediency, our simulation study is based on data observed from the Biloela site in 2006. In generating trait data for a replicate, we assume a broad based heritability of 0.6, there is a single major QTL, the polygenic component consists of 30 polygenes where the effect of a polygene is formed from an unmapped SNP which was chosen randomly, and the variance structure of the simulated trait closely follows that of the loaf volume trait. The marker data consists of those genotypes collected in the association study from the 3129 SNP across the 21 homologous chromosomes. Each replicate has the same marker data. It is the simulated trait data that varies across replicates.

Our simulation study consists of three parts. First, we conduct a null study to investigate whether our Monte Carlo sampling approach controls correctly the genome-wide type I error rate. We also evaluate the impact of assuming a block diagonal structure for $\boldsymbol{V}$. We do this by also calculating the genome-wide type I error rates when the variance matrix for the joint distribution of the test statistic is formed from loci pairings across the entire genome. Ten thousand replicates are generated where the QTL has no effect. Results are reported for QK-based association mapping implemented as a single-stage and weighted two-stage analysis.

Second, a power study is performed to investigate if there is a difference in performance between implementing QK-based association mapping as a single-stage analysis versus a two-stage analysis. Data are generated under QTL of different sizes and with varying percentages of missing marker data (1.1, 5, 9, 11 %). Missing marker genotypes are imputed for a marker locus by coding its SNP genotypes as –1 and 1, normalizing its marker data, and replacing missing genotypes with zero values. One thousand replicates are generated for each of the four different percentages of missing marker data and each differently sized QTL. A genome-wide significance level of 5 % is assumed.

Third, a study to evaluate our heuristic procedure for choosing between single-stage and two-stage analysis is performed. Data are generated under a QTL of no, moderate, and large effect. One thousand replicates are generated for each differently sized QTL. The performance of our heuristic procedure is measured by calculating the proportion of replicates for which the analysis type is inferred correctly. To determine if a single-stage analysis is truly needed, it is necessary to analyze fully the genome-wide data via single-stage and weighted two-stage analysis. Then, for each marker locus j, perform the test $|-\log p_j^{\text{adj}} - \log q_j^{\text{adj}}| > |-\log p_j^{\text{adj}} - 1.3|$, where we have assumed a genome-wide significance level of 5 %. If true, an incorrect finding occurs if our heuristic procedure determines that it is valid to implement two-stage analysis.

Implementation

Our pipeline is implemented within the R environment (R Core Team 2013), but $\boldsymbol{Q}$ and $\boldsymbol{K}$ are calculated outside of R. The elements of $\boldsymbol{Q}$, the probability of plants belonging to a given subpopulation, are calculated with the computer program STRUCTURE 2.3.4 (Pritchard et al. 2000a). The elements of $\boldsymbol{K}$ are calculated with the codominant marker-based estimator of Ritland (1996) as implemented in the computer program SPAGeDi 1.3 (Hardy and Vekemans 2002). The linear mixed models are fitted with residual maximum likelihood as implemented in the R package ASReml-R (Butler et al 2009). Calculations are distributed across multiple processors with functionality contained within the snow package (Rossini et al. 2007; Tierney et al. 2009). All analyzes are performed on a Linux machine with dual 6-core Intel Xeon Westmere cores.

Results

Control of the genome-wide type I error rate

A null simulation study was performed to evaluate our Monte Carlo sampling approach for controlling the genome-wide type I error rate. Replicates were analyzed with QK-based association mapping implemented as a single-stage and weighted two-stage analysis. The genome-wide type I error rates are given in Table 2. We found almost no difference in type I error rates when $\boldsymbol{V}$ is approximated via our block diagonal approach compared to constructing the covariance matrix from all possible loci pairings (Genome-wide). Also, our Monte Carlo sampling approach gave type I error rates that were very close to the nominal significance levels regardless of the analysis approach implemented. For completeness, we also corrected for multiple testing with the Bonferroni correction. As expected, the Bonferroni correction was conservative.

Power study

The power curves for the single-stage and weighted two-stage analyzes of the simulated data are shown in Fig. 1. As expected, for a given percentage of missing marker data, power increases as the amount of phenotypic variation explained by a QTL is increased. Conversely, power decreases as the percentage of missing marker data increases for a QTL of a certain size. It is reassuring that the close agreement between single-stage and weighted two-stage results is not affected by the size of the QTL or the amount of missing marker data.

Evaluation of our heuristic procedure for choosing the validity of results from two-stage analysis

Replications were generated under the null hypothesis of no QTL, a moderately size QTL that explains 5 % of the phenotypic variation, and a large QTL that explains 10 % of the phenotypic variation. For each replicate, we ran our heuristic procedure to determine if it was valid to implement QK-based association mapping as a two-stage analysis. A 5 % genome-wide significance and a $z = 1$ (see Sect. 2.7 for details) were assumed. We also performed genome-wide single-stage analyzes to determine if two-stage analysis was indeed yielding valid results. We found that our heuristic procedure performed extremely well. For no QTL, our procedure correctly determined with 100 % (96 %) accuracy when single-stage (two-stage) analysis was required. For a moderately sized QTL, our procedure performed with 99 % (87 %) accuracy and for a large QTL, our procedure performed with 98 % (83 %) accuracy. Our heuristic procedure did find it more difficult, for moderate to large QTL, to correctly determine that two-stage analyzes would in fact yield reliable results. However, concluding wrongly that single-stage analysis is needed when two-stage analysis would suffice only means that in some situations, we could implement QK-based association mapping more efficiently.

Table 2 Genome-wide type I error rates for single-stage and two-stage analysis with p values adjusted for multiple testing via our Monte Carlo sampling approach and via the Bonferroni correction

Multiple testing Procedure	$\boldsymbol{V}$ Calculation	Analysis approach	Genome-wide significance level		
			1 %	5 %	10 %
Monte Carlo	Linkage group	Single-stage	0.009	0.054	0.108
		Two-stage	0.010	0.055	0.109
	Genome-wide	Single-stage	0.008	0.053	0.110
		Two-stage	0.009	0.054	0.111
Bonferroni		Single-stage	0.006	0.022	0.052
		Two-stage	0.006	0.022	0.052

The variance matrix $\boldsymbol{V}$ is calculated two different ways; for the covariance of $\boldsymbol{\omega}$ between linked marker loci only (Linkage group) and between all marker loci (Genome-wide)

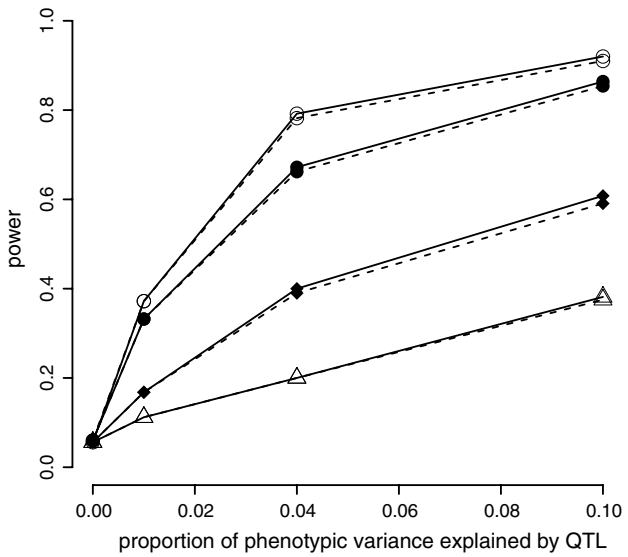

Fig. 1 Power curves for single-stage and weighted two-stage analysis of simulated data with differing percentages of missing marker data. Solid (*dashed*) lines are for single-stage (two-stage) analysis. *Unfilled circles*, *filled circles*, *filled diamonds*, and *unfilled triangles* denote results from analyzing data where the percentage of missing marker data is 1.1, 5, 9, and 11 %, respectively. The single-stage and two-stage results are near identical

Table 3 Summary statistics and heritability for the loaf volume trait for each environment

Environment		Number of lines	Mean (cm ^{3})	Range (cm ^{3})	Coefficient of variation	Heritability
Site	Year					
Biloela	2006	110	903.2	788–1043	0.048	0.39
	2007	177	930.6	830–1040	0.045	0.50
	2008	193	888.7	785–1038	0.055	0.69
Duaringa	2006	112	923.6	808–1050	0.046	0.27
Culara	2007	169	916.3	800–1030	0.047	0.44
	2008	189	890.8	745–1003	0.042	0.31

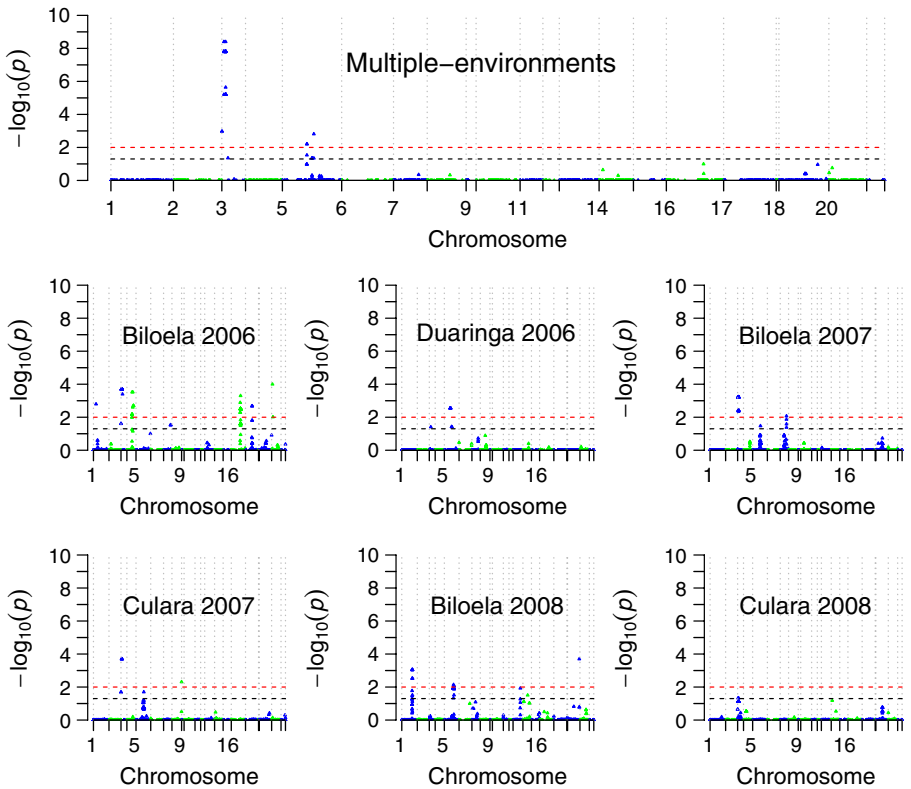

Fig. 2 Manhattan plots of adjusted $-\log p$ values from QK-based association mapping implemented as a single-stage analysis. The first plot is the Manhattan plot for the analysis of the multi-environment data. The following six plots are the Manhattan plots for the analysis of the single-environment data. The *gray vertical dash lines* denote the ends of the chromosomes. The *black horizontal dash line* is the 5 % genome-wide significance threshold. The *red horizontal dash line* is the 1 % genome-wide significance threshold. Chromosomes labeled $1, 2, 3, \ldots, 20, 21$ correspond to chromosomes $1A, 1B, 1D, \ldots, 7B, 7D$

Wheat association study

Summary statistics for loaf volume trait and heritability

The mean loaf volume trait varied significantly ($p < 0.01$) across sites and years (Table 3). The Biloela site recorded the highest (930.6 cm ^{3}) and lowest (888.7 cm ^{3}) mean loaf volumes of the six environments. The range of loaf volume values were fairly consistent across environments as were the coefficients of variation. The generalized heritability, calculated according to Eq. (7) in Oakey et al. (2006), was low to moderate (0.31–0.69) with five of six environments with heritability 0.5 or less. This is consistent with other published heritability estimates for loaf volume.

Genome-wide association mapping

Here, we present results from our genome-wide analyzes of the wheat association data. The data were analyzed with QK-based association mapping implemented as a single-stage analysis and for comparison, a weighted two-stage analysis. We began with single environment (or univariate) analyzes and then performed a multi-environment (or multivariate) analysis. The Manhattan plots with adjusted $-\log p$ values from the single-stage analyzes for each environment and the multi-environment trial are shown in Fig. 2. Manhattan plots with unadjusted $-\log p$ values from the single-stage analyzes are shown in Figure S1 in Supplementary materials. A much cleaner picture of the

marker-trait associations is obtained with Manhattan plots with adjusted p values. Also the points are immediately interpretable.

Single-stage and weighted two-stage analyzes gave very similar $-\log p$ values. The Pearson correlation between the exact and approximate $-\log p$ values for the single environment analyzes ranged from 0.982 to 0.998 and was 0.992 for the multi-environment data. Figure S2 in Supplementary materials shows the close agreement between the two implementations of QK-based association mapping for the analysis of the multi-environment data. However, there is a significant difference in computing time. Single-stage analysis of the single environment data took 13–16 min. Two-stage analysis took one to one and a half minutes. Single-stage analysis of the multi-environment data took 37.96 h. Two-stage analysis took only 3.03 h. Two-stage analysis was in excess of ten times faster than single-stage analysis. Our heuristic procedure correctly determined that two-stage analyzes would yield reliable results for the single environment and multi-environment data.

Discussion

Being able to conduct association studies on a genome-wide scale is an exciting development. Association studies map QTL with high resolution. They also have a broad genetic base, making them well suited for unlocking the genetic secrets of complex traits. However, potentially the greatest benefit to association studies in plants is rarely emphasized; they allow QTL studies to move away from costly purpose-built biparental crosses to unstructured populations. For the first time, genetic discovery can be performed opportunistically. The study presented in this paper is a case in point. It began as a study to better understand the phenotypic structure of important traits in the sponge-and-dough bread making process. However, we saw an opportunity to broaden its scope. By collecting genome-wide marker data, we turned the study into an association study. With genotyping costs continuing to drop, it will become increasingly economical to expand a study to also include genome-wide association mapping.

One of the greatest obstacles to routine implementation of QK-based association mapping in plants is its high computational cost. Genotypes are often collected in plant studies on fewer marker loci than in human studies, but plant studies generally follow a far more elaborate design structure. This makes fitting the QK model computationally expensive. We have addressed this issue statistically. We presented a new efficient (weighted) two-stage analysis for QK-based association mapping. We also developed a procedure to govern its deployment. However, there is also opportunity here to harness distributed computing. QK-based association mapping is easily parallelizable. Calculation of the unadjusted p values can be done in parallel. Even the computation of the covariances in $\boldsymbol{V}$, the most computationally demanding part of our Monte Carlo sampling approach, can be distributed. With languages such as R now having the capability to harness distributed computing, substantial reductions in computing time can be achieved relatively easily.

We acknowledge that other formulations of our genetic model (model 2) are also possible. For an interesting review of the different ways in which a mixed model can be specified for multi-environment data, see van Eeuwijk et al. (2010). For example, instead of treating a marker locus effect as fixed, we could model it as a random effect. However, we wanted to preserve the original intent of QK-based association mapping where the strength of marker-trait associations are measured via the significance of fixed effects. The impact, consequences, and benefits of different formulations of model (2) requires further investigation that is beyond the scope of this paper.

Our heuristic procedure for choosing between single-stage and two-stage analysis deserves some explanation with regard to how it was developed. We began with a procedure that chose, randomly, a small number of marker loci for analysis, calculated raw p values via single-stage and weighted two-stage analysis, adjusted the raw p values for multiple testing, and calculated the mean squared error (MSE) of the $-\log$ ratio of the adjusted p values. We surmised that two-stage analyzes, when implemented on a genome-wide scale, would yield reliable results when the MSE was small. However, analyzes of simulated data revealed this to be false. We also tried replacing the MSE calculation with the test $|-\log(p_j^{\text{adj}}) + \log(q_j^{\text{adj}})| > |-\log(p_j^{\text{adj}}) - x|$, where log is to base 10, x is the $-\log$ of the genome-wide significance level and p_j^{adj} (q_j^{adj}) is the adjusted p value for $\boldsymbol{\tau}_j$ for the weighted two-stage (single-stage) analysis. This test was motivated by the premise that some inaccuracy in the estimation of the p values from two-stage analyzes can be tolerated. As long as these errors do not culminate in wrong acceptance/rejection of single-stage findings. However, this approach also failed to guard against spurious findings. The test was sensitive to the number of marker loci being sampled and the genome-wide significance of those markers. Performance could only be guaranteed with a prohibitively large sample of marker loci.

There are several extensions to our work that we are pursuing. First, our QK-model is a multivariate model which we used to analyze multi-environment data. However, our model could also be applied to the detection of marker-trait associations in data collected on multiple correlated traits. Second, even by simplifying the structure of $\boldsymbol{V}$ to being block diagonal, its calculation can still be time consuming if

the number of linked marker loci is large. To solve this problem, we are exploring clustering techniques and principal component analysis as a means of reducing the dimension of $\boldsymbol{V}$. Results are preliminary but promising. Third, there is opportunity for a more in-depth analysis of the wheat association data. A range of different trait data were collected at each of the three study phases inviting analysis. Also, we are building a denser marker map that will improve mapping resolution. Fourth, we are working on an R package that implements the methods presented in this paper.

Over the coming years, association studies will push even harder against the boundaries of computational tractability. New next-generation based genotyping technologies such as genotyping-by-sequencing are already being perfected. These technologies give opportunity for large volumes of genotypes to be collected from plants with genomes that are not even sequenced. The continued development of new statistical and computational tools for association mapping is needed. Genome-wide association studies in plants will always be challenging to analyze, but with this challenge also comes great reward.

Author contribution statement AWG developed the methods, implemented the methods in R, performed the analyzes. CC was involved in the design and collection of the association data.

Acknowledgments We are grateful to Prof. Arunas Verbyla and Dr. Julian Taylor for many helpful discussions when developing this work, Dr. Alison Kelly for constructing base models of the phenotypic data, and the Department of Primary Industries for giving us access to the phenotypic data from the wheat study.

Conflict of interest The authors declare no conflict of interest.

References

Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66(1):279–292

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml User Guide. URL: http://www.vsni.co.uk/software/asreml/

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138(3):963–971

Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25(2):115–121

Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Stat Sci 18(1):71–103

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

George AW (2013) Controlling type 1 error rates in genome-wide association studies in plants. Heredity 111:86–87

Hardy OJ, Vekemans X (2002) Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2(4):618–620

Huang B, George A, Forrest K, Kilian A, Hayden M, MK M, Cavanagh C, (2012) A multiparent advanced generation intercross population for genetic analysis in wheat. J Plant Biot 10(7):826–839

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

Lever T, Kelly A, De Faveri J, Martin D, Sheppard J, Quail K, Miskelly D (2005) Australian wheat for the sponge and dough bread making process. Aust J Agr Res 56(10):1049–1057

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) Fast linear mixed models for genome-wide association studies. Nat Methods 8:833–835

Mohring J, Piepho HP (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci 49:1977–1988

Muller BU, Stich B, Piepho HP (2011) A general method for controlling the genome-wide type i error rate in linkage and association mapping experiments in plants. Heredity 106(5):825–831

Muller BU, Stich B, Piepho HP (2013) Response to controlling type 1 error rates in genome-wide association studies in plants. Heredity 111:88

North BV, Curtis D, Sham PC (2002) A note on the calculation of empirical p values from monte carlo procedures. Am J Hum Genet 71(2):439–441

Oakey H, Verbyla A, Pitchford W (2006) Joint modeling of additive and non-additive genetic line effects in single field trials. Theor Appl Genet 113:809–819

Piepho HP, Mohring J, Schulz-Streeck T, Ogutu J (2012) A stage-wise approach for the analysis of multi-environment trials. Biom J 54:844–860

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155(2):945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67(1):170–181

R Core Team (2013) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org/

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res 67(2):175–185

Rossini AJ, Tierney L, Li N (2007) Simple parallel statistical computing in r. J Comput Graph Stat 16(2):399–420

Sidak Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62:626–633

Smith A, Cullis B, Gilmour A (2001a) The analysis of crop variety evaluation data in australia. Aust NZ J Stat 43(2):129–145

Smith A, Cullis B, Thompson R (2001b) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138–1147

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium—the insulin gene region and insulin-dependent diabetes-mellitus (IDDM). Am J Hum Genet 52(3):506–516

Stich B, Mohring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178(3):1745–1754

Tierney L, Rossini AJ, Li N (2009) Snow: A parallel computing framework for the r system. Int J Parallel Prog 37(1):78–90

van Eeuwijk FA, Bink MCAM, Chenu K, Chapman SC (2010) Detection and use of qtl for complex traits in multiple environments. Curr Opin Plant Biol 13:193–205

Welham S, Gogel B, Smith A, Thompson R, Cullis B (2010) A comparison of analysis methods for late-stage variety evaluation trials. Aust NZ J Stat 52:125–149

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An arabidopsis example of association mapping in structured samples. Plos Genetics 3(1)

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