Theor Appl Genet (2014) 127:1753–1770
DOI 10.1007/s00122-014-2337-4

ORIGINAL PAPER

Whole-genome QTL analysis for MAGIC

Arūnas P. Verbyla · Andrew W. George · Colin R. Cavanagh · Klara L. Verbyla

Received: 17 October 2013 / Accepted: 21 May 2014 / Published online: 14 June 2014

Abstract
Key message **An efficient whole genome method of QTL analysis is presented for Multi-parent advanced generation integrated crosses.**
Abstract Multi-parent advanced generation inter-cross (MAGIC) populations have been developed for mice and several plant species and are useful for the genetic dissection of complex traits. The analysis of quantitative trait loci (QTL) in these populations presents some additional challenges compared with traditional mapping approaches. In particular, pedigree and marker information need to be integrated and founder genetic data needs to be incorporated into the analysis. Here, we present a method for QTL analysis that utilizes the probability of inheriting founder alleles across the whole genome simultaneously, either for intervals or markers. The probabilities can be found using three-point or Hidden Markov Model (HMM) methods. This whole-genome approach is evaluated in a simulation study and it is shown to be a powerful method of analysis. The HMM probabilities lead to low rates of false positives and low bias of estimated QTL effect sizes. An implementation of the approach is available as an R package. In addition, we illustrate the approach using a bread wheat MAGIC population.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2337-4) contains supplementary material, which is available to authorized users.

Communicated by Marco Bink.

A. P. Verbyla (✉)
Computational Informatics and Food Futures National Research Flagship, CSIRO, Atherton QLD 4883, Australia
e-mail: ari.verbyla@csiro.au

A. P. Verbyla
School of Agriculture, Food and Wine, The University of Adelaide, PMB 1, Glen Osmond SA 5064, Australia

A. W. George
Computational Informatics and Food Futures National Research Flagship, CSIRO, Dutton Park QLD 4102, Australia

C. R. Cavanagh
Plant Industry and Food Futures National Research Flagship, CSIRO, Canberra ACT 2601, Australia

K. L. Verbyla
Computational Informatics and Food Futures National Research Flagship, CSIRO, Canberra ACT 2601, Australia

Introduction

Food security is a major challenge facing the human population. Essential to feeding the world's growing population is an increase in agricultural productivity under ever-changing environmental conditions. Critically important is the ability to identify and produce crops with improved quality and yield attributes along with robust disease resistance and broad adaptability in a changing climate. For decades, bi-parental populations have been the mainstay in identifying quantitative trait loci (QTL), which is the first stage in quantifying genes and gene pathways responsible for these economically important traits. However, the restricted nature of such populations has meant that identified QTL can generally only be mapped to wide genomic regions and have not been able to produce the results desired in breeding programs. Conversely, using breeding populations to identify QTL is also not ideal. The increased complexity can introduce difficulties including unknown population structures, thereby distorting the relationships between markers and trait when applying association mapping.

Recently, the generation of multi-parent advanced generation integrated cross (MAGIC) populations has provided an additional option for QTL mapping; this overcomes the complexity of a breeding program, while not being constrained by the limitations of bi-parental populations. MAGIC was first described by Mott et al. (2000), and was used for fine-mapping QTL in heterogeneous mice stock generated from an eight-way cross of diverse strains (HS). Another mice population, the collaborative cross (CC) (Threadgill et al. 2002), was established in 2002, also using an eight-way cross. Cavanagh et al. (2008) first discussed the potential of MAGIC in crops. MAGIC populations provide the ability to capture large proportions of the genetic diversity present in a population through the selection of diverse founder lines, ideally increasing the number of QTL segregating in the cross. The founder lines in a classic MAGIC population are inter-crossed (generally for $n_f/2$ generations where n_f is the number of founder lines) until all founders have an equal probability of contributing to the genetic makeup of a line. This is followed by multiple generations of selfing to create recombinant inbred lines. Such a structure leads to an amplified number of recombination events which increases the mapping resolution of detected QTL. Note that more complex inter-crossing patterns can be used, but the methods developed in this paper are for the classic MAGIC design.

In plants, many populations are being developed across a wide range of species. In durum wheat, Trebbi et al. (2008) report on the development of a 4-way population, while Huang et al. (2012) describe the two MAGIC populations developed in wheat. MAGIC populations have also been generated for *Arabidopsis thaliana* (Kover et al. 2009) and rice (Bandillo et al. 2013) with the purpose of fine-mapping QTL in these species. A review of integrated crosses is provided by Rakshit et al. (2012).

Developing methods of analysis that utilizes the additional information available in a MAGIC population is essential to maximize the usefulness of such a resource. QTL analysis for MAGIC populations was initially discussed by Xu (1996) who uses an interval mapping approach on marker scores for a four-way cross. The regression method (Haley and Knott 1992) was the basis for this analysis. However, Mott et al. (2000) demonstrate that simple marker regression using marker scores fails in the CC population and describes a multi-point method (known as HAPPY). This method begins by constructing the probabilities that an allele has been inherited from the founders, which is an alternative to the marker scores, that is made possible through the population structure of the MAGIC population (see Fig. 1). This can provide valuable information as marker genotypes may not be fully informative with multiple founders. The probabilities that a marker has come from each founder are then used to identify QTL and to establish the size of the QTL for each founder allele. Mott et al. (2000) utilize a Hidden Markov Model (HMM), also described in Broman (2006), to calculate the founder probabilities at each interval. Additional aspects of the multi-point method are considered by Valdar et al. (2006a) and Valdar et al. (2006b), including the potential power of the method in a MAGIC population and approaches for the determination of thresholds.

Kover et al. (2009) extend the work of Mott et al. (2000), Valdar et al. (2006a) and Valdar et al. (2006b) by developing three methods for QTL analysis and then applying the methods to data on *Arabidopsis thaliana*. The first method extended the approach of Mott et al. (2000). A forward selection approach was used after an initial single (fixed effects) marker analysis to allow for multiple QTL and also for possible population structure. The analysis was then repeated by re-sampling the data 500 times and identifying QTL for each re-sampling. The support for a QTL was determined by the fraction of times a QTL was identified in the 500 analyses. In the second approach of Kover et al. (2009), kinship information was included to account for the population structure, similar to that used by Malosetti et al. (2011) in a 3-way barley cross. The third approach used by Kover et al. (2009) was a hierarchical Bayes method. In the analyses of the data presented in Kover et al. (2009), the results of the three approaches were very similar.

The methods of Mott et al. (2000) and Kover et al. (2009) are ultimately based on a genome scan, where each interval is tested independently for evidence of a QTL. Similarly, Huang et al. (2012) use a simple interval mapping approach in a wheat MAGIC population to identify QTL for plant height and hectolitre weight. The use of such approaches has been shown to cause bias in bi-parental populations which has led to the development of composite interval mapping (CIM) (Zeng 1994; Jansen 1994) and multiple interval mapping (Kao et al. 1999). The methods of Mott et al. (2000) could be extended to CIM, however, a viable alternative is whole-genome average interval mapping (WGAIM). Described initially in Verbyla et al. (2007) and modified in Verbyla et al. (2012), WGAIM has been demonstrated to outperform CIM and has been extended to multivariate situations (Verbyla and Cullis 2012).

The WGAIM approach for QTL analysis simultaneously incorporates all marker information in the analysis, overcoming the need for repeated genome scans. In addition, WGAIM uses a forward selection approach greatly reducing the number of analyses. The approach allows for population structure to be modeled through the inclusion of pedigree information and for any non-genetic effects, such as experimental design terms to be easily included. A simple random effects working model is used in which all intervals are allowed to contain a possible QTL. Note that rather than using intervals, the markers themselves can be used in

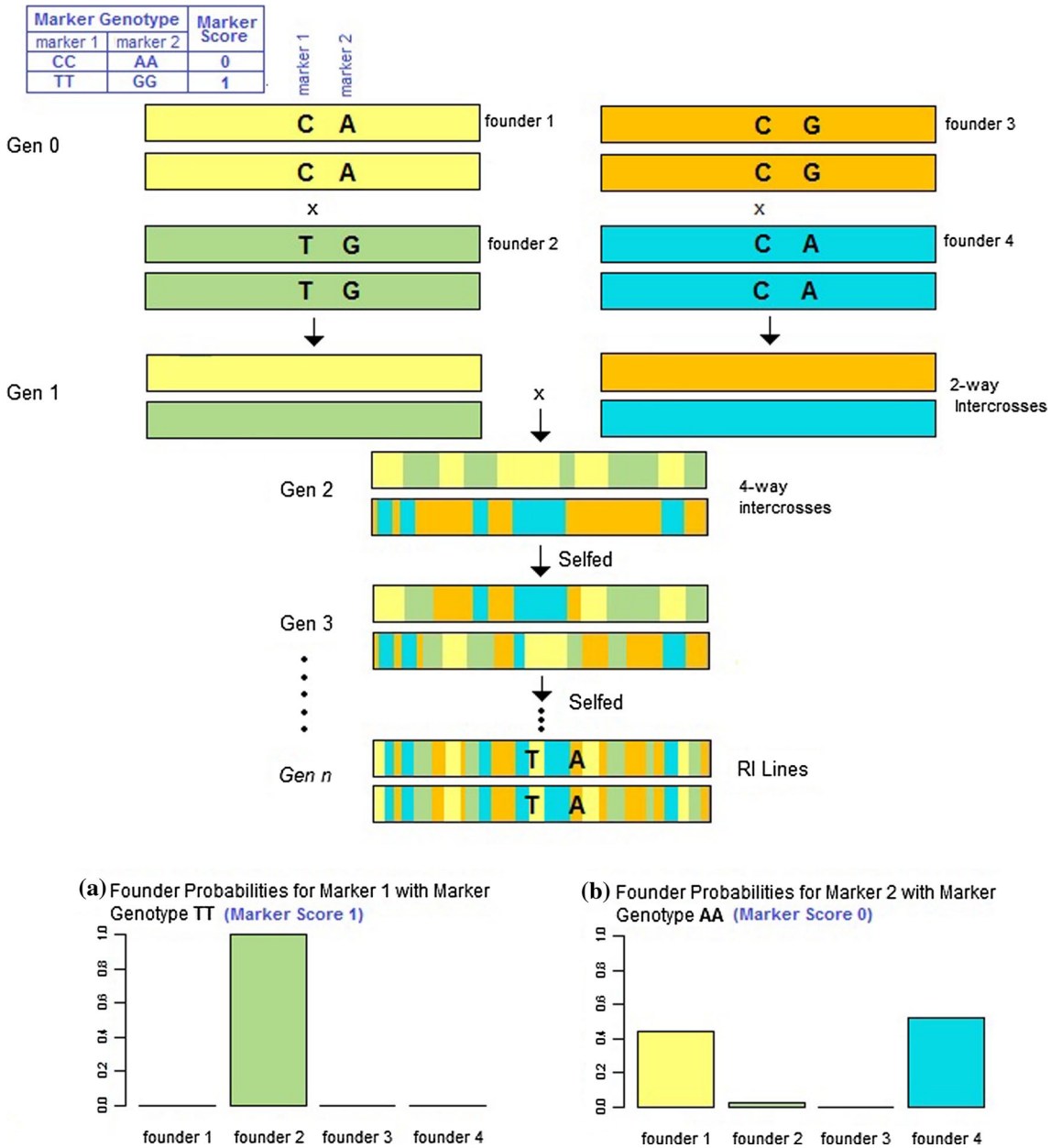

Fig. 1 MAGIC population generation for four founders and two loci tracked through the process. The histograms in **a** and **b** present the probability that the marker represents each founder allele

the analysis, for example when the marker density is high; however, marker effects are assumed uncorrelated in the WGAIM approach whereas using intervals introduces a correlation between markers (Verbyla et al. 2007). A likelihood ratio test of significance of the random effects working model is conducted to decide if selection of a putative QTL is warranted. This formulation means a threshold for QTL detection is readily available.

In this study, we extend WGAIM for use in MAGIC populations, describing a powerful method utilizing founder probabilities through a simulation study and a real data example. Two different approaches for constructing the necessary probabilities of inheriting the founder alleles for each line and each locus, namely the HMM approach and a three-point probability approach, are explored for use both at the markers and within intervals. The simulation study is used to examine the Type I error rate, accuracy of predicted effect sizes and power of the method based on the structure of a real MAGIC wheat population. The simulation study demonstrated the power of the approach when identifying QTL; interestingly this included consistently identifying two QTL when

simulated in close linkage. Utilizing HMM probabilities either with an interval- or marker-based approach is a key recommendation, due to a high rate of false discoveries and biased effect sizes when using three-point probabilities. A comparison with HAPPY (Mott et al. 2000) is also presented in the simulation study and it is shown in that the extension of WGAIM finds more QTL and has far fewer false positives than HAPPY. The paper finishes with the real data example, demonstrating the performance of the method for QTL analysis in a 4-way wheat MAGIC population.

Methods

Linear mixed model

Mixed models form the basis of our analysis. This allows inclusion of both fixed and random effects that are either part of the design of the trial generating the data or are required for an efficient analysis. Thus, if $\mathbf{y}$ is the $n \times 1$ vector of trait data, we suppose

$$\mathbf{y} = \mathbf{X}\tau + \mathbf{Z_0 u_0} + \mathbf{Z_g u_g} + \mathbf{e} \tag{1}$$

where the design matrices $\mathbf{X}$, $\mathbf{Z}_0$ and $\mathbf{Z}_g$ are known. The first two reflect the experimental design of the trial and any other variation that requires modeling (Smith et al. 2005, 2006); τ is the vector of fixed effects parameters, and $\mathbf{u}_0$ is a vector of random effects. The matrix $\mathbf{Z}_g$ assigns each observation to one of n_g varieties or genetic lines. Thus, $\mathbf{u}_g$ is the $n_g \times 1$ vector of genetic effects; the model for $\mathbf{u}_g$ is described below. Finally, $\mathbf{e}$ is the vector of residuals. The two terms $\mathbf{u}_\mathrm{o}$ and $\mathbf{e}$ are assumed to be distributed as

$$\begin{bmatrix} \mathbf{u}_\mathrm{o} \\ \mathbf{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_\mathrm{o} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

where the covariance matrices $\mathbf{G}_\mathrm{o}$ and $\mathbf{R}$ will depend on the study. It is important to model the non-genetic components well to determine the genetic effects of the model correctly.

Multi-parent whole-genome QTL analysis

The approach for MAGIC is based on the whole-genome average interval mapping (WGAIM) presented originally by Verbyla et al. (2007) and more recently by Verbyla et al. (2012). The original WGAIM approach uses all intervals on a linkage map in the analysis and a forward selection approach to choose putative QTL. It is a powerful approach for QTL analysis in bi-parental populations. We call the current approach multi-parent WGAIM or MPWGAIM. MPWGAIM involves using probabilities of inheriting founder alleles for each potential QTL locus and each line. This is the major modification of the original WGAIM approach; there are consequences in the outlier detection step, in the percentage of genetic variance accounted for by each putative QTL and in defining a measure of strength of the QTL.

While the original formulation of WGAIM involves using intervals, the observed markers can be used instead. The fundamental difference in doing so, is that the interval approach induces a simple correlation structure between marker effects, something to be expected if a QTL is linked to markers in a region of the genome; using the methods of this paper, WGAIM applied to markers does not incorporate correlation of marker effects.

The approach presented here allows for a QTL in every interval or at each marker on the linkage map; we will call the position the location, be it in the interval or at a marker. The QTL sizes are assumed to be a random effect and a test of significance of that random effect is conducted to determine if a putative QTL can be selected. If significant, an outlier detection technique is used to determine the most likely location for the putative QTL, otherwise the process terminates. If there is a selection, the selected putative QTL is added to the model as a random effect and the process continues. Thus, a forward selection process is conducted. At the end of the selection process the QTL sizes and their strength are assessed using methods appropriate for QTL taken as random effects, in particular the LOGP score, namely $-\log_{10} P$ where P is a probability reflecting the strength of the QTL, and percentage of genetic variance for each QTL can be determined and reported. The details of these steps are presented as follows.

Suppose, we have c linkage groups (or chromosomes) and r_k markers on linkage group k, $k = 1, 2, \ldots, c$. The genetic model for line i, $i = 1, 2, \ldots, n_g$, in the multi-parent (or founder) situation is given by

$$u_{gi} = \sum_{k=1}^{c} \sum_{j=1}^{r_k^*} \mathbf{q}_{ikj}^\mathrm{T} \mathbf{a}_{kj} + u_{pi} \tag{2}$$

where $r_k^* = r_k - 1$ if intervals are used and $r_k^* = r_k$ if markers are used in the analysis. Model (2) allows for a QTL at each location (interval or marker) on the linkage map. If there are n_f founders, $\mathbf{q}_{ikj}$ is the $n_\mathrm{f} \times 1$ vector which indicates the founder allele for line i for a potential QTL at location j on linkage group k. The elements of $\mathbf{q}_{ikj}$ are q_{ikjl} for $l = 1, 2, \ldots, n_\mathrm{f}$ and one of these is 1 and the rest are zero. Thus, if $\mathbf{1}_{n_\mathrm{f}}$ is a vector of ones,

$$\mathbf{q}_{ikj}^\mathrm{T} \mathbf{1}_{n_\mathrm{f}} = \sum_{l=1}^{n_\mathrm{f}} q_{ikjl} = 1$$

The term $\mathbf{a}_{kj}$ is the vector of QTL sizes for location j on linkage group k, one for each founder, reflecting the

possible differential expression of a QTL for each founder. Note that most of the vectors of sizes will be zero. Our aim is to determine which locations have non-zero QTL sizes.

The last component of (2) is u_{pi} which is an independent polygenic or residual genetic effect, and typically $u_{pi} \sim N(0, \sigma_p^2)$ although pedigree information can also be included, see Oakey et al. (2006). This term allows for possibly a large number of small QTL that cannot be detected individually.

Of course at the QTL the founder allele is unknown and hence which element of $\mathbf{q}_{ijk}$ is 1 is unknown. To overcome this problem, the regression approach is used (Haley and Knott 1992; Martinez and Curnow 1992) for analysis. Thus, $\mathbf{q}_{ijk}$ is replaced by its expected value. Let $\mathbf{p}_{ikj}$ be the $n_\mathrm{f} \times 1$ vector of founder probabilities for a potential QTL for line i at location j on linkage group k. That is, the lth element p_{ikjl} is the probability that the QTL allele is from founder l. Determining these probabilities is discussed below. First, note that

$$\mathbf{p}_{ikj}^\mathrm{T}\mathbf{1}_{n_\mathrm{f}} = \sum_{l=1}^{n_\mathrm{f}} p_{ikjl} = 1$$

just like $\mathbf{q}_{ikj}$. Second, $\mathbf{q}_{ikj}$ has a multinomial distribution with index number equal to 1 (the "sample" size) and probabilities $\mathbf{p}_{ikj}$, that is $\mathbf{q}_{ikj} \sim Mn(1, \mathbf{p}_{ikj})$. Then

$$\mathrm{E}\left(\mathbf{q}_{ikj}\right) = \mathbf{p}_{ikj}$$

and if $\mathrm{diag}\left(\mathbf{p}_{ikj}\right)$ is a diagonal matrix with elements given by the vector $\mathbf{p}_{ikj}$,

$$\mathrm{var}\left(\mathbf{q}_{ikj}\right) = \mathrm{diag}\left(\mathbf{p}_{ikj}\right) - \mathbf{p}_{ikj}\mathbf{p}_{ikj}^\mathrm{T} \tag{3}$$

a result that will be used when finding percentage of variance explained by a QTL.

In accordance with the regression method, we consider the model

$$u_{gi} = \sum_{k=1}^{c}\sum_{j=1}^{r_k^*} \mathbf{p}_{ikj}^\mathrm{T}\mathbf{a}_{kj} + u_{pi}$$

or in vector-matrix form

$$\mathbf{u}_g = \mathbf{P}\mathbf{a} + \mathbf{u}_p \tag{4}$$

In addition, we assume $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2\mathbf{I})$; note that $\mathbf{a}$ has $(r - c)n_\mathrm{f}$ elements ($r = \sum_{k=1}^{c} r_k$) if intervals are used and rn_f elements if markers are used.

Founder probabilities

To proceed with any analysis requires $\mathbf{P}$ and hence $\mathbf{p}_{ikj}$. Two approaches for determining $\mathbf{p}_{ikj}$ are presented. The first involves three-point probabilities and extends the results of Broman (2005); these results have been independently derived by the authors of the R (R Development Core Team 2013) package mpMap (Huang and George 2011). The second approach involves the use of a HMM as presented by Broman (2006) and implemented in the qtl package (Broman et al. 2003, 2012) in R.

The assumptions underlying these developments are that we have inbred lines through selfing and we, therefore, focus on haplotypes of homozygous lines.

Three-point probabilities: interval-based approach

We begin with some notation. First, we consider a single line or individual and a single (generic) interval flanked by M_L, the left-hand marker locus and M_R, the right-hand marker locus. Let Q denote the potential QTL locus in the interval defined by M_L and M_R. Let the unobserved or hidden founder alleles be denoted by $A_1, A_2, \ldots, A_{n_\mathrm{f}}$ where the marker or QTL locus index is suppressed. The observed marker score for M_L is denoted by M_LS with a similar definition for M_RS. The unobserved allele at a marker or QTL has an additional index A, so that, for example, Q_A denotes the unobserved allele at the QTL locus Q, and M_LA and M_RA denote the unobserved allele at the left- and right-hand marker locus, respectively.

Because Q_A is unknown, we use an interval analysis approach and condition on the observed marker phenotypes or scores for the flanking markers or loci, that is, M_LS and M_RS. The probabilities required for the QTL analysis involve the conditional probability that the QTL allele is one of the founder alleles, A_i, given the marker scores of the two flanking markers; that is $\Pr(Q_A = A_i | M_\mathrm{LS} = m_1 \cap M_\mathrm{RS} = m_2)$. Let F_L and F_R denote the set of founder alleles indices (the index for A_l is l) that are consistent with the marker scores m_1 and m_2, respectively. Then, the probability can be expressed as

$$\begin{aligned}
&\Pr(Q_A = A_i | M_\mathrm{LS} = m_1 \cap M_\mathrm{RS} = m_2) \\
&= \frac{\Pr(Q_A = A_i \cap M_\mathrm{LS} = m_1 \cap M_\mathrm{RS} = m_2)}{\Pr(M_\mathrm{LS} = m_1 \cap M_\mathrm{RS} = m_2)} \\
&= \frac{\sum_{j\in F_\mathrm{L}}\sum_{k\in F_\mathrm{R}} \Pr(M_\mathrm{LA} = A_j \cap Q_A = A_i \cap M_\mathrm{RA} = A_k)}{\sum_{j\in F_\mathrm{L}}\sum_{k\in F_\mathrm{R}} \Pr(M_\mathrm{LA} = A_j \cap M_\mathrm{RA} = A_k)} \\
&= \frac{\sum_{j\in F_\mathrm{L}}\sum_{k\in F_\mathrm{R}} \Pr(M_\mathrm{LA} = A_j \cap Q_A = A_i \cap M_\mathrm{RA} = A_k)}{\sum_{i=1}^{n_\mathrm{f}}\sum_{j\in F_\mathrm{L}}\sum_{k\in F_\mathrm{R}} \Pr(M_\mathrm{LA} = A_j \cap Q_A = A_i \cap M_\mathrm{RA} = A_k)}
\end{aligned} \tag{5}$$

To find these probabilities, we require the three-point haplotype probabilities $\Pr(M_\mathrm{LA} = A_j \cap Q_A = A_i \cap M_\mathrm{RA} = A_k)$.

Let the recombination fraction between M_L and M_R be denoted by r_{13}, between M_L and Q by r_{12} and between Q and M_R by r_{23}. Broman (2005) provides expressions for three-point haplotype probabilities when $r_{12} = r_{23}$.

Following Broman (2005), the four- and eight-way cross three-point probabilities are built up in stages. Thus, the two-way two-point probabilities given by Haldane and Waddington (1931) and noted by Broman (2005) are used to find the two-way three-point probabilities. Using the notation of Broman (2005), we begin by defining a prototype for haplotypes that share the same probability. Thus, for the two-way cross, the three-point prototype $A_1A_1A_1$ corresponds to haplotypes in which all alleles originate from the same founder; if the founder alleles are A_1 and A_2 , these are the haplotypes $A_1A_1A_1$ and $A_2A_2A_2$ and the prototype is simply the first possibility. Thus, the count of the number of haplotypes that share the same probability for prototype $A_1A_1A_1$ in Table 1 is 2. For the two-way cross, there are four prototype probabilities (one more than the case $r_{12} = r_{23}$). and all have count 2 because there are two possible haplotypes for each prototype.

The probabilities for the four-way and eight-way crosses are also given in Table 1 and follow using the arguments of Broman (2005). The details in that table depend on the structure of the MAGIC population. For example, the prototypes, the haplotypes in the prototype, and their probabilities for the four-way cross depend on the structure presented in Fig. 1, where founders 1 and 2 are crossed, founders 3 and 4 are crossed, and their respective F1 progeny is crossed before the resulting lines are selfed. Thus, alleles A_1 and A_2 form an equivalence class due to the initial crossing as do alleles A_3 and A_4. This structure determines the number of haplotypes in a prototype. For example, there are four haplotypes for prototype $A_1A_2A_2$ because A_1 is one of four possible founder alleles and once chosen the other allele must be the other allele in the same equivalence class. For prototype $A_1A_3A_4$, the first allele can be chosen in four ways, and the next allele must come from the other equivalence class and may be chosen in two ways; the last allele is then the remaining allele in the other equivalence class. This results in eight possible haplotypes as given in Table 1. The same argument follows for the eight-way cross, where it has been assumed pairs of founders are crossed initially to form four families. The resulting F1 lines are crossed between pairs of two families (families 1 and 2 or families 3 and 4 only) to form F2 lines and then the resulting lines crossed between this second level of family structure to form F3 lines before selfing. Thus, for prototype $A_1A_3A_5$ we have 8 times 2 times 4 possible haplotypes, corresponding to 8 possible founders for the first allele, 2 possible alleles in the F2 structure, and lastly four possible alleles in the F2 structure in the crossing of families 3 and 4. All the counts for the number of haplotypes in Table 1 are found in a similar manner.

The probabilities in Table 1 form the basis of one possible approach to calculations to be described below for use in QTL analysis.

Table 1 Three-point probabilities for two-way, four-way and eight-way crosses with selfing

Prototype	Count	Probability
Two-way cross		
$A_1A_1A_1$	2	$x_1 = \frac{1}{2}\left\{\frac{1}{2(1+r_{12})} - \frac{r_{13}}{1+2r_{13}} + \frac{1}{2(1+2r_{23})}\right\}$
$A_1A_1A_2$	2	$x_2 = \frac{1}{2}\left\{\frac{1}{2(1+r_{12})} - \frac{1}{2(1+2r_{13})} + \frac{r_{23}}{1+2r_{23}}\right\}$
$A_1A_2A_2$	2	$x_3 = \frac{1}{2}\left\{\frac{r_{12}}{1+r_{12}} + \frac{r_{13}}{1+2r_{13}} - \frac{r_{23}}{1+2r_{23}}\right\}$
$A_1A_2A_1$	2	$x_4 = \frac{1}{2}\left\{\frac{r_{12}}{1+r_{12}} - \frac{r_{13}}{1+2r_{13}} + \frac{r_{23}}{1+2r_{23}}\right\}$
Four-way cross		
$A_1A_1A_1$	4	$a_1 = x_1\frac{(1-r_{12})(1-r_{23})}{2}$
$A_1A_1A_2$	4	$a_2 = x_1\frac{(1-r_{12})r_{23}}{2}$
$A_1A_2A_2$	4	$a_3 = x_1\frac{r_{12}(1-r_{23})}{2}$
$A_1A_2A_1$	4	$a_4 = x_1\frac{r_{12}r_{23}}{2}$
$A_1A_1A_3$	8	$a_5 = x_2\frac{(1-r_{12})}{4}$
$A_1A_3A_3$	8	$a_6 = x_3\frac{(1-r_{23})}{4}$
$A_1A_3A_1$	8	$a_7 = x_4\frac{(1-r_{13})}{4}$
$A_1A_2A_3$	8	$a_8 = x_2\frac{r_{12}}{4}$
$A_1A_3A_4$	8	$a_9 = x_3\frac{r_{23}}{4}$
$A_1A_3A_2$	8	$a_{10} = x_4\frac{r_{13}}{4}$
Eight-way cross		
$A_1A_1A_1$	8	$b_1 = a_1\frac{(1-r_{12})(1-r_{23})}{2}$
$A_1A_1A_2$	8	$b_2 = a_1\frac{(1-r_{12})r_{23}}{2}$
$A_1A_2A_2$	8	$b_3 = a_1\frac{r_{12}(1-r_{23})}{2}$
$A_1A_2A_1$	8	$b_4 = a_1\frac{r_{12}r_{23}}{2}$
$A_1A_1A_3$	16	$b_5 = a_2\frac{(1-r_{12})}{4}$
$A_1A_3A_3$	16	$b_6 = a_3\frac{(1-r_{23})}{4}$
$A_1A_3A_1$	16	$b_7 = a_4\frac{(1-r_{13})}{4}$
$A_1A_1A_5$	32	$b_8 = a_5\frac{(1-r_{12})}{4}$
$A_1A_5A_5$	32	$b_9 = a_6\frac{(1-r_{23})}{4}$
$A_1A_5A_1$	32	$b_{10} = a_7\frac{(1-r_{13})}{4}$
$A_1A_2A_3$	16	$b_{11} = a_2\frac{r_{12}}{4}$
$A_3A_1A_2$	16	$b_{12} = a_3\frac{r_{23}}{4}$
$A_1A_3A_2$	16	$b_{13} = a_4\frac{r_{13}}{4}$
$A_1A_2A_5$	32	$b_{14} = a_5\frac{r_{12}}{4}$
$A_5A_1A_2$	32	$b_{15} = a_6\frac{r_{23}}{4}$
$A_1A_5A_2$	32	$b_{16} = a_7\frac{r_{13}}{4}$
$A_1A_3A_5$	64	$b_{17} = a_8\frac{1}{8}$
$A_1A_5A_7$	64	$b_{18} = a_9\frac{1}{8}$
$A_1A_5A_3$	64	$b_{19} = a_{10}\frac{1}{8}$

Count indicates the number of haplotypes in the prototype. The four-way cross probabilities depend on the two-way cross probabilities and the eight-way cross probabilities depend on the four-way cross probabilities

HMM probabilities: interval-based analysis

An alternative approach for calculating haplotype probabilities is to use HMM as outlined by Broman (2006). The

fundamental difference in using a HMM is that the whole linkage group is used in calculating probabilities at a particular location and hence the method is multi-point rather than three-point. We shall call the resulting probabilities HMM probabilities for the remainder of the paper. Suppose $M_1, M_2, \ldots, M_{r_k}$ are the markers on chromosome k and Q is a putative QTL on that chromosome. Let M_{lS} and M_{lA} be the marker scores and genotypes, respectively, for marker locus M_{l} for an individual line, and Q_S and Q_A be the corresponding quantities for the QTL; Q_S is missing. Similar to the three-point calculation, we require

$$\Pr(Q_A = A_i | M_{1S} = m_{1s} \cap \cdots \cap M_{r_k S} = m_{r_k s}). \tag{6}$$

Broman (2006) notes that the M_{lA} form a (hidden) Markov Chain under the assumption of no interference. The calculation of (6) can then proceed efficiently using the so-called forward–backward equations for HMM, rather than the analog of (5); for details see Broman (2006). Thus multi-point probabilities can be found for any location on the chromosome for each individual line using the HMM approach.

Average interval mapping

The three-point and HMM probabilities depend on the unknown r_{12} and r_{23} and the known r_{13}. In fact using Trow's formula (Trow 1913), this reduces to one unknown recombination fraction r_{12}. Thus, for each potential QTL, and our model (4) allows for one in every interval, we have both a size and location that require estimation. To determine both turns out to be prohibitive and following Verbyla et al. (2007) we eliminate all the r_{12} as follows. Using Haldane's distance

$$d_{12} = -\frac{1}{2}\log(1 - 2r_{12})$$

it is assumed the distance d_{12} is uniformly distributed on the full interval $[0, d_{13}]$. Then, each of the four-way or eight-way probabilities is integrated over this prior distribution. The algebra is tedious for the three-point probabilities and numerical integration is used (Piessens et al. 1983). For the HMM case, numerical integration using the trapezoidal rule is used and requires a grid of distances or recombination frequencies across each interval and hence a decision on their spacing. This does not appear to be too crucial. In practice, a grid with a step of 0.1 cM is used uniformly across the map, but this will need to be reduced for maps where some distances are less than 0.1 cM between markers. In addition, for narrow intervals, the probabilities will not vary greatly across the interval and so the number of evaluations need not be as big as for wider intervals where probabilities may change across the interval; thus there does not appear to be any need to vary the width with varying density of markers. To evaluate the stability of the average probabilities, stepsizes of 0.1, 0.05 and 0.01 cM were used to find average probabilities for the linkage map described in the Methods section. Comparing the 0.1 cM to the 0.05 cM stepsize, the maximum difference between the corresponding probabilities was 0.00027 (found for an interval of total width 0.5 cM on the map, with smallest width 0.388 cM) with a relative error of 0.047 %. If a step of 0.01 cM was compared to a stepsize of 0.1 cM, the maximum difference was 0.00049 (for the same interval as with stepsize 0.05 cM) with a relative error of 0.085 %. The conclusion from this small examination was that using 0.1 cM was sufficient no matter how wide the interval.

The average probabilities need only be calculated once for each linkage map. The revised genetic model is given by

$$\mathbf{u}_g = \mathbf{P}_A \mathbf{a} + \mathbf{u}_p \tag{7}$$

where $\mathbf{P}_A$ is the $n_g \times (r - c)n_{\mathrm{f}}$ matrix of average probabilities.

Haplotype probabilities: marker-based analysis

The development above is based on interval mapping ideas and potential QTL in the interval. Alternatively, a marker-based analysis could be carried out, where it is assumed putative QTL are at marker loci. The three-point probability calculations outlined above can be carried out at a marker locus, but there is additional information on the observed marker scores at the locus that can be used. We consider marker locus l on chromosome k as discussed in the HMM subsection above and consider the QTL allele indicator $Q_A = M_{lA}$; thus the potential QTL is at marker l and it is sufficient to consider M_{lA}. For locus l, let F_l denote the set of unobserved marker phenotype indices consistent with the observed marker score M_{lS}. Then, the probability equivalent to the three-point probabilities (5) is for $i \in F_l$ (so that the allele A_i is consistent with the marker score $M_{lS} = m_2$),

$$\begin{aligned}
&\Pr(M_{lA} = A_i | M_{l-1,S} = m_1 \cap M_{lS} = m_2 \cap M_{l+1,S} = m_3) \\
&= \frac{\Pr(M_{lA} = A_i \cap M_{l-1,S} = m_1 \cap M_{lS} = m_2 \cap M_{l+1,S} = m_3)}{\Pr(M_{l-1,S} = m_1 \cap M_{lS} = m_2 \cap M_{l+1,S} = m_3)} \\
&= \frac{\sum_{j \in F_{l-1}} \sum_{k \in F_{l+1}} \Pr(M_{l-1,A} = A_j \cap M_{lA} = A_i \cap M_{l+1,A} = A_k)}{\sum_{j \in F_{l-1}} \sum_{k \in F_{l+1}} \Pr(M_{l-1,A} = A_j \cap M_{l+1,A} = A_k)} \\
&= \frac{\sum_{j \in F_{l-1}} \sum_{k \in F_{l+1}} \Pr(M_{l-1,A} = A_j \cap M_{lA} = A_i \cap M_{l+1,A} = A_k)}{\sum_{i \in F_l} \sum_{j \in F_{l-1}} \sum_{k \in F_{l+1}} \Pr(M_{l-1,A} = A_j \cap M_{lA} = A_i \cap M_{l+1,A} = A_k)}
\end{aligned} \tag{8}$$

while for $i \notin F_l$ the probability is zero.

For the HMM based approach, the appropriate conditional probability involves the full set of observed marker scores. The HMM approach thus requires evaluation of

$$\Pr(M_{lA} = A_i | M_{1S} = m_{1s} \cap \cdots \cap M_{lS} = m_s \cap \cdots \cap M_{r_k S} = m_{r_k s})$$

In both these cases, the matrix $\mathbf{P}_A$ is now of size $n_g \times rn_\mathrm{f}$ and contains the probabilities for each marker and each individual line.

Test for a putative QTL

To determine if a putative QTL exists, two models are fitted using (1). These are

$$\mathbf{u}_g = \mathbf{u}_p \tag{9}$$

and

$$\mathbf{u}_g = \mathbf{P}_A \mathbf{a} + \mathbf{u}_p \tag{10}$$

where $\mathbf{a} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{I}_{(r-c)n_\mathrm{f}})$. Note that with MAGIC populations the dimension $(r-c)n_\mathrm{f}$ can potentially be very large and lead to the $p > n$ problem. Verbyla et al. (2012) propose an approach that reduces the dimension for model fitting and that same approach can be used for MPWGAIM. This reduces the dimension for model fitting to n_g, in the same manner as for the bi-parental situation; see the Appendix for details.

As for bi-parental crosses, a test of $H_0 : \sigma_a^2 = 0$ is used to establish if a putative QTL can be selected. A residual likelihood ratio test is conducted to test model (9) against (10). The test is non-standard and the null distribution is a mixture of a point probability of 0.5 at zero and half of a chi-square on one degree of freedom (Stram and Lee 1994). If H_0 is rejected, there is sufficient potential QTL size variance to warrant the selection of a putative QTL. If H_0 is retained, there is insufficient potential QTL size variance and hence the selection process is terminated.

Selection of a putative QTL

If H_0 is rejected, an outlier statistic is used to select the most likely marker for the putative QTL, just like the bi-parental situation (Verbyla et al. 2007, 2012). For interval (or marker) j on linkage group k, the alternative outlier model is given by

$$\mathbf{u}_g = \mathbf{P}_A(\mathbf{a} + \mathbf{D}_{kj}\delta_{kj}) + \mathbf{u}_p \tag{11}$$

where $\mathbf{D}_{kj}$ is an $(r-c)n_\mathrm{f} \times n_\mathrm{f}$ matrix with a one for each founder in interval j on linkage group k and zero elsewhere; if markers are used in the analysis $\mathbf{D}_{kj}$ is an $rn_\mathrm{f} \times n_\mathrm{f}$ matrix. This model inflates the additive effects by δ_{kj} which is assumed to follow a $N(\mathbf{0}, \sigma_{akj}^2 \mathbf{I}_{n_\mathrm{f}})$. The outlier statistic is based on the score test (Cox and Hinkley 1974) of $H_0 : \sigma_{akj}^2 = 0$. If $\tilde{a}_{jkl}$ is the best linear unbiased predictor (BLUP) of the size of a potential QTL effect for founder l in interval j on linkage group k with variance $\mathrm{var}(\tilde{a}_{jkl})$, the outlier statistic is given by

$$t_{jk}^2 = \frac{\sum_{l=1}^{n_\mathrm{f}} \tilde{a}_{jkl}^2}{\sum_{l=1}^{n_\mathrm{f}} \mathrm{var}(\tilde{a}_{jkl})} \tag{12}$$

so that the effects are summed over the founders. The statistic incorporates all the sizes for founder alleles at potential QTL.

The outlier statistic is calculated for each interval (or marker) and the interval (or marker) having the largest statistic is deemed to be the putative QTL interval (or marker).

The selected putative QTL is added as a random effect. Thus, the two models to be compared for possible selection of another putative QTL are, extending (9) and (10),

$$\mathbf{u}_g = \mathbf{P}_{A1}\mathbf{a}_1 + \mathbf{u}_p \tag{13}$$

and

$$\mathbf{u}_g = \mathbf{P}_{A1}\mathbf{a}_1 + \mathbf{P}_{A,-1}\mathbf{a}_{-1} + \mathbf{u}_p \tag{14}$$

where $\mathbf{P}_{A1}$ is a matrix of probabilities (one for each founder for each line) for the first putative QTL (interval or marker) and $\mathbf{P}_{A,-1}$ is the matrix of founder probabilities for each line omitting the selected interval or marker probabilities. Note that the vector of the putative QTL size effects is assumed $\mathbf{a}_1 \sim N(\mathbf{0}, \sigma_{a1}^2 \mathbf{I}_{n_\mathrm{f}})$ and $\mathbf{a}_{-1} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{I}_{(r-c-1)n_\mathrm{f}})$.

Selection of additional QTL

The above process is continued in a forward selection manner. Thus, after s selections of putative QTL the two models to be compared for possible selection of another putative QTL are

$$\mathbf{u}_g = \sum_{j=1}^{s} \mathbf{P}_{Aj}\mathbf{a}_j + \mathbf{u}_p \tag{15}$$

$$\mathbf{u}_g = \sum_{j=1}^{s} \mathbf{P}_{Aj}\mathbf{a}_j + \mathbf{P}_{A,-s}\mathbf{a}_{-s} + \mathbf{u}_p \tag{16}$$

Note that $\mathbf{a}_j \sim N(\mathbf{0}, \sigma_{aj}^2 \mathbf{I}_{n_\mathrm{f}})$ so that each putative QTL is assumed to come from its own distribution to reduce bias in the size of estimated effects (as in the bi-parental case).

Final assessment of significance of QTL

To provide the level of significance of each QTL, the approach outlined by Verbyla et al. (2012) is followed. Under the normality assumptions of the linear mixed model, the approach involves the conditional distribution of the QTL size vector given the data, namely,

$$\mathbf{a}_j | \mathbf{y}_2 \sim N(\tilde{\mathbf{a}}_j, \Sigma_{\mathrm{PEV},j})$$

where $\mathbf{y}_2$ is the component of the data free of fixed effects (Verbyla 1990). The mean of this conditional distribution

is the BLUP of $\mathbf{a}_j$, that is the estimated size of the QTL $\tilde{\mathbf{a}}_j$, and $\Sigma_{\mathrm{PEV},j}$ is the prediction error variance matrix (PEV) of $\mathbf{a}_j$. If $\Sigma^{-}_{\mathrm{PEV},j}$ is a generalized inverse of $\Sigma_{\mathrm{PEV},j}$, the distance measure

$$d_j^2 = (\mathbf{a}_j - \tilde{\mathbf{a}}_j)^{\mathrm{T}} \Sigma^{-}_{\mathrm{PEV},j} (\mathbf{a}_j - \tilde{\mathbf{a}}_j)$$

has a chi-squared distribution with $n_{\mathrm{f}} - 1$ degrees of freedom and if

$$c_j^2 = \tilde{\mathbf{a}}_j^{\mathrm{T}} \Sigma^{-}_{\mathrm{PEV},j} \tilde{\mathbf{a}}_j$$

a measure of the strength of the putative QTL is given by

$$p_j = \Pr(d_j^2 > c_j^2).$$

Notice that this probability can be calculated for the QTL as a whole, that is for all founders together, and for individual founders. This enables "significance" of QTL effects both at the overall and at the founder level to be reported. In this paper, the LOGP score, defined by

$$\mathrm{LOGP}_j = -\log_{10}(p_j)$$

is used to give an indication of the strength of the putative QTL effects.

Percentage of genetic variance for a QTL

Lastly, the percentage of genetic variance accounted for by each putative QTL is of interest. This can be determined approximately as follows.

Once all putative QTL have been found, suppose s is the number of putative QTL. Rather than using (16), consider the genetic effect for line i in terms of the indicator variable $\mathbf{q}_{ij}$ for QTL j,

$$u_{gi} = \sum_{j=1}^{s} \mathbf{q}_{ij}^{\mathrm{T}} \mathbf{a}_j + \mathbf{p}_{Ai,-s}^{\mathrm{T}} \mathbf{a}_{-s} + u_{pi}$$

where $\mathbf{p}_{Ai,-s}^{\mathrm{T}}$ is the ith row of $\mathbf{P}_{A,-s}$. Then, the variance of u_{gi} is given by

$$\mathrm{var}(u_{gi}) = \sum_{j=1}^{s} \mathbf{a}_j^{\mathrm{T}} \mathrm{var}(\mathbf{q}_{ij}) \mathbf{a}_j + \sigma_a^2 \mathbf{p}_{Ai,-s}^{\mathrm{T}} \mathbf{p}_{Ai,-s} + \sigma_p^2$$

where $\mathrm{var}(\mathbf{q}_{ij})$ is given by (3). We have ignored linkage because it is very difficult to allow for linkage in the calculations and the calculations are approximate in nature.

Each line has its own (different) variance because the variance depends on $\mathbf{p}_{ij}$ and $\mathbf{p}_{Ai,-s}$ and these can be different for each i. To make progress, it is necessary to define an "average" line with ("average") QTL indicator $\bar{\mathbf{q}}_j$ so that an overall representative variance can be found. To do so, consider the average founder probabilities, $\bar{\mathbf{p}}_j$, defined as

$$\bar{\mathbf{p}}_j = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{p}_{ij}.$$

Then, we define

$$\mathrm{var}(\bar{\mathbf{q}}_j) = \mathrm{diag}(\bar{\mathbf{p}}_j) - \bar{\mathbf{p}}_j \bar{\mathbf{p}}_j^{\mathrm{T}}.$$

In the same way, for the term involving the intervals (or markers) not selected, we simply average the founder probabilities over all the lines for the non-selected intervals (or markers), $\bar{\mathbf{p}}_{A,-s}$ say. Then, the total variance of an "average" line effect, u_g^* is

$$\mathrm{var}\left(u_g^*\right) = \sum_{j=1}^{s} \mathbf{a}_j^{\mathrm{T}} \mathrm{var}(\bar{\mathbf{q}}_j) \mathbf{a}_j + \sigma_a^2 \bar{\mathbf{p}}_{A,-s}^{\mathrm{T}} \bar{\mathbf{p}}_{A,-s} + \sigma_p^2. \quad (17)$$

Equation (17) specifies an approximate total genetic variance. The percentage genetic variance attributed to the jth QTL is then

$$\mathrm{PV}_j = 100 \frac{\mathbf{a}_j^{\mathrm{T}} \mathrm{var}(\bar{\mathbf{q}}_j) \mathbf{a}_j}{\mathrm{var}\left(u_g^*\right)}. \quad (18)$$

In practice the unknown sizes $\mathbf{a}_j$ and variance components σ_a^2 and σ_p^2 are replaced by their estimates.

The definition of the "average" line is somewhat arbitrary, but it was chosen for simplicity of both concept and calculation. The true underlying percentage genetic variance will vary from line to line in the same way and hence suffers from the same lack of uniqueness. Given the definition used, the accuracy of the percentage variance will depend on the accuracy of the estimated QTL sizes to the true QTL sizes, and also on the accuracy of the estimated polygenic variance. This is because the averaging of the probabilities would be necessary to determine a measure of the percentage variance even if the true sizes were known. Thus assessing the accuracy of percentage genetic variance in the simulation study to be discussed below amounts to investigating the accuracy of the estimated QTL sizes.

Computation

The MPWGAIM approach has been implemented in the package mpwgaim in the R environment (R Development Core Team 2013) and is available from the authors. This package has dependencies on the R packages wgaim (Taylor et al. 2011), asreml (Butler et al. 2011), qtl (Broman et al. 2003, 2012) and mpMap (Huang and George 2011).

Table 2 Summary of the linkage map for a four-way wheat MAGIC population

Chromosome	Number of markers	Length
1A	465	292
1B	345	326
1D	112	145
2A	364	307
2B	440	436
2D	123	172
3A	350	340
3B	217	258
3D	31	153
4A	357	363
4B	150	158
4D	25	106
5A	383	328
5B	560	379
5D	67	238
6A	444	253
6B	425	318
6D	49	229
7A	467	355
7B	281	312
7D	36	259
Unlinked1	37	14
Unlinked2	13	35
Unlinked3	22	12

In particular, it is the power of asreml that allows the complex models to be fitted.

Materials

Linkage map

A linkage map for a four-way cross in wheat was developed in CSIRO, Australia. An early version of the map is discussed in Huang et al. (2012), where DArTs and microsatellites were used, while Cavanagh et al. (2013) present a map based on SNPs. The map used in this paper is based on a combination of both sets of markers. The map was developed using the mpMap package as well as manual intervention and consists of 5,763 markers, mostly SNPs, with 755 DArTs and 39 multi-allelic microsatellites. These were grouped into the 21 chromosomes of wheat plus three additional linkage groups of markers. The full map is available in the supplementary material. A summary of the map is presented in Table 2 where the number of markers is presented for each chromosome together with the length of each chromosome in cM. We expect longer chromosomes

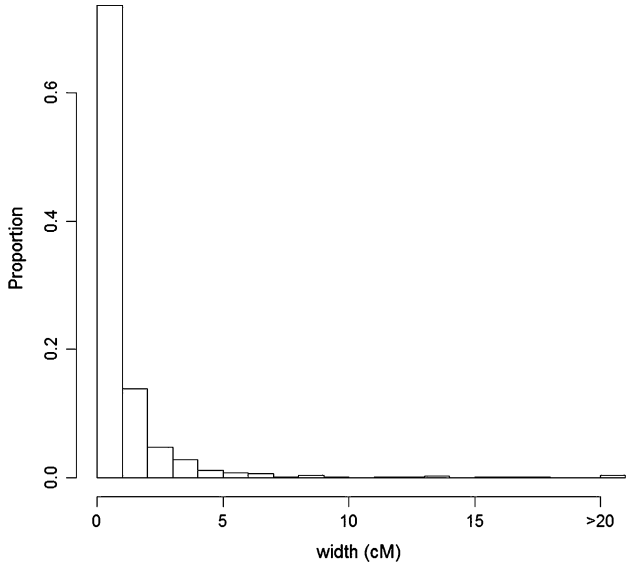

Fig. 2 Histogram of interval distances (in cM) between markers for the four-way wheat MAGIC map. Some intervals were wider than 20 cM and these have been grouped into a class at 21 cM

because of increased recombination events in MAGIC designs. Chromosome 2B has a translocation and is the longest of the chromosomes because of an introgression of a wild relative. The total map length was 5,788 cM. Note that the D-genome was most sparse.

A histogram of the interval widths across the genome is presented in Fig. 2 (there were some large interval widths that were set at 21 to enable a useful histogram to be presented). The widths are predominantly under 5 cM.

Many of the markers were co-located at the same position (see the supplementary material) on the map and are, therefore, not useful in a QTL analysis. Markers were therefore culled prior to QTL analysis to ensure non-zero recombination fractions between the remaining markers. Thus, the final map for QTL analysis consisted of 3,230 markers (including 620 DArTs and 37 microsatellites).

Simulation study

To assess the performance of MPWGAIM, a simulation study was conducted. The linkage map used was based on a subset of the four-way wheat MAGIC map discussed above and presented in the supplementary material. The subset was created by removing the D-genome and all genomes for chromosomes 2 and 7 (and the small additional linkage groups Unlinked1, Unlinked2 and Unlinked3). The main reason for removing chromosomes was to avoid excessive computations in the simulation, while retaining a reasonable number of chromosomes. The D-genome was removed because the number of markers was small, while chromosome 2 contains a translocation and in the simulation

Table 3 Specification for the power simulation

Seven QTL on chromosomes 1 to 5 with sizes as specified for the founder alleles

QTL	Chromosome	Position (cM)	Founder			
			1	2	3	4
1	1	141	0.354	−0.354	−0.354	0.354
2	2	160	0.354	−0.118	−0.118	−0.118
3	3	174	0.354	−0.354	0.000	0.000
4	4	187	0.354	−0.354	−0.354	0.354
5	4	207	0.354	−0.354	−0.354	0.354
6	5	77	0.354	−0.354	−0.354	0.354
7	5	97	−0.354	0.354	0.354	−0.354

study we wished to avoid complications. This left ten chromosomes for use in the simulation study. There were 2,041 markers across the ten chromosomes remaining; see Table 2.

Two simulation studies were conducted; the genetic data were generated using the mpMap package (Huang and George 2011) in R (R Development Core Team 2013) using the reduced linkage map and the available pedigree for the 1,088 lines in the MAGIC population. First, a simulation with no QTL was conducted to examine the Type I error rate (at least one QTL found when none exist) and the false discovery rate (number of QTL found when none exist). The model generating the data for analysis for each simulation was based on the simple model

$$y_{ij} = \mu + u_{gi} + e_{ij}$$

for $i = 1, 2, \ldots, 1{,}088$ and $j = 1, 2$. Thus there are two replicates of 1,088 genetic lines. The constant $\mu = 10$, while $u_{gi} \sim N(0, \sigma_g^2)$ are assumed independent over i with $\sigma_g^2 = 0.5$. This component of genetic variation is therefore assumed to arise from the so-called infinitesimal model. The residuals e_{ij} were simulated from independent standard normal variables. Thus, the heritability was 0.5. Five hundred simulations were conducted.

Second, a power study was carried out when seven QTL were present. Single QTL were included on chromosomes 1, 2 and 3 while two QTL were included on chromosomes 4 and 5. No QTL were simulated for chromosomes 6 to 10. The sizes and locations of the QTL are given in Table 3. Again 500 simulations were conducted and the model was

$$y_{ij} = \mu + \sum_{j=1}^{7} \mathbf{q}_{ij}^{\mathrm{T}} \mathbf{a}_j + u_{gi} + e_{ij}$$

where $\mathbf{q}_{ij}$ is the indicator variable for the founder allele for line i for QTL j, with sizes $\mathbf{a}_j$ given in Table 3. Note that the heritability is either 73, 72 or 70 % depending on whether lines have QTL alleles from founders 2 to 4 for the QTL on chromosome 2 or alleles from founders 3 and 4 for the QTL on chromosome 3.

To assess the performance of MPWGAIM, a comparison with HAPPY (Mott et al. 2000) was also conducted for the power study. This involved running the publicly available software for each simulation, available at http://mus.well.ox.ac.uk/magic/. First, HAPPY founder probabilities were determined using the software. These probabilities are different from the three-point probabilities and the HMM probabilities of Broman (2006). The number of effective generations was set to 4. To determine the genome-wide threshold (using a p value of 0.05) for determining putative QTL for each simulation, 1,000 permutations of the data were carried out, again as provided in the software. In addition, as inspection of 500 plots was not ideal (but the way such genome scans are normally interpreted), the output provided was slightly modified allowing for islands (stretches of chromosome above the genome-wide threshold) where QTL were identified to be merged if there was one marker below the threshold that separated the islands; the largest peak was still declared the QTL for this composite island. Lastly, the summary function was modified to allow detection of QTL in coupling as the existing software did not provide for such determination. This involved searching for multiple peaks in an island.

In the power study, a QTL was deemed to be detected if either the interval or a chosen marker was 5 cM either side of the true QTL. This is more stringent than Broman and Speed (2002) and Verbyla et al. (2007, 2012). For HAPPY, a QTL was deemed to be found if the selected peak was in an interval that contained the point 5 cM either side of the true QTL position.

MAGIC experimental data

A wheat MAGIC population was grown in several experiments conducted by Food Futures National Research Flagship, CSIRO, Australia. One such field trial was conducted at Yanco, New South Wales Australia. The trial layout was 16 rows by 78 ranges in a partially replicated design (Cullis et al. 2006) but data were not collected on the first five ranges. The data collected in the trial involved 672

Table 4 Proportion of correct determination (within 5 cM of the true QTL position) for each of the seven QTL, together with a total out of seven for marker- and interval-based analysis using 3 point (3 pt) and HMM probabilities and for HAPPY

Method	Probabilities	QTL							Total
		1	2	3	4	5	6	7	
Marker	3 pt	1.000	0.644	0.892	0.996	0.996	0.774	0.754	6.056
	HMM	1.000	0.816	0.972	0.994	0.996	0.762	0.766	6.306
Interval	3 pt	1.000	0.870	0.984	0.998	0.994	0.812	0.656	6.314
	HMM	1.000	0.818	0.982	1.000	0.994	0.730	0.734	6.258
HAPPY	HAPPY	0.974	0.614	0.914	0.800	1.000	0.156	0.172	4.630

four-way MAGIC lines plus parents, and 205 other lines with varying replication, namely parents, standard wheat varieties and 121 eight-way lines. Several traits were of interest but here we focus on lodging, scored 1 (upright, no lodging) to 9 (horizontal, complete lodging) for each plot. The MAGIC population provides a diverse founder base in terms of yield and lodging is a limiting factor for yield. Thus the MAGIC population was seen as a natural platform for determining putative QTL for lodging and hence for one component of yield.

The statistical model that provides the baseline for the QTL analysis was given by (1) with

$$\mathbf{X} = \mu \mathbf{1}_n + \tau \mathbf{r}$$

where the fixed effects design $\mathbf{X}$ allows for an overall mean μ and a linear row effect τ with associated explanatory vector $\mathbf{r}$ specifying the row position of each plot; there were no additional random effects in the experiment. The error was taken as $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \Sigma_r \otimes \Sigma_c)$, where Σ_r and Σ_c are both correlation matrices for an autoregressive process of order 1, one in the row direction and the other in the range (or column) direction and $\otimes$ is the Kronecker product; see Gilmour et al. (1997). This latter structure allows for spatial variation in the field. The parameter estimates for the baseline model were $\tau = -0.10$ with a standard error of 0.017, $\sigma_g^2 = 4.65, \sigma^2 = 1.84$ and the spatial correlations $\rho_r = 0.27$ and $\rho_c = 0.39$ in the row and column directions, respectively. The heritability for lodging, using the approach of Oakey et al. (2006), was 0.65.

Results

Simulation study

The first part of the simulation study was to examine the Type I error and false discovery rate (FDR) when no QTL were present. The nominal type I error rate was set at 0.05 and so it is expected that the probability of finding QTL when none were present would be around 0.05. For both marker and interval analyses using either three-point or HMM probabilities the test was found to be conservative. The realized Type I error rates were 0.036 and 0.044 for three-point and HMM probabilities for analyses based on markers or intervals. FDR values were higher, 0.064 and 0.066 for three-point and HMM probabilities for markers and 0.064 and 0.062 for interval-based analysis. These figures reflect the fact that sometimes more than one QTL were found in a null analysis. The results are similar to those given in Verbyla et al. (2007) for bi-parental populations.

The second component of the simulation study was to investigate the power of QTL detection in a setting that mirrors the actual MAGIC population developed in CSIRO, Australia. Thus, seven QTL were simulated for a population of 1,088 lines for a linkage map consisting of 10 linkage groups.

The rate of detection of each QTL (and the overall total across QTL) is presented in Table 4 for five methods of analysis; the methods were MPWGAIM using markers or intervals, using three-point or HMM probabilities and using HAPPY. QTL 1, 4 and 5 were detected in almost all simulations using MPWGAIM. For HAPPY QTL 4 was found 80 % of the time.

The results for QTL 2 and 3 are inconsistent across methods. For MPWGAIM analysis based on markers, using HMM probabilities is preferable whereas for intervals the rates of detection are better when using the three-point probabilities. For HAPPY the proportion of times QTL 2 and 3 were found was lower than MPWGAIM apart from the rate for QTL 3, where HAPPY is better than the marker-based MPWGAIM analysis using three-point probabilities.

The rates for QTL 6 and 7 are similar across all methods apart from HAPPY where there is very poor determination of these two QTL. Overall all MPWGAIM methods, marker or interval using three-point or HMM probabilities find about 6 out of the 7 QTL; the rates are very similar. In contrast, using HAPPY results in 4.6 QTL (on average) being found. Thus, MPWGAIM is an improvement on the analysis using HAPPY.

Note that differences between a marker-based analysis and an interval-based analysis are to be expected because the model using markers assumes marker effects are uncorrelated whereas using intervals induces a correlation between markers.

Table 5 Two by two tables for QTL in repulsion and coupling containing the number of times QTL are selected for marker- and interval-based analysis using 3 point (3 pt) and HMM probabilities and for modified HAPPY

In the table, D stands for detected, while $\bar{D}$ stands for not detected

	Method									
	Marker 3pt		Marker HMM		Interval 3pt		Interval HMM		HAPPY HAPPY	
	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$	D	$\bar{D}$	D
Coupling										
$\bar{D}$	0	2	1	1	0	3	0	3	0	100
D	2	496	2	496	1	496	0	497	0	400
Repulsion										
$\bar{D}$	83	40	117	0	85	87	131	2	340	74
D	30	347	2	381	9	319	4	363	82	4

Table 6 Proportion of false positives over 500 simulations for each QTL chromosome and chromosomes 6–10 which contained no QTL, with totals across all chromosomes for marker- and interval-based analysis using 3 point (3 pt) and HMM probabilities for MPWGAIM, and HAPPY

Method	Probabilities	Chromosome						Total
		1	2	3	4	5	6–10	
Marker	3 pt	0.046	0.288	0.296	0.080	0.528	0.148	1.186
	HMM	0.024	0.030	0.030	0.024	0.006	0.054	0.168
Interval	3 pt	0.028	0.062	0.186	0.088	0.530	0.118	1.012
	HMM	0.026	0.012	0.018	0.010	0.008	0.050	0.124
HAPPY	HAPPY	2.438	0.482	0.804	1.572	0.366	0.012	5.674

Table 5 gives two-way tables for the four combinations of approaches based on MPWGAIM and HAPPY for the two QTL on chromosome 4 that are in coupling and the two QTL on chromosome 5 that are in repulsion. For the two QTL in coupling, all MPWGAIM methods basically find the two QTL in almost every simulation. This is in contrast to the bi-parental situation (Verbyla et al. 2007, 2012) where finding QTL in coupling is very difficult. For HAPPY, using the modified summary that allows for multiple peaks in an island to be recognized, both QTL were detected in 400 of the 500 simulations. Interestingly, in the other 100 simulations the second QTL (in terms of position on the linkage group) was detected. The first QTL was never detected on its own.

For QTL in repulsion, the rates are much lower as shown in Table 5. First, the results for marker and interval are very similar within the three-point and HMM probability cases for MPWGAIM. The rate at which both QTL were detected was higher for the HMM approach, but that approach failed to detect either QTL at a higher rate than using three-point probabilities. In addition, using HMM probabilities essentially leads to both or neither QTL being detected, unlike the three-point probabilities where single QTL were sometimes found. In comparison, finding QTL in repulsion appears very difficult using HAPPY.

Perhaps the biggest differences arose in the proportions of false positives found using the methods; the details are presented in Table 6. It is clear that using MPWGAIM and three-point probabilities leads to high numbers of false positives, effectively one per simulation. In contrast, using HMM probabilities in MPWGAIM leads to very low rates of false positives. The rates for intervals are marginally lower than when using markers. HAPPY resulted in many false positives, on average 5.7 false positives per simulation. It is not clear why using three-point probabilities leads to a higher rate of false positives, but it is conjectured that the probabilities are not accurate when limited scores are available for the markers and there are multiple founders. This is overcome when using probabilities based on the HMM because effectively haplotypes across each linkage group are being used. The high rates for HAPPY are puzzling because the chosen thresholds were calculated using permutation; the determination was based on the output that was produced from analysis using HAPPY. Note, however, that the false positives for HAPPY are on the linkage groups which contain QTL. The rate of false positives on linkage groups not containing QTL is lower than MPWGAIM.

One feature that arose in the simulations when using MPWGAIM was finding multiple QTL in the 5 cM window each side of the QTL. The number of times more than one QTL was found as given in Table 7. Again the clear pattern is that using three-point probabilities leads to confounding of potential QTL for both marker and interval analyses and large numbers of multiple QTL are found. The rates when using HMM probabilities are very small.

Table 7 Number of times in the 500 simulations that two QTL are selected in the 5 cM window for each QTL for marker- and interval-based analysis using 3 point (3 pt) and HMM probabilities

Method	Probabilities	QTL							Total
		1	2	3	4	5	6	7	
Marker	3 pt	5	0	17	1	8	3	71	105
	HMM	4	0	0	5	1	0	0	10
Interval	3 pt	1	0	42	2	9	2	47	103
	HMM	2	0	0	2	0	0	0	4

Lastly, the estimated size of effect for QTL was examined across all simulations and the means and empirical standard errors are presented in Table 8. For each QTL the true sizes are presented for each founder together with the results found for the simulations where the QTL was found. In general, the means are very similar across all methods. Two points of difference that arise are for QTL on chromosome 3 and the two QTL on chromosome 5. The QTL on linkage group 3 had founders 3 and 4 with sizes equal to zero. Using HMM probabilities results in very small means across simulations but when using three-point probabilities the mean sizes are quite different from zero. Using HAPPY, one mean size was close to zero while the other was not. On linkage group 5, the means for QTL 6 are negatively biased when using three-point probabilities while for QTL the mean probabilities for founders 2 and 4 are very close to zero. The means for HAPPY are very different to those simulated across all founders. In contrast the results for analyses using the HMM probabilities appear consistent with the true probabilities. It is not clear why sizes of effects for analysis based on three-point or HAPPY probabilities sometimes are quite different to those simulated. We can only conjecture that for the particular founder, those probabilities are not well estimated and do not discriminate the lines very well. Hence their small estimated sizes. Lastly, standard errors vary in their sizes across the methods and there is no consistent pattern.

Because the QTL sizes tend to be well estimated using MPWGAIM with HMM probabilities, the percentage variance calculations outlined in the methods section are likely to be accurate and will provide a realistic appraisal of the contribution of each QTL to the genetic variance.

Analysis of MAGIC experimental data

The analysis of lodging data from the field trial discussed in the Materials section was conducted using the approaches of the paper.

Putative QTL were found for the lodging data using both interval- and marker-based MPWGAIM analyses, with both three-point and HMM probabilities as outlined in the Methods. Results are presented only for the interval-based analysis using HMM probabilities because of the findings in the simulation study. However, the corresponding results for a marker-based analysis using HMM based probabilities and also using three-point probabilities can be found in the Supplementary online material.

For the interval-based analysis using HMM probabilities 10 putative QTL were found; see Table 9. The two largest QTL (in terms of percentage variance and LOGP score) are the Green Revolution height reducing genes *Rht-B1b* and *Rht-D1b* (formerly known as *Rht1* and *Rht2*). *Rht-D1b* was placed at the start of chromosome 4D but had many missing values for the four-way lines and a neighboring region was selected in the QTL analysis. The remaining putative QTL explain small percentages of the genetic variance but have reasonable LOGP scores and the size of QTL effects suggest these additional QTL would be useful in improving lodging if the appropriate alleles were selected.

If the supplementary material is examined it can be seen that for the marker-based analysis using HMM probabilities 12 QTL are found with 9 being essentially the same as for the interval analysis. The analyses based on the three-point probabilities lead to 13 and 12 putative QTL for the interval- and marker-based approaches, respectively, again with 9 in common with the results in Table 9. Thus, the results are very similar across the approaches.

Discussion

An important aspect of QTL analysis in MAGIC populations is that marker scores may not be fully informative because there are multiple founders. The WGAIM approach of Verbyla et al. (2007) has been extended to a multi-parent population (MAGIC) with inter-crossing and selfing in a natural way using probabilities of inheriting founder alleles.

The current approach is based on Verbyla et al. (2012) where QTL effects are assumed random rather than fixed. This reduces the bias of estimated QTL sizes. The approach allows assessment of the impact of the putative QTL through a probability value of the extremity of the impact, through percentage genetic variance and a LOGP score. For an analysis using markers, the use of three-point and HMM probabilities for founder effects was incorporated using an averaging approach akin to the original WGAIM method.

Table 8 Means and standard errors of the mean for each QTL and each founder size for marker- and interval-based analysis using 3 point (3 pt) and HMM probabilities and using HAPPY

Founder									
QTL	Method	Probabilities	1		2		3		4	
			Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	Simulated		0.354		−0.354		−0.354		0.354	
	Marker	3 pt	0.345	0.0012	−0.345	0.0012	−0.345	0.0012	0.345	0.0012
	Marker	HMM	0.368	0.0026	−0.336	0.0021	−0.357	0.0022	0.326	0.0026
	Interval	3 pt	0.350	0.0012	−0.350	0.0012	−0.350	0.0012	0.350	0.0012
	Interval	HMM	0.375	0.0026	−0.340	0.0021	−0.360	0.0022	0.325	0.0026
	HAPPY	HAPPY	0.361	0.0013	−0.346	0.0027	−0.377	0.0016	0.362	0.0013
2	Simulated		0.354		−0.118		−0.118		−0.118	
	Marker	3 pt	0.324	0.0025	−0.116	0.0019	−0.119	0.0016	−0.090	0.0024
	Marker	HMM	0.339	0.0022	−0.132	0.0025	−0.105	0.0020	−0.102	0.0022
	Interval	3 pt	0.340	0.0022	−0.121	0.0022	−0.119	0.0015	−0.100	0.0021
	Interval	HMM	0.347	0.0021	−0.134	0.0025	−0.107	0.0020	−0.106	0.0022
	HAPPY	HAPPY	0.349	0.0021	−0.078	0.0020	−0.139	0.0018	−0.132	0.0016
3	Simulated		0.354		−0.354		0.000		0.000	
	Marker	3 pt	0.296	0.0024	−0.106	0.0015	−0.082	0.0025	−0.109	0.0014
	Marker	HMM	0.339	0.0023	−0.324	0.0031	0.000	0.0023	−0.016	0.0030
	Interval	3 pt	0.297	0.0039	−0.129	0.0016	−0.038	0.0060	−0.129	0.0016
	Interval	HMM	0.345	0.0023	−0.328	0.0032	0.000	0.0023	−0.017	0.0031
	HAPPY	HAPPY	0.300	0.0055	−0.153	0.0014	0.006	0.0035	−0.153	0.0014
4	Simulated		0.354		−0.354		−0.354		0.354	
	Marker	3 pt	0.406	0.0033	−0.350	0.0027	−0.364	0.0022	0.308	0.0032
	Marker	HMM	0.343	0.0026	−0.358	0.0027	−0.334	0.0024	0.349	0.0027
	Interval	3 pt	0.384	0.0021	−0.347	0.0030	−0.410	0.0031	0.373	0.0018
	Interval	HMM	0.364	0.0027	−0.376	0.0028	−0.352	0.0026	0.364	0.0029
	HAPPY	HAPPY	0.384	0.0055	−0.447	0.0043	−0.321	0.0074	0.384	0.0054
5	Simulated		0.354		−0.354		−0.354		0.354	
	Marker	3 pt	0.300	0.0016	−0.314	0.0027	−0.286	0.0023	0.300	0.0016
	Marker	HMM	0.322	0.0025	−0.328	0.0026	−0.333	0.0025	0.339	0.0025
	Interval	3 pt	0.300	0.0017	−0.312	0.0028	−0.288	0.0024	0.300	0.0017
	Interval	HMM	0.308	0.0025	−0.321	0.0028	−0.318	0.0026	0.331	0.0026
	HAPPY	HAPPY	0.433	0.0014	−0.417	0.0024	−0.448	0.0032	0.433	0.0014
6	Simulated		0.354		−0.354		−0.354		0.354	
	Marker	3 pt	0.299	0.0036	−0.221	0.0039	−0.289	0.0041	0.212	0.0048
	Marker	HMM	0.338	0.0032	−0.312	0.0029	−0.349	0.0034	0.322	0.0034
	Interval	3 pt	0.294	0.0035	−0.244	0.0043	−0.308	0.0037	0.257	0.0043
	Interval	HMM	0.345	0.0033	−0.326	0.0031	−0.355	0.0035	0.336	0.0035
	HAPPY	HAPPY	0.140	0.0035	−0.070	0.0054	−0.203	0.0040	0.133	0.0052
7	Simulated		−0.354		0.354		0.354		−0.354	
	Marker	3 pt	−0.283	0.0040	0.011	0.0035	0.260	0.0074	0.011	0.0035
	Marker	HMM	−0.339	0.0030	0.309	0.0034	0.346	0.0033	−0.316	0.0037
	Interval	3 pt	−0.295	0.0038	0.022	0.0040	0.251	0.0080	0.022	0.0040
	Interval	HMM	−0.341	0.0031	0.318	0.0036	0.345	0.0033	−0.322	0.0038
	HAPPY	HAPPY	−0.225	0.0033	0.003	0.0028	0.220	0.0040	0.002	0.0029

An important component of MPWGAIM is the dimension reduction that allows large problems to be analyzed efficiently. The dimension is reduced to the number of lines having marker information and this can be much smaller than the effective dimension in multi-parent situations. There are relatively simple results that allow the required calculations (of the outlier statistics) to be carried out. These ideas are being extended to both the multivariate

Table 9 Putative QTL for the MAGIC lodging data found using intervals and HMM probabilities

Chr	Left dist	Right dist	Founder	Size	Founder		Prob	%var	LOGP
					Prob	LOGP			
1D	59.15	72.88	Yitpi	0.408	0.052	1.29	0.039	1.2	1.41
			Chara	−0.199	0.233	0.63			
			Baxter	0.06	0.411	0.39			
			Westonia	−0.309	0.129	0.89			
2B	262.54	263.04	Yitpi	−0.47	0.041	1.39	0.068	1.2	1.17
			Chara	0.159	0.302	0.52			
			Baxter	0.18	0.302	0.52			
			Westonia	0.082	0.412	0.38			
3A	84.35	86.92	Yitpi	−0.442	0.094	1.03	0	3.7	4.41
			Chara	−0.316	0.199	0.7			
			Baxter	−0.068	0.423	0.37			
			Westonia	0.736	0.012	1.92			
3A	183.94	187.57	Yitpi	−0.488	0.037	1.43	0.007	1.9	2.13
			Chara	0.03	0.457	0.34			
			Baxter	0.471	0.044	1.35			
			Westonia	−0.067	0.406	0.39			
3A	309.24	309.74	Yitpi	0.706	0.011	1.95	0	3.4	4.36
			Chara	−0.5	0.053	1.28			
			Baxter	−0.161	0.298	0.53			
			Westonia	−0.125	0.346	0.46			
4B	120	121.01	Yitpi	0.843	0.049	1.31	0	15	17.78
			Chara	−0.487	0.18	0.75			
			Baxter	−1.469	0.003	2.58			
			Westonia	0.785	0.063	1.2			
4D	11.04	18	Yitpi	−1.804	0.002	2.77	0	29.7	30.37
			Chara	1.348	0.013	1.89			
			Baxter	1.03	0.046	1.34			
			Westonia	−1.193	0.028	1.55			
5B	111.53	113.05	Yitpi	0.237	0.189	0.72	0.007	1.8	2.15
			Chara	0.389	0.076	1.12			
			Baxter	−0.256	0.177	0.75			
			Westonia	−0.42	0.059	1.23			
7B	196.94	197.45	Yitpi	0.618	0.014	1.86	0.001	2.4	3.17
			Chara	−0.08	0.391	0.41			
			Baxter	−0.317	0.136	0.87			
			Westonia	−0.282	0.167	0.78			
Unlinked2	31.06	32.07	Yitpi	0.053	0.41	0.39	0.018	1.2	1.75
			Chara	−0.262	0.134	0.87			
			Baxter	−0.234	0.155	0.81			
			Westonia	0.407	0.037	1.43			

situation (multi-trait, multi-environment and multiple treatments) and the analysis of epistatic interactions in both bi-parental and multi-parent situations where the dimension reduction is vital.

An important issue is the difference between a marker-based and an interval-based analysis. Verbyla et al. (2007) show that using intervals, the underlying marker effects are correlated. This is to be expected because markers that are in close linkage with a QTL would be expected to exhibit similar association with the trait. In contrast, using a marker-based analysis assumes independent marker effects (for the models of this paper). This will lead to differences in analysis of QTL, though we might expect similar results for both methods as was found in the simulation study and

the example. For an analysis based on markers, it may be preferable to propose a correlated set of effects rather than an uncorrelated set. This was not explored for the current paper, but is a general issue to be considered.

The simulation study showed that MPWGAIM was very good at finding QTL, whether interval or marker based, and whether three-point or HMM probabilities were used. However, the latter probabilities seemed to produce more unbiased QTL size effects than the corresponding effects found when using three-point probabilities. In addition, false positive rates were very low when using HMM probabilities and unacceptably high when using three-point probabilities. These latter probabilities tended to give rise to linked spurious QTL close to the true QTL (which was also found). We conjecture that the reason for the increased rate of false positives when using three-point probabilities is the reduced information content of three-point haplotypes for multiple founders in comparison to using HMM methods where the effective haplotype length is the entire linkage group.

The comparison with HAPPY (Mott et al. 2000) in the power simulation study showed that MPWGAIM detects more true QTL, and that for HAPPY it was difficult to detect QTL in repulsion. In addition, there were many false positives when using HAPPY. MPWGAIM outperforms HAPPY.

The lodging data for a wheat MAGIC 4-way population was analyzed and confirmed the similarity of analyses using intervals or markers. There have been several papers in which QTL for lodging have been determined for bi-parental populations. Keller et al. (1999) consider lodging data at three environments (analyzed separately) and find 13, 8 and 8 QTL with five in common across all environments and three in common at two environments. Of the QTL found in our analysis, it appears that one of the QTL on 3A, and the QTL on 5B and 7B may be in the same region as those found by Keller et al. (1999). The remainder are different. Verma et al. (2005) consider analysis over 2 years of data (separate analyses for each year) and found QTL on 4B, 4D, 6D and 7D in both years and on 2B and 1D in 1 year. The QTL on 4B and 4D are the height genes, also found in our analyses. Smaller plants are less prone to lodging. QTL on 2B and 1D were found in our analyses and may correspond to those found by Verma et al. (2005). In their study, McCartney et al. (2005) find three QTL, on 4B, 4D and 3D, so again the height genes play an important role. Lastly, Marza et al. (2006) find three QTL on 1B, 4AL and 5A, none of which were found in our analyses.

In conclusion, MPWGAIM is a powerful approach for QTL analysis that performed well in the simulation study. MPWGAIM is a better method than HAPPY, again based on the simulation study. Using HMM probabilities leads to low rates of false positives and so the recommendation based on this paper is to use HMM probabilities with either the interval or marker-based MPWGAIM approach.

Acknowledgments The authors thank the reviewers and the editor for their constructive comments which lead to major improvements in the paper.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

The size of $\mathbf{P}_A$ and hence the number of effects in $\mathbf{a}$, $(r-c)n_{\mathrm{f}}$ for interval-based analyses and rn_{f} for marker-based analyses, can be very large. This usually results in the number of effects required to be estimated being bigger than the sample size n. Verbyla et al. (2012) propose an approach that reduces the dimension for model fitting and the same approach can be used for MPWGAIM. This reduces the dimension for model fitting to the number of lines n_g, in the same manner as for the bi-parental situation.

If $\mathbf{a}^* \sim N(\mathbf{0}, \sigma_a^2 \mathbf{I}_{n_g})$, the model

$$\mathbf{u}_g = (\mathbf{P}_A \mathbf{P}_A^{\mathrm{T}})^{1/2} \mathbf{a}^* + \mathbf{u}_p$$

results in the same variance model as (7), but the number of effects in $\mathbf{a}^*$ equals the number of lines n_g. Then as in Verbyla et al. (2012), the BLUPs of $\mathbf{a}^*$ and $\mathbf{a}$, denoted by $\tilde{\mathbf{a}}^*$ and $\tilde{\mathbf{a}}$ respectively, are related by

$$\tilde{\mathbf{a}} = \mathbf{P}_A^{\mathrm{T}} (\mathbf{P}_A \mathbf{P}_A^{\mathrm{T}})^{-1/2} \tilde{\mathbf{a}}^*$$

with variance matrix

$$\mathrm{var}(\tilde{\mathbf{a}}) = \mathbf{P}_A^{\mathrm{T}} (\mathbf{P}_A \mathbf{P}_A^{\mathrm{T}})^{-1/2} \mathrm{var}(\tilde{\mathbf{a}}^*) (\mathbf{P}_A \mathbf{P}_A^{\mathrm{T}})^{-1/2} \mathbf{P}_A$$

and only the diagonal elements of this matrix are required in the calculation of the outlier statistics (12). Thus, an efficient computational approach exists for high-dimensional situations.

References

Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, Tung CW, McCouch S, Thomson M, Mauleon R, Singh RK, Gregorio G, Redoa E, Leung H (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:11

Broman K (2005) The genomes of recombinant inbred lines. Genetics 169:1133–1146

Broman KW (2006) Use of hidden Markov models for QTL mapping. Technical report. John Hopkins University, Department of Biostatistics, working paper 125

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc Ser B 64:641–656

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890

Broman KW, Wu H, Churchill G, Sen S, Yandell B, et al (2012) qtl: Tools for analyzing QTL experiments. http://CRAN.R-program.org/package=qtl R package version 1.23-16

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2011) Mixed models for S language environments: ASReml-R reference manual. Technical report. Queensland Department of Primary Industries, http://www.vsni.co.uk/software/asreml/

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci 110:8057–8062. doi:10.1073/pnas.1217133110

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11:381–393

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2:269–293

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Huang BE, George AW (2011) R/mpmap: A computational platform for the genetic analysis of multi-parent recombinant inbred lines. Bioinformatics 27:727–729

Huang BE, George AW, Forrest KL, Kilian A, Hayden M, Morell MK, Cavanagh CR (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10:826–839

Jansen RC (1994) Controlling the Type I and Type II errors in mapping quantitative trait loci. Genetics 138:871–881

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152:1203–1216

Keller M, Karutz C, Schmid JE, Stamp P, Winzeler M, Keller B, Messmer MM (1999) Quantitative trait loci for lodging resistance in a segregating wheat x spelt population. Theor Appl Genet 98:1171–1182

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine- map quantitative traits in *Arabidopsis thaliana*. PLoS Genet 5(e1000):551

Malosetti M, van Eewijk FA, Boer MP, Casas MAM, Elia M, Moralejo M, Bhat PR, Ramsey L, Molina-Cano JL (2011) Gene and QTL detection in a three-way barley cross under selection ny a mixed model with kinship information using SNPs. Theor Appl Genet 122:1605–1616

Martinez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Marza F, Bai GH, Carver BF, Zhou WC (2006) Quantitative trait loci for yield and related traits in the wheat population Ning7840 x Clark. Theor Appl Genet 112:688–698

McCartney CA, Somers DJ, Humphreys DG, Lukow O, Ames N, Noll J, Cloutier S, McCallum BD (2005) Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross rl4452 x 'ac domain'. Genome 48:870–883

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci USA 97:12,649–12,654

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

Piessens R, de Doncker-Kapenga E, Uberhuber CW, Kahaner DK (1983) QUADPACK: a subroutine package for automatic integration. Springer, Berlin

R Development Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org, ISBN 3-900051-07-0

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

Smith A, Cullis B, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agric Sci (Cambridge) 143:449–462

Smith AB, Lim P, Cullis BR (2006) The design and analysis of multi-phase quality trait experiments. J Agric Sci (Cambridge) 144:393–409

Stram DO, Lee JW (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177

Taylor JD, Diffey S, Verbyla AP, Cullis BR (2011) wgaim: Whole genome average interval mapping for QTL detection using mixed models. http://CRAN.R-program.org/package=wgaim R package version 1.1

Threadgill DW, Hunter KW, Williams RW (2002) Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. Mamm Genome 13:175–178

Trebbi D, Maccaferri M, Giuliani S, Sorensen A, Sanquineti MC, Massi A, Tuberosa R (2008) Development of a multi-parental (four-way cross) mapping population for multi-allelic QTL analysis in durum wheat. In: Appels R, Eastwood R, Lagudah E, Langridge P, Mackay M, McIntyre L (eds) The 11th International Wheat Genetics Symposium proceedings. University of Sydney

Trow A (1913) Forms of reproduction: primary and secondary. J Genet 2:313–324

Valdar W, Flint J, Mott R (2006a) 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

Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JNP, Mott R, Flint J (2006b) Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 38:879–887

Verbyla AP (1990) A conditional derivation of residual maximum likelihood. Aust J Stat 32:227–230

Verbyla AP, Cullis BR (2012) Multivariate whole genome average interval mapping: QTL analysis for multiple traits and/or environments. Theor Appl Genet 125:933–953

Verbyla AP, Cullis BR, Thompson R (2007) The analysis of QTL by simultaneous use of the full linkage map. Theor Appl Genet 116:95–111

Verbyla AP, Taylor JD, Verbyla KL (2012) RWGAIM: An efficient high dimensional random whole genome average (QTL) interval mapping approach. Genet Res 94:291–306

Verma V, Worland AJ, Sayers EJ, Fish L, Caligari PDS, Snape JW (2005) Identification and characterization of quantitative trait loci related to lodging resistance and associated traits in bread wheat. Plant Breed 124:234–241

Xu S (1996) Mapping quantitative trait loci using four-way crosses. Genet Res 68:175–181

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468