Theor Appl Genet (2012) 124:1097–1113
DOI 10.1007/s00122-011-1772-8

ORIGINAL PAPER

QTL linkage analysis of connected populations using ancestral marker and pedigree information

Marco C. A. M. Bink · L. Radu Totir ·
Cajo J. F. ter Braak · Christopher R. Winkler ·
Martin P. Boer · Oscar S. Smith

Received: 20 April 2011 / Accepted: 15 December 2011 / Published online: 7 January 2012
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract The common assumption in quantitative trait locus (QTL) linkage mapping studies that parents of multiple connected populations are unrelated is unrealistic for many plant breeding programs. We remove this assumption and propose a Bayesian approach that clusters the alleles of the parents of the current mapping populations from locus-specific identity by descent (IBD) matrices that capture ancestral marker and pedigree information. Moreover, we demonstrate how the parental IBD data can be incorporated into a QTL linkage analysis framework by using two approaches: a Threshold IBD model (TIBD) and a Latent Ancestral Allele Model (LAAM). The TIBD and LAAM models are empirically tested via numerical simulation based on the structure of a commercial maize breeding program. The simulations included a pilot dataset with closely linked QTL on a single linkage group and 100 replicated datasets with five linkage groups harboring four unlinked QTL. The simulation results show that including parental IBD data (similarly for TIBD and LAAM) significantly improves the power and particularly accuracy of QTL mapping, e.g., position, effect size and individuals' genotype probability without significantly increasing computational demand.

Communicated by M. Frisch.

M. C. A. M. Bink (✉) · C. J. F. ter Braak · M. P. Boer
Department of Biometris, Wageningen University and Research
Centre, Droevendaalsesteeg 1, 6708 PB Wageningen,
The Netherlands
e-mail: marco.bink@wur.nl

L. R. Totir · C. R. Winkler · O. S. Smith
Pioneer Hi-Bred International, A DuPont Company,
Johnston, IA 50131, USA

Introduction

The quantitative dissection of complex traits into underlying genetic components has been the stated goal of many generations of quantitative geneticists (Falconer 1989). Recently, increased availability of molecular markers combined with enhanced statistical analysis techniques has given quantitative geneticists new tools. One simple approach to achieve the goal is to use quantitative trait locus (QTL) detection methods that exploit phenotypic and molecular marker data collected in designed bi-parental mapping populations of large size (Boer et al. 2007). However, such an approach has a serious limitation in that it explores only a small fraction of the genetic variance available in the reference population from which the two parents of the bi-parental mapping population are sampled. Additionally, analyses of mapping populations from different parents for the same trait can give inconsistent estimates of QTL positions and effect sizes (Beavis et al. 1991). QTL analysis of connected populations has been advocated as an alternative to increase the amount of genetic variability accounted for in the statistical model (Bink et al. 2002; Blanc et al. 2006). This approach is also expected to yield more consistent QTL mapping results. However, a common assumption in this approach is that the parents of the connected populations are unrelated and thus can be treated as independent (Blanc et al. 2006; Crepieux et al. 2005; Fang et al. 2011; Hayashi and Iwata 2009). While this assumption is convenient from the standpoint of the statistical analysis, it does not reflect the reality of most breeding programs and leads to loss of power in QTL estimation when the parents are in fact related.

The mapping resolution in QTL linkage studies depends, among other factors, on the number of meioses events accounted for in the statistical model. Therefore, accounting for meioses events that occurred in the ancestors of the parents of the current mapping population should be beneficial in the detection of QTL and the precise placement of these QTL on the genetic map. Meuwissen and Goddard (2000) proposed such methodology for the precise mapping of loci affecting quantitative traits. This methodology combines linkage and linkage-disequilibrium information where the latter is a function of the historical/ancestral recombinations. While Meuwissen and Goddard (2000) made use of population genetics theory to model linkage disequilibrium, an alternative scenario is possible when highly accurate pedigree and marker information exists for the recent ancestral generations of the parents of connected populations. This is especially true when the ancestral pedigree has been genotyped for many more genetic markers than the connected populations to be used for QTL mapping. However, explicit inclusion of the marker and pedigree information collected on ancestors into the dataset to be analyzed can create a significant missing marker data problem which requires significant imputation efforts when used for mapping experiments of a size typical for breeding programs working with commercial elite germplasm. Instead of including the high-density genotyped ancestors themselves into the statistical analysis, we propose an approach that collapses the marker and pedigree information from the ancestors into parental identity by descent (PIBD) information.

This article presents a novel Bayesian approach to combine PIBD information into a QTL linkage analysis framework. The PIBD information pertains to the parents of the connected populations that form the analysis dataset and this information may be obtained in various ways. However, it is assumed that this information is in the form of an IBD matrix specific to a particular genomic position. Here we extend the Bayesian hierarchical framework of Bink et al. (2008) to allow for latent ancestral alleles that are derived from the locus-specific PIBD matrices. The approach is empirically tested using simulated phenotypic and marker data conditional on a pedigree specific to a maize mapping population. Extensions and implications to other QTL mapping experiments are discussed.

Materials and methods

In conventional QTL linkage mapping it is common to assume independence among the parental alleles of the mapping population(s) (Blanc et al. 2006; Crepieux et al. 2005; Fang et al. 2011; Hayashi and Iwata 2009). Here, this assumption is replaced by allowing for putative dependencies between the parents based on ancestral pedigree and marker information.

In the description of the methodology and implementation, we will concentrate on mapping populations containing inbred lines, i.e., individuals are homozygous at all loci. Consequently, we may use the terms allele, haplotype, and individual synonymously. The theoretical concepts presented in this article can be readily adapted for outbred populations.

Consider a set of n_j multiple mapping populations ($J_1 - J_{\mathrm{m}}$) that have a connectedness through a set of parents ($I_1 - I_{\mathrm{n}}$) as shown in Fig. 1a. We adopt a bi-allelic additive QTL model where Q (q) denotes the allele that increases (decreases) the quantitative trait value. The frequency of allele Q is denoted by p for which we assume a uniform prior distribution between 0 and 1. Furthermore, let vector $\boldsymbol{\lambda}$ denote the positions of all QTL in the model, where the number of putative QTL (N_{QTL}) is treated as a random variable in our Bayesian approach and the prior distribution of the positions of the putative QTL along the marked genome is assumed uniform and continuous.

Three types of data are available, i.e., phenotypic trait data ($\mathbf{Y}_{\boldsymbol{T}}$), low-density marker data on mapping populations ($\mathbf{Y}_{\boldsymbol{M}}$), and parental IBD data ($\mathbf{Y}_{\boldsymbol{D}}$). The parental IBD data are IBD probabilities among the parents of the mapping populations, available as symmetric matrices $\mathbf{Q}$ for a set of n_Q positions along the genome. At each genomic position an element Q_{ij} of $\mathbf{Q}$ is the probability that parents I_i and I_j are IBD.

Modeling QTL genotypes

Let n_i denote the number of parents, n_j the number of mapping populations (crosses), $n_o[j]$ the number of offspring in the jth mapping population, and n_{a} the number of ancestral alleles. Then, we denote $\mathbf{G}$ as the ($n_i \times N_{\mathrm{QTL}}$) matrix of parental alleles with $i = 1, \ldots, n_i$) and $qtl = 1, \ldots, N_{\mathrm{QTL}}$. Similarly, let $\mathbf{S}$ denote the ($n_O \times N_{\mathrm{QTL}}$) matrix of segregation (or meiosis) indicators (Donnelly 1983; Lander and Green 1987), where n_O is the total number of offspring across all mapping populations, i.e., $n_O = \sum_{j=1}^{n_j} n_o[j]$; let $\mathbf{C}$ denote the ($n_i \times N_{\mathrm{QTL}}$) matrix of ancestral class indicators; and let $\mathbf{A}$ denote the ($n_{\mathrm{a}} \times N_{\mathrm{QTL}}$) matrix of ancestral alleles. Finally, for ease of readability we will suppress the subscripts pertaining to qtl and describe the concept as if only one QTL is assumed.

Original framework used to model QTL genotypes

In the original framework of Bink et al. (2008) the parental genotypes are assumed to be unrelated or independent (Fig. 1c). We denote this model as UNR in the remainder of this study. The prior distribution for the parental genotypes at a QTL is

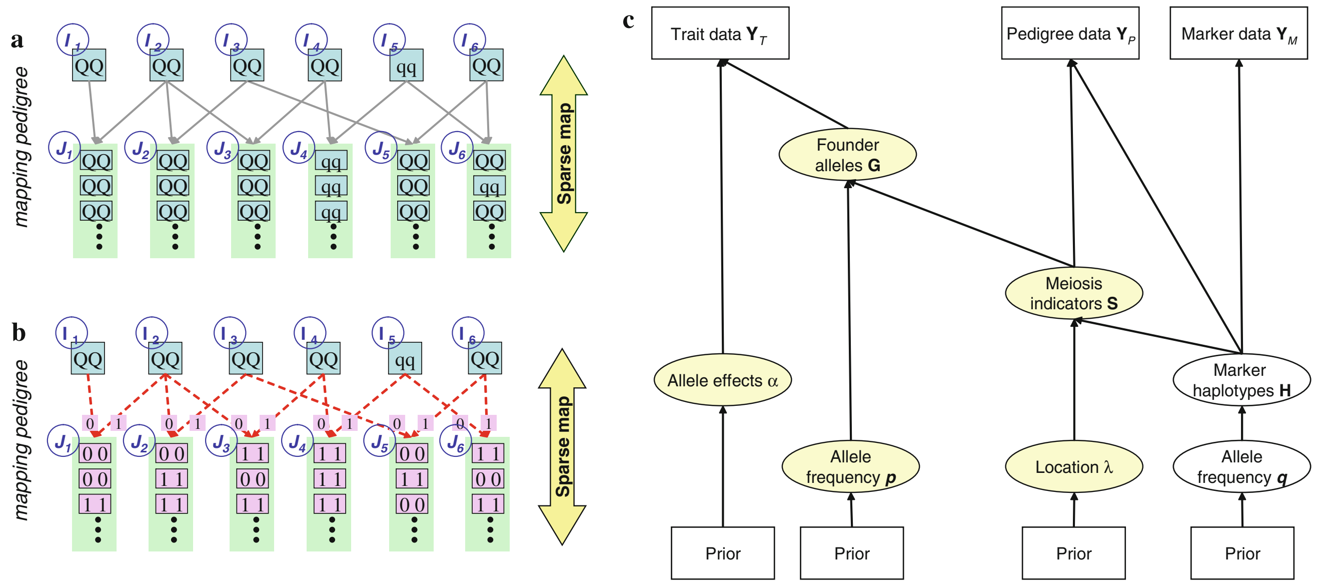

Fig. 1 A graphical representation of QTL genotypes (*QQ* and *qq*) in an inbred pedigree of six connected mapping populations (J_1–J_6) derived from crosses between unrelated inbred parents (I_1–I_6). **a** provides the actual genotypes whereas **b** is a representation using 0/1 segregation indicators to describe parental origin in the mapping offspring. **c** shows the corresponding directed acyclic graphs for unrelated mapping parents with founder alleles (**G**)

$$P(\mathbf{G}|p) = \prod_{i=1}^{n_i} P(g_i|p) \tag{1}$$

where g_i denotes the genotype of parent i and p denotes the allele frequency at a QTL. The QTL genotypes of the offspring in the mapping populations ($J_1 - J_{\rm m}$) are defined as functions of the parental genotypes and segregation (meiosis) indicators (Donnelly 1983; Lander and Green 1987). For an inbred offspring o, the interpretation of a binary segregation indicator $s_{\rm o}$ is that 0 (1) pertains to the 1st (2nd) inbred parent of that mapping population (Fig. 1b). The prior distribution for the QTL segregation indicators of all offspring of all mapping populations is

$$P(\mathbf{S}|\lambda, \mathbf{Y}_M) = \prod_{j=1}^{n_j} \prod_{o=1}^{n_o[j]} P(s_o|\lambda, lfm_o, rfm_o) \tag{2}$$

where $n_o[j]$ is the number of offspring from population j and $\mathbf{Y}_M$ pertains to marker data and *lfm* (*rfm*) denotes the left-sided (right-sided) informative flanking markers for inbred offspring o.

Framework used to model dependence of parental QTL genotypes—ancestral alleles

In contrast to the original framework of Bink et al. (2008) (Fig. 1), we now consider ancestral relationships in modeling QTL alleles of the parents of the mapping population (Fig. 2). In Fig. 2a an example of a complete ancestral pedigree is depicted using an assigned descent path of QTL genotypes from ancestors to the parents of the mapping population. Alternatively, if the ancestral pedigree is not modeled explicitly the QTL genotypes in the parents are assumed to be copies of anonymous ancestral alleles (Fig. 2b). The ancestral alleles are anonymous as the ancestral pedigree is not modeled explicitly in the proposed approach. For known pedigrees, it may be possible to link the ancestral alleles to real ancestors in the recorded pedigree (ter Braak et al. 2010). In our example, this holds for ancestor alleles A_1, A_2, and A_3 (Fig. 2b). Furthermore, the number of ancestral alleles (n_a) is considered random and may vary at different loci along the genome. We denote the identity of these ancestral copies as observed in the parents of the mapping population via ancestral class indicators I, II, etc. (Fig. 2b). The number of ancestral alleles equals three in the example in Fig. 1 where ancestral allele I, II, and III have, respectively, 2, 1, and 3 copies present in the parental genotypes. For an inbred parent i, the interpretation of an integer ancestral class indicator c_i is that its value a, $a = 1, \ldots, n_a$, pertains to the ath ancestral allele. Analogous to the segregation indicators, **S**, in (2), the prior distribution for the ancestral class indicators is

$$P(\mathbf{C}|\lambda, \mathbf{Y}_D) = \prod_{i=1}^{n_i} P(c_i|\lambda, Y_D(l), Y_D(r)) \tag{3}$$

where $Y_D(l)$,($Y_D(r)$) denotes the immediate left (right) flanking PIBD information. The additional source of information $\mathbf{Y}_D$ is used to account for locus-specific relationships among parents as generated by the ancestral pedigree and marker information. It will be discussed in detail below. Finally, the prior distribution for ancestral QTL alleles is

$$P(\mathbf{A}|p) = \prod_{l=1}^{n_a} P(a_l|p) \tag{4}$$

where a_l is the allele of ancestor l. This prior is similar to the prior of Eq. 1 of the parental genotypes in the original model of Bink et al. (2008), namely it assumes independence among ancestral alleles (Fig. 2a).

Models for parental identity by descent (PIBD) data

The locus-specific PIBD matrices jointly form the PIBD data (denoted $\mathbf{Y}_D$, Fig. 2); thus $\mathbf{Y}_D$ represents an array of $\mathbf{Q}$ matrices. An example of such a $\mathbf{Q}$ matrix is presented in Table 1A. These IBD probabilities cannot be used directly in the Bayesian sampling algorithms of Bink et al. (2008). However, two alternative models have been suggested to capture the information provided by these data (ter Braak et al. 2010). We discuss their implementation in a Bayesian framework below.

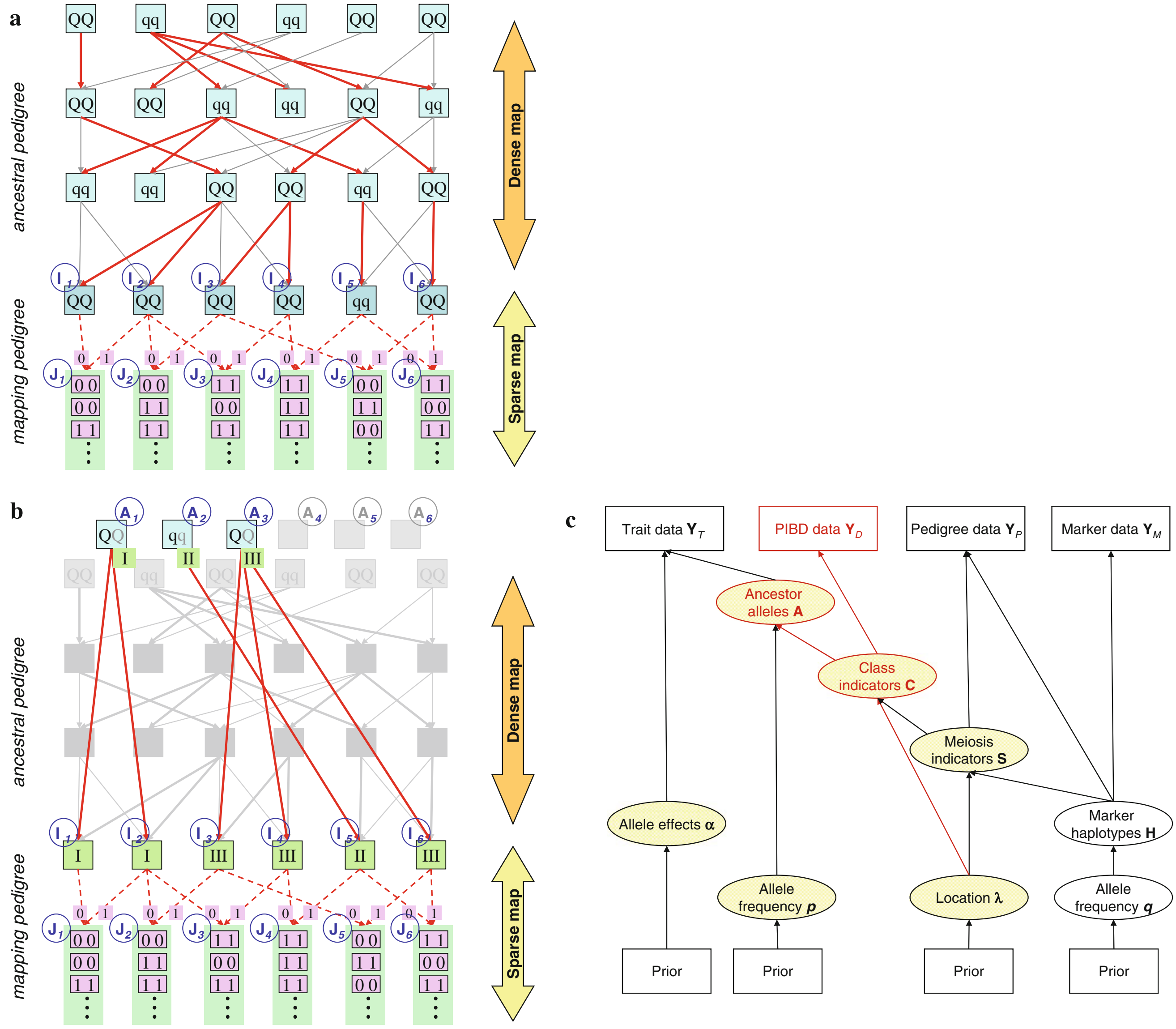

Fig. 2 A graphical representation of an known ancestral pedigree with known transmission of QTL alleles (*red arrows*) is included in (**a**). This same information is condensed in (**b**) using assignments (I–III) linking the alleles of the mapping parents to the ancestral alleles (A_1–A_6). Note that ancestral alleles may or may not coincide with real ancestors in the pedigree. The assignment to ancestral alleles is based on the TIBD classification of Table 1. **c** shows the corresponding Directed Acyclic Graphs for related mapping parents. Parent IBD data (an array of **Q**-matrices) to account for relatedness by replacing the founder alleles (**G**) by Ancestor alleles (**A**) and Class indicators (**P**)

Table 1 Numerical example of IBD probability matrix **Q** among six inbred individuals (A) and the corresponding probability matrix **P** for the latent ancestor model (LAAM) (B). The IBD status matrix $\mathbf{Q}_{\mathrm{TIBD}}$ (C) and the corresponding ancestor assignments ($\mathbf{P}_{\mathrm{TIBD}}$) for the threshold model (D), where the IBD status of pair I$_4$–I$_6$ has been adjusted for reason of transitivity. The assignments correspond to the example given in Fig. 1

A **Q**

	I$_1$	I$_2$	I$_3$	I$_4$	I$_5$	I$_6$
I$_1$	1.00	0.95	0	0	0.17	0
I$_2$	0.95	1.00	0	0	0.03	0
I$_3$	0	0	1.00	0.81	0	0.80
I$_4$	0	0	0.81	1.00	0	0.65
I$_5$	0.17	0.03	0	0	1.00	0
I$_6$	0	0	0.80	0.65	0	1.00

B $\mathbf{P}_{\mathbf{LAAM}}$

	A$_1$	A$_2$	A$_3$	A$_4$	A$_5$
I$_1$	0.96	0.04	0	0	0
I$_2$	0.99	0.01	0	0	0
I$_3$	0	0	1.00	0	0
I$_4$	0	0	0.81	0.19	0
I$_5$	0.07	0.93	0	0	0
I$_6$	0	0	0.80	0.01	0.19

C $\mathbf{Q}_{\mathbf{TIBD}}$

	I$_1$	I$_2$	I$_3$	I$_4$	I$_5$	I$_6$
I$_1$	1	1	0	0	0	0
I$_2$	1	1		0	0	0
I$_3$	0	0	1	1	0	1
I$_4$		0	1	1	0	*1*
I$_5$	0	0	0	0	1	0
I$_6$	0	0	1	*1*	0	1

D $\mathbf{P}_{\mathbf{TIBD}}$

	A$_1$	A$_2$	A$_3$
I$_1$	1	0	0
I$_2$	1	0	0
I$_3$	0	0	1
I$_4$	0	0	1
I$_5$	0	1	0
I$_6$	0	0	1

Threshold IBD model (TIBD)

This first model uses a threshold on the IBD probabilities, for example a threshold value of 0.80 and is denoted as the TIBD model. If the IBD probability of two inbred individuals (*i* and *j*) sharing the same ancestral allele exceeds this threshold, the alleles are assumed to have the same ancestral allele and their IBD probability is substituted with a value of 1.0, i.e.,

$$\begin{array}{ll} Q_{T_{/\mathrm{BD}},ij} = 0 & \text{if } Q_{ij} < T_{\mathrm{IBD}} \\ Q_{T_{\mathrm{IBD}},ij} = 1 & \text{otherwise} \end{array} \tag{5}$$

Values below the threshold are replaced by 0 unless a transitivity problem arises (ter Braak et al. 2010). In case of a transitivity problem, the threshold is locally lowered to create consistency in IBD patterns (Table 1C). This threshold-based approach results in a crisp 0/1 matrix; for example the TIBD model for the example in Table 1A yields three ancestral allele classes (Table 1D). The inbreds I_1 and I_2 are copies of ancestral allele A_1, inbreds I_3, I_4, and I_6 are copies of ancestral allele A_3, and inbred I_5 is a copy of ancestral allele A_2. Note that the IBD probability between inbreds I_4 and I_6 is below the threshold but set to 1 because of the transitivity rule.

The position of a putative QTL is assumed to be continuous along the genome and can thus be in between positions at which PIBD data are available. The **Q** matrix pertaining to the putative position λ_{qtl} may be calculated as the weighted average of the **Q** matrices at two flanking positions λ_{l} and λ_{r}, (cf., Eq. 3)

$$\mathbf{Q}_{\lambda_{\mathrm{qtl}}} = \left(\left(\lambda_{\mathrm{qtl}} - \lambda_{\mathrm{l}}\right)\mathbf{Q}_{l} + \left(\lambda_{\mathrm{r}} - \lambda_{\mathrm{qtl}}\right)\mathbf{Q}_{r}\right) / \left(\lambda_{\mathrm{r}} - \lambda_{\mathrm{l}}\right) \tag{6}$$

This implies that matrix $\mathbf{Q}_{\lambda}$ is fixed at any position along the genome and consequently the class indicators ($\mathbf{P}_{\mathrm{TIBD}}$) are also fixed in this threshold model. The calculation of $\mathbf{Q}_{\lambda}$ is performed at every sampling step in the Markov chain Monte Carlo (MCMC) simulation. The application of the weighted **Q** matrix (Eq. 6) to some initial datasets led to spurious results due to erroneous clustering of individuals based on the averaged probabilities. Another approach is to sample the **Q** matrix for a putative position λ_{qtl} between two flanking positions λ_{l} and λ_{r} as follows:

$$\mathbf{Q}_{\lambda_{\mathrm{qtl}}} = \begin{cases} \mathbf{Q}_{\mathrm{l}} & \text{with} \quad \text{Prob} = \left(\lambda_{\mathrm{qtl}} - \lambda_{\mathrm{l}}\right) / \left(\lambda_{\mathrm{r}} - \lambda_{\mathrm{l}}\right) \\ \mathbf{Q}_{\mathrm{r}} & \text{with} \quad \text{Prob} = \left(\lambda_{\mathrm{r}} - \lambda_{\mathrm{qtl}}\right) / \left(\lambda_{\mathrm{r}} - \lambda_{\mathrm{l}}\right) \end{cases} \tag{7}$$

This sampling approach seems more robust as it does not suffer the problem that occurs in the weighted-average implementation (Eq. 6). In addition, a computational advantage of the sampling-based implementation is that the PIBD matrices can be processed once prior to the MCMC simulation to cluster individuals, given the threshold value, and thus have crisp 0/1 matrix results available for all PIBD positions instead of the original IBD probability matrices.

Latent Ancestor Allele Model (LAAM)

This approach starts with a simple model with K disjoint latent classes in which each parent belongs to precisely one latent class, i.e., one latent ancestor allele. We extend this model with probabilities. Let **P** be an $n \times K$ matrix with elements p_{ik} being the probability that parent i belongs to class k.

$$p_{ik} \geq 0 \text{ and } \sum_{k=1}^{K} p_{ik} = 1(i = 1,\ldots,n; k = 1,\ldots,K). \tag{8}$$

By drawing the class memberships for each parent i from the ith row of **P** independently, the probability that parents i and j fall in the same class is

$$q^*_{ij} = \sum_{k=1}^{K} P(i \in \text{class}(k) \wedge j \in \text{class}(k)) = \sum_{k=1}^{K} p_{ik}p_{jk} \quad \forall i \neq j. \tag{9}$$

We have recently proposed several algorithms to find a matrix **P** such that q^*_{ij} is close to the observed Q_{ij} for all $i \neq j$ (ter Braak et al. (2009, 2010). For the **Q** matrix in Table 1A, the **P** matrix with five latent ancestor allele classes that corresponds with PIBD matrix **Q** (zero RMSE between **Q** and $\mathbf{Q}^*$) is also given (Table 1B).

In the Bayesian algorithm we need 0/1 matrices such as the one in Table 1D, expressing a unique assignment of each parent to a single ancestral allele. We can sample such matrices from this prior model by sampling for each parent i independently its ancestral allele (class membership) according to its (row-wise) probabilities $\{p_{ik}, k = 1,..,K\}$ in Table 1B.

The $\mathbf{P}_{\text{LAAM}}$ matrix pertaining to position λ_{qtl} may be calculated from the matrix $\mathbf{Q}_\lambda$, where $\mathbf{Q}_\lambda$ is obtained from Eq. 6. However, this is often computationally demanding within the MCMC simulation and the weighted-average $\mathbf{Q}_\lambda$ (6) may lead to unreliable analysis results. Similar to the sampling of $\mathbf{Q}_\lambda$ for $\mathbf{P}_{\text{TIBD}}$, the approach taken to compute the $\mathbf{P}_{\text{LAAM}}$ matrices follows (7) for all n_Q positions along the genome, i.e., a finite number, prior to the actual MCMC simulation. Let $\mathbf{P}_{\text{l}}$ and $\mathbf{P}_{\text{r}}$ be the **P** matrices pertaining to the flanking positions λ_{l} and λ_{r}. Then, the **P** matrix at the QTL position λ_{qtl}

$$\mathbf{P}_{\lambda_{\text{qtl}}} = \begin{cases} \mathbf{P}_{\text{l}} & \text{with} \quad \text{Prob} = (\lambda_{\text{qtl}} - \lambda_{\text{l}})/(\lambda_r - \lambda_{\text{l}}) \\ \mathbf{P}_{\text{r}} & \text{with} \quad \text{Prob} = (\lambda_{\text{r}} - \lambda_{\text{qtl}})/(\lambda_r - \lambda_{\text{l}}) \end{cases} \tag{10}$$

Sampling from P_{l} and P_{r} with these probabilities yields precisely the average IBD probabilities Q_λ of Eq. 6 if P_{l} and P_{r} perfectly fit Q_{l} and Q_{r}, respectively (see "Appendix A"). In the case of a perfect fit, the sampling approach of Eq. 10 is therefore equivalent to that of calculating the **P** matrix corresponding to $\mathbf{Q}_\lambda$ (6). In the case of a non-perfect fit, the two approaches are almost equivalent. Note that the latter sampling approach is not the same as sampling from an average of $\mathbf{P}_{\text{l}}$ and $\mathbf{P}_{\text{r}}$.

Effective number of latent ancestors

The prior model introduces correlation among the alleles of the parents because parents with similar rows in **P** are likely to be assigned as offspring from the same latent ancestor and will thus receive more often the same allele than under the independence model. Thus, ter Braak et al. (2010) also propose to use the effective number of latent classes as a measure for genetic diversity (see "Appendix B").

Bayesian Markov chain Monte Carlo QTL analysis

The utilization of PIBD data adds a new layer in the Bayesian hierarchical framework described by Bink et al. (2008) as presented in Fig. 2. We now present the linear model for the phenotypes and the joint posterior distribution of all random variables.

Data likelihood

The probability model for the trait phenotypes $(\mathbf{Y}_{\text{T}})$ is assumed to be

$$P(\mathbf{Y}_{\text{T}}|\mu, N_{\text{QTL}}, \mathbf{W}, \alpha) = N\left(\mathbf{1}\mu + \sum_{k=1}^{N_{\text{QTL}}} \mathbf{W}_k\alpha_k, \sigma^2_e\right), \tag{11}$$

where **1** is a unity vector pertaining to an overall mean effect (μ) and σ^2_e is the residual variance for the trait of interest; the incidence matrix **W** (see also "Appendix C") pertains to the QTL genotypes with additive effects α. Treating the number of bi-allelic QTL (N_{QTL}) as a random variable, the number of columns and the length of vector α are a priori unknown. Let $\theta = \{\mu, \alpha, \lambda, \mathbf{p}, \mathbf{q}, \sigma^2_e\}$, i.e., the set of location and dispersion variables given a particular number of QTL in the model. Vector **q** comprises the frequencies of marker alleles, which are included in the prior on the marker haplotypes (**H**) as shown in Fig. 2. Further details and prior assumptions on the linear model are as described by Bink et al. (2008) and are omitted here.

Joint posterior distribution

The probability model for the phenotypes and the prior distributions yield the joint posterior distribution. Let $\mathbf{Y}_P$ and $\mathbf{Y}_M$ denote the pedigree and marker data, respectively. The joint posterior distribution of all unknowns is now written as,

$$\begin{aligned} &P\left(\theta, N_{\mathrm{QTL}}, \mathbf{W}|\mathbf{Y}_T, \mathbf{Y}_P, \mathbf{Y}_M\right) \\ &\quad \propto P(\mathbf{Y}_T|\theta, N_{\mathrm{QTL}}, \mathbf{W})P(\mathbf{Y}_P|\mathbf{H}, \mathbf{W})P(\mathbf{Y}_M|\mathbf{H}) \\ &\qquad \times P(\mathbf{W}|p, \lambda, \mathbf{H})P(\mathbf{H}|q)P(\theta|N_{\mathrm{QTL}})P(N_{\mathrm{QTL}}) \end{aligned} \tag{12}$$

Note that the incidence matrix **W** is fully determined by variables **A**, **C**, and **S** ("Appendix C"). For MCMC simulation, the sampling distributions of the random variables are derived from this joint posterior distribution by conditioning on all other variables. These conditional distributions are as described by Bink et al. (2008), except for the QTL ancestor class indicator variables (**C**) that will be presented in the following.

Posterior conditional distributions of ancestral class indicators

In the Markov chain simulation the sampling distribution of the ancestral class indicators is derived by treating all other random variables in the joint posterior distribution as fixed. The posterior (12) then reduces to multiplying the likelihood function and prior on class indicators (3),

$$P(\mathbf{C}|\lambda, \mathbf{Y}_T, \mathbf{Y}_F) \propto P(\mathbf{Y}_T|\theta, N_{\mathrm{QTL}}, \mathbf{A}, \mathbf{C}, \mathbf{S})P(\mathbf{C}|\lambda, \mathbf{Y}_F) \tag{13}$$

In the TIBD model with the weighted **Q** (6) this sampling distribution is actually deterministic because of the conditioning on λ. Consequently, the ancestral class indicators are updated jointly with the position of the QTL. In the TIBD model with the sampled **Q** (7) we follow the same approach as presented below for the LAAM model.

In the LAAM model the sampling distribution of **C** is stochastic. We have implemented a Metropolis–Hastings updating algorithm for the joint sampling distribution of **C** and λ, i.e.,

$$P(\lambda, \mathbf{C}|\mathbf{Y}_F, \mathbf{Y}_T) \propto P(\mathbf{Y}_T|\theta, N_{\mathrm{QTL}}, \mathbf{A}, \mathbf{G}, \mathbf{S})P(\mathbf{G}|\lambda, \mathbf{Y}_T)P(\lambda) \tag{14}$$

where the algorithm starts with a normal random walk proposal distribution for λ and then proposes a **C**' from the prior as follows. With the probabilities given in (10) we sample either the left-flanking matrix $\mathbf{P}_\mathrm{l}$ or the right-flanking matrix $\mathbf{P}_\mathrm{r}$ of the new position and call the resulting matrix **P** with elements $\{p_{ik}\}$. We then sample for each parent i independently its ancestral allele (class membership) according to its (row-wise) probabilities $\{p_{ik}, k = 1, .., K\}$. The allele that is then proposed is the allele assigned to the ancestral class. Because we sample **C** from the prior and because the prior of λ is uniform, the acceptance ratio for the proposed values is then simply a ratio of the likelihoods of the current and proposed values.

The method of updating of the alleles of the ancestral classes is identical to that for updating alleles of parents in the model were parents are unrelated (UNR), which is Gibbs sampling. Note that the alleles of the parents are correlated in the TIBD and LAAM models because of the **P** matrix. Models TIBD and LAAM thus shift the independence assumption upward in the pedigree structure, namely from the parents to the latent ancestors.

Markov chain Monte Carlo simulation and posterior inference

The calculation of the above joint posterior distribution is analytically intractable, and we apply computer-intensive MCMC simulation (Gilks et al. 1996) to obtain draws from the joint posterior distribution. Different MCMC sampling algorithms are used, i.e., the Gibbs sampler (Gelman et al. 1995) when the conditional sampling distribution has a recognizable kernel and can directly be sampled from, and the Metropolis–Hastings algorithm (Gelman et al. 1995) when the conditional distribution cannot be sampled from directly. The sampling of ancestral class indicators under our new models TIBD and LAAM has been detailed above. To allow changes in model dimension, i.e., to increase or decrease the number of QTL in the model, we use the reversible jump MCMC method (Green 1995), similar to previous QTL model selection studies (Bink et al. 2002; Heath 1997; Sillanpaa and Arjas 1998). For each model we performed a Markov chain simulation of 500,000 (200,000) cycles for the pilot dataset (replicated datasets) and stored every 200th sample for posterior inference.

For all three models (UNR, TIBD, and LAAM), three values (1, 3, and 5) are evaluated for the mean of the Poisson distribution being the prior on the N_{QTL} in the analyses of our simulated data. The stored draws from the joint posterior distribution were used for posterior inference on the variables of interest, most importantly the characteristics of QTL (number, position, size, genotypes). A linkage group was divided into 1-centiMorgan (cM) bins and the number of QTL per bin per cycle was used to calculate the posterior QTL intensity (Sillanpaa and Arjas 1998).

For model selection in the pilot dataset we used Bayes factors (Kass 1993; Kass and Raftery 1995) as a measure of evidence coming from the data for different QTL models. More precisely, we used the statistic $2 \times \ln(BF)$ that scales similar to a LOD score test statistic (Kass and Raftery 1995).

In the replicated datasets we adopt the approach of Hayashi and Iwata (2009) to assess the power and accuracy of the three models. The posterior QTL intensity for 1-cM bins along the linkage groups was calculated. Subsequently, the Summed QTL Intensity (SQI) was calculated by summing the QTL intensity over a single linkage group (Hayashi and Awata 2008). Thresholds of SQI were determined from empirical null distributions of the maximum SQI over all linkage groups obtained from 100 null data sets (no QTL were modeled on any linkage group). When SQI exceeds these thresholds for any linkage group, detection of a QTL was declared. For declared QTL the position and effect were calculated as the weighted average over the linkage group where the weights were equal to the QTL intensity. We also examined an alternative method: to declare a QTL, a SQI threshold value of 0.50 must be exceeded (regardless of model), and the posterior mode estimate of QTL location (and the estimated QTL effect pertaining to that location mode) is used for every declared QTL. Taking SQI threshold values other than 0.50—we explored a range between 0.2 and 0.8—yielded similar patterns in relative performance among the three models (results not presented). Furthermore, the SQI approach works well for linkage groups with only 1 (or no) QTL but cannot be applied to a linkage group with multiple QTL as in the pilot dataset.

Simulated data

To empirically test our models we use one pilot dataset and 100 replicated datasets with the same pedigree data and marker densities but with different trait architectures.

Pedigree of connected mapping populations

The simulated data set mimics a real QTL mapping experiment from an ongoing maize breeding program. That is, 16 inbred parents were crossed in an incomplete mating design to produce 30 mapping populations of Recombinant Inbred Lines (RILs) (Table 2). The number of crosses per parent ranged from 1 (parent *863*) to 8 (parent *773*) and the number of RILs per population ranges from 4 to 50. The entire dataset contains 1,072 RILs.

Table 2 Simulated QTL genotypes and estimated posterior genotype probabilities in the pilot dataset for map intervals with positive QTL evidence of the 16 mapping parents

Parent	No. of crosses	No. of progeny	Simulation						Model[a]					
			UNR						TIBD			LAAM		
			Position (cM)			Intervals (cM)			Intervals (cM)			Intervals (cM)		
			30	60	140	28–37	58–64	134–143	25–32	57–61	136–142	26–33	58–62	138–143
761	4	83	1	0	1	0.3	0.6	0.5	**0.9**	**0.1**	**1.0**	0.7	**0.1**	**1.0**
766	2	59	0	0	0	0.3	0.3	**0.1**	**0.1**	**0.0**	0.2	**0.1**	**0.0**	0.2
773	8	248	1	0	1	0.3	0.5	0.8	**0.9**	**0.1**	**1.0**	0.7	**0.1**	**1.0**
775	3	145	0	1	1	0.3	0.8	0.7	**0.1**	**1.0**	**1.0**	**0.1**	**0.9**	**1.0**
822	2	97	1	0	1	0.3	0.7	0.7	**0.9**	**0.0**	**1.0**	**0.9**	**0.0**	**1.0**
847	4	197	0	0	0	**0.1**	0.4	**0.1**	**0.1**	**0.0**	**0.0**	**0.1**	**0.0**	**0.0**
851	6	230	1	0	0	0.8	0.3	**0.1**	**1.0**	**0.0**	**0.0**	**0.9**	**0.0**	**0.0**
853	5	190	1	0	0	0.4	**0.1**	**0.1**	**0.9**	**0.0**	**0.0**	**0.9**	**0.0**	**0.0**
855	4	119	0	0	1	0.4	0.4	0.3	**0.1**	**0.0**	**1.0**	**0.1**	**0.0**	**1.0**
857	7	283	0	1	1	0.2	0.8	**0.9**	**0.1**	**1.0**	**1.0**	**0.1**	**0.9**	**1.0**
859	3	100	1	0	0	0.3	0.5	**0.1**	**0.9**	**0.1**	**0.0**	**0.9**	**0.1**	**0.0**
861	3	69	1	0	1	0.5	0.8	0.8	**0.9**	0.8	**0.9**	**0.9**	0.2	**0.9**
863	1	40	1	1	0	0.7	0.8	**0.1**	**1.0**	**1.0**	**0.0**	**0.9**	**0.9**	**0.0**
865	2	23	0	1	0	0.7	0.7	**0.1**	**0.1**	**1.0**	**0.0**	**0.1**	**0.9**	**0.0**
867	2	82	1	0	0	0.4	0.6	**0.1**	**0.9**	**0.0**	**0.0**	0.8	**0.1**	**0.0**
869	4	179	1	1	0	0.8	0.8	**0.1**	**0.9**	**1.0**	**0.0**	**0.9**	**1.0**	**0.0**
Average $\lvert P_{true} - P_{est}\rvert$						0.44	0.39	0.18	0.08	0.08	0.04	0.11	0.06	0.05

The absolute difference between true (simulated) and estimated QTL genotype is given as the parents' average

The posterior probabilities pertain to the increasing QTL genotype, denoted "*QQ*" or "++". The probabilities for the decreasing QTL genotype, denoted "*qq*" or "−−" is equal to one minus the printed probability. The probabilities that are smaller than or equal to 0.1 and probabilities that are larger than or equal to 0.9 are bold printed

The map regions with positive QTL evidence are intervals with bins having a 2ln(BF) > 2. These regions are also depicted in Fig. 4

[a] Model abbreviations: *UNR* model with unrelated parents (without Ancestral Classes), *TIBD* model with ancestral allele classes and Threshold sampling approach, *LAAM* model with ancestral allele classes and Latent Ancestor Allele approach

Pedigree of ancestral generations

The known ancestral pedigree of the 16 parents of the connected mapping populations contains 162 inbred lines. Of these, 32 are founders, i.e., their parentages (pedigree) are assumed unknown.

Marker data

We simulated genetic data for 32 independent founders and their descendants according to the known (ancestral and mapping) pedigree structure. The pedigree contains multiple loops and the longest lineage is nine generations for any of the resulting 16 inbred parents. A gene-dropping method (Maccluer et al. 1986) was used to simulate Mendelian inheritance of marker and QTL alleles from parents to offspring while Haldane's mapping function (Haldane 1919) was used to transform linkage distances into recombination fractions.

Pilot dataset One linkage group of 150 cM was simulated that was covered by 16 equidistantly spaced bi-allelic SNP markers for the mapping populations (=sparse map). In contrast, the 16 parents and their ancestors were genotyped for 151 markers covering the same genome length (1 cM distance, = dense map).

Replicated datasets Five linkage groups of 100 cM were simulated, each group covered by 11 and 101 equidistantly spaced bi-allelic SNP markers for the mapping populations and parents and ancestral pedigree members, respectively. So, the total numbers of SNP markers were 55 and 505 SNPs for the two subsets of the pedigree.

Phenotypic trait data

Pilot dataset The phenotypes of all mapping individuals were simulated by assuming three QTL residing at positions 30, 60, and 140 cM on the linkage group. The distance between the first and second QTL was relatively small to assess differences in detection power and accuracy of estimates of closely linked QTL. The size of the additive effect for all of the three simulated QTL was set to 1.0.

Replicated datasets Linkage groups 1–4 contained a single QTL at positions 22, 44, 66, and 88, respectively. Linkage group 5 did not harbor a QTL and was included to evaluate the false positives rate. The size of the additive effects for the QTL at the subsequent linkage groups were 1.3, 1.1, 0.9, and 0.7, respectively.

In both datasets, the allele frequency of all QTL was 0.5 in the ancestral founder population and linkage equilibrium among all loci was assumed. No other genetic or non-genetic variables were simulated and the residual variance was equal to 16.0. In the pilot dataset, the realized phenotypic variance was equal to 19.3 and the heritability of each of the three QTL was approximately 5%. The phenotypic distributions (not shown) for the whole population and the individual mapping populations were continuous and uni-modal as expected from the relatively large residual variance. The QTL allele frequency of 0.5 in the founders did not yield the same frequency in the mapping populations (or their parents) due to drift in the pedigree ancestral to the mapping parents. For the replicated datasets we calculated the fraction of segregation in the n_j mapping populations by

$$F_{\text{Segr}} = \left(\sum_{j=1}^{n_j} n_o[j] \times s_o[j] \right) \bigg/ n_O \qquad (15)$$

where $s_o[j]$ is a binary indicator whether the jth mapping population is segregating ($s_o = 1$) or not ($s_o = 0$). The segregation indicator is weighted by the size of the mapping population ($n_o[j]$) and the total number of mapping offspring is given by n_O (defined previously).

Parent IBD data

For every marker position on the 1 cM map, IBD probabilities among the 16 inbred parents were calculated by using the FlexQTL software (Bink et al. 2008). Note that these 16 × 16 PIBD matrices were calculated (and stored) only once for a given dataset regardless of the number of traits to be analyzed or model specification in the analysis. The resulting 151 (505) PIBD matrices of the pilot dataset (replicated datasets) were used as the new additional data source in the Bayesian analysis (data $\mathbf{Y}_D$ in Fig. 2). In the LAAM approach we used the least-squares approximation to obtain Latent Class probabilities allowing a maximum of 16 classes (ter Braak et al. 2010).

Results

Pilot dataset

The effective number of ancestral classes (see "Appendix B") was computed for the TIBD and LAAM models and significant variation in this number was observed. That is, the mean (standard deviation) for the TIBD and LAAM models were 3.4 (0.96) and 3.1 (0.76), respectively. The lowest number was 1.6, implying a substantial probability (0.625) that two randomly chosen individuals belong to the same ancestral allele class. The highest number was more than 6. The effective number in the LAAM model was always smaller or equal to the effective number in the

TIBD model, indicating that parents in the LAAM model may have a higher probability of sharing the same ancestral allele.

The posterior mean estimates for the overall mean and residual variance component for the UNR model deviated more from the simulated values than those from the TIBD and LAAM models, irrespective of the prior value for the number of QTL (Table 3). In all three models, the posterior estimates for the number of QTL and the total QTL variance (and consequently heritability) were inflated for higher mean values of the prior number of QTL. The Bayes Factors estimates correctly identified the proper QTL numbers for the TIBD and LAAM models, i.e., large Bayes Factor values (>20) for model with 3 QTL over a model with 2 QTL. For the UNR model the Bayes Factor estimate for the 3 over 2 QTL model was only moderate (values between 3.1 and 4.1).

The peaks of the posterior intensity profiles for QTL position were in general close to the simulated values (30, 60, and 140 cM) for all three models (Fig. 3). The intensity profiles were clearly more peaked and higher for models TIBD and LAAM (Fig. 3). In these models, the intensity peaks were often somewhat left-shifted from the simulated QTL positions, especially for the QTL at 60 cM. For model UNR the intensity profile near the first two simulated QTL positions was relatively flat and stretched as these two QTL were closely linked. The value of the prior number of QTL seemed most influential on the QTL intensity profile for the UNR model (Fig. 3) where intensity profiles were lifted near QTL 1 and 2 while the profile broadened near QTL 3. These posterior intensity plots pointed to an increased accuracy in positioning the QTL on the linkage groups when adding ancestral information into the analysis.

Bin-wise Bayes Factor estimates with a cut-off value of 2.0 are used to identify the most probable QTL regions. For each bin in these QTL regions, the estimates of the mean and 90% probability credible regions for QTL effect size are shown in Fig. 4 (for prior number of QTL equal to 3). The posterior mean estimates for all three models were very close to the simulated value of 1.0. The credible regions were smaller and more accurate in the TIBD and LAAM models than in the UNR model.

For these most probable QTL regions we computed the posterior probabilities for QTL genotypes for the 16 parents of the mapping population (Table 2). The posterior probability estimates were often inconclusive ($0.3 \leq P \leq 0.7$) for the UNR model, especially for the QTL at 30 and 60 cM, while for the QTL at 140 cM ten out of 16 parents had conclusive probability estimates ($0.1 \leq P$ or $P \geq 0.9$). On the other hand, the genotype probability estimates for the TIBD and LAAM models were almost always conclusive for

Table 3 Posterior mean estimates for overall mean (μ), residual variance (σ_e^2), the number of QTL (N_{QTL}), the QTL variance (σ_{QTL}^2), and heritability (h^2) for the pilot dataset

Variable	μ	σ_e^2	N_{QTL}	σ_{QTL}^2	h^2	2*ln* (Bayes factor)[a]			
						1/0	2/1	3/2	4/3
Simulation	0	16	3	3	0.16				
Model[b]									
Prior(N_{QTL}) = 1									
UNR	1.1	16.6	3.3	3.9	0.19	na	26	3.7	1.7
TIBD	−0.1	16.2	3.6	3.5	0.18	na	na	29	1.7
LAAM	0.0	16.2	3.7	3.6	0.18	na	na	28	2.3
Prior(N_{QTL}) = 3									
UNR	0.9	16.5	4.8	4.2	0.20	na	9.5	3.1	1.5
TIBD	0.0	16.1	4.6	4.0	0.20	na	na	24	1.6
LAAM	0.1	16.1	4.9	4.4	0.21	na	na	24	2.2
Prior(N_{QTL}) = 5									
UNR	0.8	16.4	6.1	4.3	0.21	na	na	4.1	1.6
TIBD	0.1	16.1	5.7	4.5	0.22	na	na	21	2.0
LAAM	0.0	16.1	5.7	4.7	0.23	na	na	21	1.6

For model selection estimates for twice the natural log of Bayes Factors are given and three values (1, 3 and 5) for the prior mean of number of QTL are used

na not available due to insufficient number of posterior samples from one or both models to estimate Bayes Factor accurately

[a] 2ln(Bayes Factor) : Bayesian statistic representing the evidence for favoring model M_q over M_{q-1} where q represents the number of QTL in the model ($q = 1, 2, 3, 4$)

[b] Model abbreviations: *UNR* model with unrelated parents (without Ancestral Classes), *TIBD* model with ancestral allele classes and Threshold sampling approach, *LAAM* model with ancestral allele classes and Latent Ancestor Allele approach

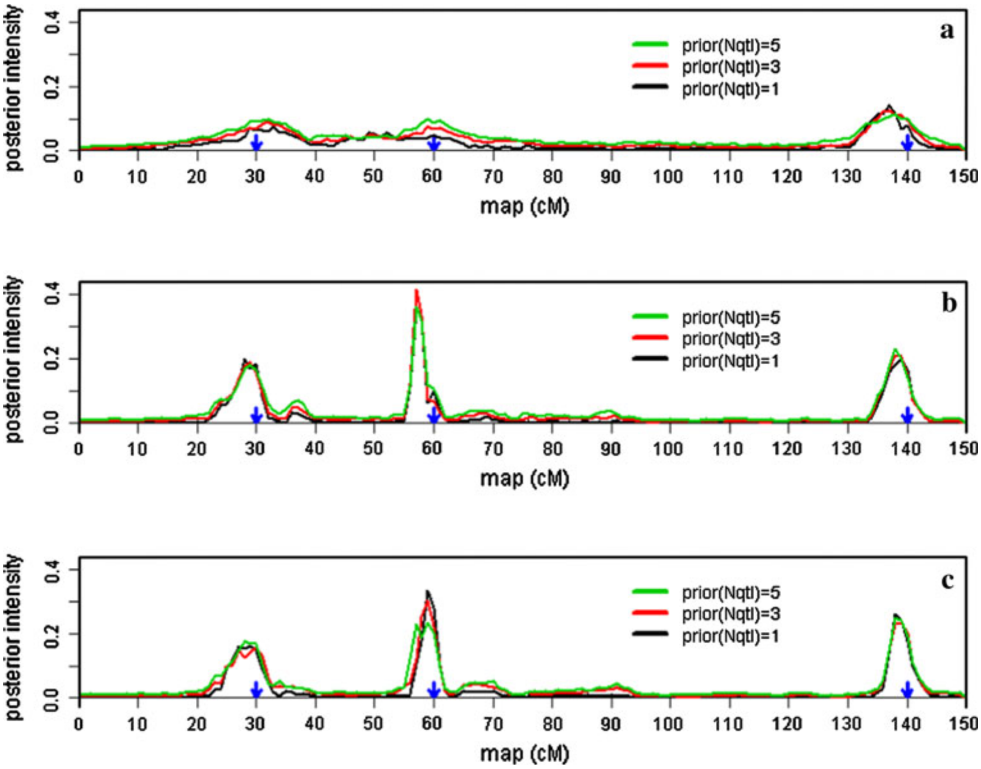

Fig. 3 Marginal posterior intensity profile estimates for QTL position along the 150 cM linkage map for three types of analyses, i.e., **a** unrelated (UNR) mapping parents; **b** related mapping parents with Threshold IBD (TIBD) algorithm; and **c** related mapping parents with Latent Ancestor Allele Model (LAAM) algorithm. The QTL positions of the simulated dataset are indicated by arrows on the *x*-axis (30, 60, and 140 cM). Three values (1, 3, and 5) for the prior number of QTL were studied

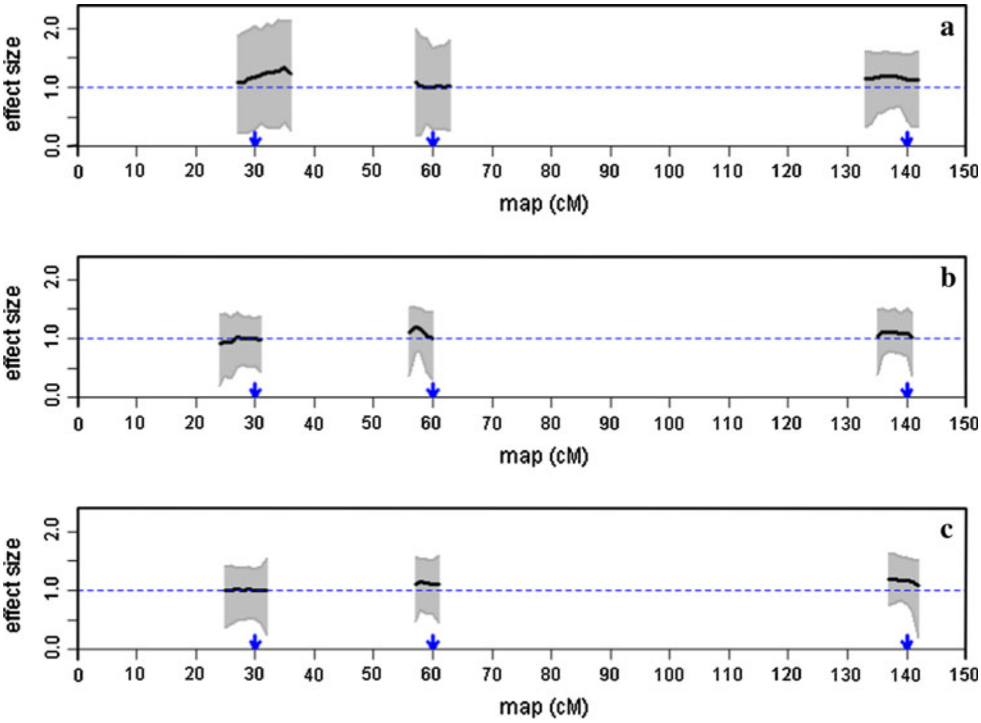

Fig. 4 Marginal posterior mean (*solid black line*) and 0.90 probability credible region estimates (*gray surfaces*) for QTL effect size along the 150 cM linkage map for three types of analyses, i.e., **a** unrelated (UNR) mapping parents; **b** related mapping parents with Threshold IBD (TIBD) algorithm; and **c** related mapping parents with Latent Ancestor Allele Model (LAAM) algorithm, and prior $E(N_{\mathrm{QTL}} = 3)$. Estimates are depicted only for areas with positive QTL signal ($2 \times \ln(\text{Bayes Factor}_{\text{bin}}) \geq 2.0$). The QTL positions of the simulated dataset are indicated by arrows on the *x*-axis (30, 60, and 140 cM). The *dashed horizontal line* indicates the simulated effect size

all three QTL positions (Table 2). These differences between models were also summarized by calculating the average absolute difference between the true and estimated QTL genotype probabilities. These summary values show that including PIBD information leads to four times more accurate genotype probability estimates.

Replicated datasets

The fraction of segregation in the mapping populations for the 100 replicated datasets is presented in Fig. 5. Note that the values were ordered within each QTL and ranged from 0.0 to 0.76. When the fraction was equal to zero the QTL was fixed for one of the alleles and consequently the QTL could not be detected in that particular replicate. For the QTL on linkage groups 1–4 the number of non-segregating QTL was equal to 18, 8, 12, and 13, respectively (Table 4), implying that for example more QTL could be detected for linkage group 2 than for linkage group 1.

Summed QTL Intensity threshold based on empirical null distribution

For the prior $E(N_{QTL} = 1)$, the 5% significance level of these empirical distributions were 0.127, 0.206, and 0.196 for the models UNR, TIBD, and LAAM, respectively. A plausible cause of the higher threshold for the TIBD and LAAM models is the following. Some (segments of) linkage groups are not segregating in the mapping parents (similar to fixed QTL in Fig. 5). These monomorphic regions are excluded in the TIBD and LAAM models which increases the prior probability for QTL on other linkage groups. This creates a higher variability in SQI among linkage groups and thus higher values for the maximum SQI of the linkage groups. The power of QTL detection was higher for the UNR model than for the other two models when using the SQI threshold based on 100 null datasets, except for the QTL on linkage group 3 (Table 4). The UNR model also yielded a much higher false discovery rate, i.e., 37 QTL were declared for linkage group 5. The posterior estimates for location were most biased for the UNR model and for the QTL at the extremes of the linkage groups. The bias is pointing to the middle of the linkage group and seems to be caused by the estimation protocol of Hayashi and Iwata (2009). Since all intervals along the linkage group are included in the estimation procedure, QTL at the extremes will have their estimated location biased toward the center of the linkage group. Especially for the QTL with smaller effect size the location cannot be precisely determined and positions further away are plausible as well. This bias in estimates of QTL location may be strongly reduced by considering the mode of the posterior mode estimates. For example, for the QTL position at the 4th linkage group the estimated mode was equal to 89 cM in all scenarios and both thresholds (results not shown). The accuracy, as represented by the standard deviation of QTL location, was always lower for the UNR model than for the TIBD and LAAM models. The accuracy of location estimates decreased (standard deviation estimates increased) for smaller QTLs (Table 4). The TIBD and LAAM models yielded very similar results for power and accuracy. The effect sizes were always underestimated for all three models, more severely for the UNR model, and this was also likely due to the estimation procedure of averaging along the whole linkage group.

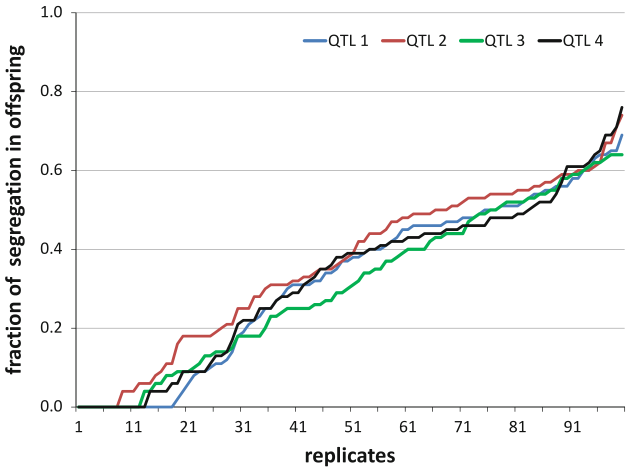

Fig. 5 Fraction of QTL segregation in offspring populations in the 100 replicated simulations (replicates in ascending order within QTL)

SQI threshold equal to 0.5

Relative to the empirical null distribution threshold, the SQI > 0.5 threshold decreased the power of QTL detection for all three models (Table 5). The UNR model yielded lowest power and the power decreased consistently with QTL effect size. The highest power for the QTL on linkage group 2 for the TIBD and LAAM models can be explained by the highest number of replicates with segregation (92). The posterior mode estimates for QTL location yielded almost unbiased results, except for the smallest QTL (linkage group 4). Also, the estimates of QTL effects at the mode of QTL location were close to the simulated values for the models TIBD and LAAM, especially for the larger QTL effect sizes. The results for the UNR model were always inferior, especially for accuracy of the estimates as expressed by the standard deviations in Table 5. These latter results indicated that the original SQI protocol of Hayashi and Iwata (2009) may be further improved and extended to compare relative performance of different methods and models with respect to power and accuracy of QTL mapping for complex traits.

Discussion

We present a novel approach to efficiently include genome-wide ancestral IBD information on parent alleles into the QTL analyses of multiple connected populations. Analysis of simulated data indicates improvement of mapping

Table 4 Posterior inferences results on replicated datasets (100 replicates), using the SQI thresholds from 100 null datasets

	LG 1	LG 2	LG 3	LG 4	LG 5
Simulation					
Position	22.0	44.0	66.0	88.0	
Effect	1.30	1.10	0.90	0.70	
Segregation[a]	82	92	88	87	0
UNR					
Power[b]	81	88	69	67	37[e]
Location[c]	29.5 (9.7)	45.8 (9.4)	55.2 (11.5)	60.0 (17.8)	
Effect[d]	0.86 (0.49)	0.55 (0.43)	0.35 (0.46)	0.24 (0.39)	
TIBD					
Power	77	83	69	54	7[e]
Location	25.5 (5.5)	43.5 (6.6)	61.8 (9.9)	75.4 (13.0)	
Effect	1.10 (0.25)	0.90 (0.24)	0.73 (0.34)	0.58 (0.32)	
LAAM					
Power	77	83	71	55	10[e]
Location	25.4 (6.2)	43.2 (6.9)	61.8 (9.3)	75.0 (14.3)	
Effect	1.10 (0.25)	0.91 (0.25)	0.72 (0.34)	0.59 (0.32)	

[a] Number of replicates with QTL segregation

[b] Significance threshold: 95% quantile of SQI (summed QTL intensity) from 100 null datasets. Threshold values were 0.127,0.206, and 0.196 for UNR, TIBD and LAAM respectively

[c] Mean (standard deviation) of the estimated posterior mean of QTL location

[d] Mean (standard deviation) of the estimated posterior mean of QTL effect across genome

[e] Type I error as no QTL was simulated on Linkage Group 5

Table 5 Posterior inferences results on replicated datasets (100 replicates), using the same SQI threshold of 0.50 for all scenarios of analysis

	LG 1	LG 2	LG 3	LG 4	LG 5
Simulation					
Position	22.0	44.0	66.0	88.0	
Effect	1.30	1.10	0.90	0.70	
Segregation[a]	82	92	88	87	n.r.
UNR					
Power[b]	66	63	29	21	2[e]
Location[c]	22.4 (7.7)	44.9 (17.0)	64.3 (14.3)	78.5 (30.2)	
Effect[d]	1.40 (0.23)	1.16 (0.29)	1.27 (0.31)	1.22 (0.46)	
TIBD					
Power	73	77	56	43	1[e]
Location	22.0 (4.8)	42.3 (7.4)	64.4 (11.2)	80.6 (19.1)	
Effect	1.31 (0.21)	1.11 (0.21)	1.01 (0.29)	0.89 (0.31)	
LAAM					
Power	73	77	56	43	1[e]
Location	21.8 (4.1)	42.1 (7.8)	63.6 (11.2)	80.4 (22.2)	
Effect	1.30 (0.20)	1.12 (0.18)	1.02 (0.29)	0.89 (0.28)	

[a] Number of replicates with QTL segregation

[b] Significance threshold: SQI (summed QTL intensity) >0.50

[c] Mean (standard deviation) of the estimated posterior mode of QTL location

[d] Mean (standard deviation) of the estimated QTL effect at posterior mode of QTL location

[e] Type I error as no QTL was simulated on Linkage Group 5

accuracy and power when genetic relationships between parents are modeled as opposed to treating the parents as independent. Two algorithms were implemented and tested. The threshold-algorithm benefits from ease of implementation and interpretation but may yield a crude classification of founder alleles, especially when PIBD probabilities are more intermediate between 0 and 1. Furthermore, consistency of classification needs to be checked via transitivity rules. The latent class algorithm is conceptually more appealing as it provides a more precise representation of the original IBD information along the genome. In our simulated datasets these two algorithms yielded the same posterior conclusions. Our current results indicate that the threshold values of 0.90 and 0.80 in the TIBD model yield very similar posterior results and mixing behavior; however, results and performance may become different with further lowering of this threshold. These implementation issues of the TIBD model are subject to further research.

The comparison of our proposed PIBD approach to a full pedigree analyses was impractical as the high marker density in the ancestral pedigree (Fig. 1) creates a major missing data problem in the mapping pedigree. The progeny in the mapping populations in the pilot dataset would have missing marker scores for 135 out of 151 loci. A comparison of our novel approach with a full pedigree analysis with all individuals genotyped for the sparse density map showed that the full pedigree analysis was almost as powerful, but that computation time was dramatically increased because of the added number of individuals and the additional number of generations in the pedigree which makes the sampling algorithms more time consuming. Even in a much smaller simulated example, i.e., considering a single biparental mapping population, the full pedigree analysis required over 50 times more computation time (results not shown).

The novel approach to include genome-wide ancestral IBD information in QTL mapping can be further extended to include a polygenic component which may account for QTL that cannot be picked up in the linkage detection, cf. Bink et al. (2008). When modeling the polygenic component, the use of a marker-based genome-wide average coancestry among founder individuals could be obtained from the PIBD matrices calculated for each chromosomal segment, cf. (NejatiJavaremi et al. 1997), and recently applied to genomic selection, e.g., (Habier et al. 2007). An alternative to this approach would be to use known pedigree relationships to construct the coancestry relationship matrix needed to account for the polygenic term.

In this study, we assume two alleles at a QTL which allows a straightforward extension to include non-additive effects, e.g., dominance and epistatic interactions. When primary interest is in additive gene actions, QTL models with many alleles may be advantageous to allow greater flexibility for panmictic populations (Hoeschele et al. 1997). However, two important implementation issues must be addressed. First, a multiple allele model may contain effects for alleles with little supporting phenotypic data and is thereby prone to less accurate results for QTL allelic effects (Hayashi and Iwata 2009). To draw accurate inference on the number of allelic effects, the allelic effects must differ substantially from each other (Jannink and Wu 2003). Second, the extension to dominance and higher-order interactions is not straightforward as many interaction effects will not be realized in the phenotypic data. The extension of our new approach to outbred populations will be straightforward in case dense marker data are available to unambiguously assign haplotypes to all parents of the mapping population. Then the dimensionality of the PIBD matrices simply doubles and the number of rows in matrix **P** also doubles. In our study we had access to accurate pedigree and marker data on ancestors of the mapping populations. When ancestral pedigree is unknown or DNA is not available on the members, the LD-based estimation method of Meuwissen and Goddard (2000) utilizing very dense marker data can be applied to obtain the location-specific parental IBD matrices in outbred and inbred populations (Bink and Meuwissen 2004).

The UNR model was the point of departure in our Bayesian approach and this model already accounts for the sharing of one or more common parents by multiple populations using a pedigree linkage approach. Other recent Bayesian approaches have been proposed that take unique QTL allelic effects for each of the mapping parents (Hayashi and Iwata 2009) with their simulated datasets containing a common reference parent in a star design. Fang et al. (2011) assumed the QTL alleles of all mapping individuals as samples from a normal distribution with a covariance matrix proportional to the IBD matrix that was calculated from the marker information on the mapping offspring and their parents. The modeling of the QTL as a random effect in a mixed model can be solved more efficiently using restricted maximum likelihood approaches. Mixed models for QTL mapping in real connected plant populations have been successful in wheat (Arbelbide and Bernardo 2006; Crepieux et al. 2005; Rosyara et al. 2009) and maize (van Eeuwijk et al. 2010), but the treatment of multiple QTL models is less straightforward for (closely) linked QTL as was the case in our pilot dataset.

The simulated datasets used in this study reflect a typical connected population structure as they contained 30 mapping populations derived from 16 connected parents with known ancestry up to 32 original founder individuals. These characteristics can easily be varied without changing the applicability of the method. For example, the idea of ancestral allele classes can also be applied to a single mapping population derived from two inbred parents.

In that case, the ancestral PIBD data will indicate which genomic regions are shared by the two inbred parents and these regions can be excluded a priori to harbor QTL. This type of information may substantially increase mapping precision but is fully ignored in other linkage methods. Furthermore, plant breeders may consider a large number of small mapping populations derived from a large number of parents where these parents inherited a limited number of (unknown) ancestral alleles. The additional layer of ancestral allele classes will facilitate substantial power to associate phenotypic trait variation with genomic polymorphisms. The increase in numbers may require more efficient algorithms to include ancestral IBD information. The practicality of our new approach was well illustrated by the successful mapping of QTLs in hybrid selection programs (van Eeuwijk et al. 2010).

The increasing availability of cheap and abundant markers opens new ways to advance genetic progress in plant and animal breeding programs, such as whole genome selection approaches (Bernardo and Yu 2007; Meuwissen et al. 2001). However, the application of high-density (SNP) genotyping to all mapping populations grown within commercial breeding programs might still not be feasible due to economic reasons. Therefore, a substantial discrepancy in marker density between elite (selected) breeding lines and regular breeding populations can occur. Our approach tackles this potential discrepancy and exploits the available sources of information efficiently to map important genomic regions affecting complex traits.

Acknowledgments We acknowledge useful suggestions and comments of our colleagues at Biometris and Pioneer on this study and manuscript. The manuscript has greatly benefitted from the reviewers' comments.

Conflict of interest The authors declare no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix A

Sampling ancestral alleles from a weighted average of IBD matrices Q_l and Q_r. The average IBD matrix is

$$\mathbf{Q}_\lambda = \alpha\mathbf{Q}_\mathrm{l} + (1-\alpha)\mathbf{Q}_2 \quad \text{for} \quad 0 \le \alpha \le 1$$

with $\alpha = (\lambda_\mathrm{qtl} - \lambda_\mathrm{l})/(\lambda_\mathrm{r} - \lambda_\mathrm{l})$

Theorem If P_l perfectly fits Q_l ($\mathbf{Q}_\mathrm{l} = \mathbf{P}_\mathrm{l}\mathbf{P}_\mathrm{l}^T$ for the off-diagonals of $\mathbf{Q}_\mathrm{l}$) and P_r perfectly fits Q_r ($\mathbf{Q}_\mathrm{r} = \mathbf{P}_\mathrm{r}\mathbf{P}_\mathrm{r}^T$ for the off-diagonals of $\mathbf{Q}_\mathrm{r}$) then the following sampling rule gives samples that fit Q_λ:

Sample ancestral alleles with probability α from $\mathbf{P}_l$ and with probability $1-\alpha$ from $\mathbf{P}_r$

Proof

$$\begin{aligned}
&\Pr(i \text{ and } j \text{ fall in same class}|\text{sampling scheme})\\
&= \alpha\sum_{k=1}^{K_1}\Pr(i \in \text{class}(k) \wedge j \in \text{class}(k)|\mathbf{P}_l)\\
&\quad + (1-\alpha)\sum_{k=1}^{K_2}\Pr(i \in \text{class}(k) \wedge j \in \text{class}(k)|\mathbf{P}_r)\\
&= \alpha\sum_{k=1}^{K_1} p_{lik}p_{ljk} + (1-\alpha)\sum_{k=1}^{K_2} p_{rik}p_{rjk}\\
&= \alpha q_{lij} + (1-\alpha)q_{lij} = q_{\lambda ij} \quad \forall i \ne j
\end{aligned}$$

Note that this theorem cannot be stated in terms of an average of **P** matrices.

Appendix B

Effective number of ancestral allele classes

In the TIBD model the number of ancestral classes is known along the genome. That is, given the threshold value and matrix $\mathbf{Q}_\lambda$ pertaining to position λ the number of ancestral classes is fixed. However, the number of ancestral classes in the LAAM model is more difficult to infer. One approach is to determine for each λ the optimal number of ancestor classes by an approach as in (ter Braak et al. 2009, 2010). The approach we take here is to set the number of ancestral classes equal to its maximum (the number of individuals) and then use the **P** matrix to calculate the effective number of the latent ancestral classes as described by (Hill 1973).

$$n_\mathrm{eff} = 1\Big/\sum_{k=1}^{K}\left(p_{+k}/p_{++}\right)^2,$$

where + used as an index indicates the sum over the index; for example p_{+k} is the column sum. Note that $1/n_\mathrm{eff}$ is equal to Simpson's index of diversity (Hill 1973). The Simpson index of diversity can be interpreted in our context as the probability that two randomly chosen individuals belong to the same class. Note that this effective number of classes can be calculated for both the TIBD and LAAM models.

Appendix C

Construction of incidence matrix of QTL effects to phenotypes

Let **W** denote the incidence matrix that links the quantitative trait phenotypes of the offspring in the mapping

$$\mathbf{T}_S = \begin{array}{c|cccccc|} & I_1 & I_2 & I_3 & I_4 & I_5 & I_6 \\ \hline J_{1,1} & 1 & & & & & \\ J_{1,2} & 1 & & & & & \\ J_{1,3} & & 1 & & & & \\ \ldots & & & & & & \\ \ldots & & & & & & \\ J_{6,1} & & & & & & 1 \\ J_{6,2} & & & & & 1 & \\ \ldots & & & & & & \end{array} \quad \mathbf{T}_C = \begin{array}{c|cccccc|} & A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \\ \hline I_1 & 1 & & & & & \\ I_2 & 1 & & & & & \\ I_3 & & & 1 & & & \\ I_4 & & & 1 & & & \\ I_5 & & 1 & & & & \\ I_6 & & & 1 & & & \end{array} \quad \mathbf{T}_A = \begin{array}{c|cc|} & Q & q \\ \hline A_1 & 1 & \\ A_2 & & 1 \\ A_3 & 1 & \\ A_4 & & \\ A_5 & & \\ A_6 & & \end{array}$$

populations to the QTL effects in a linear model. This matrix **W** can be calculated by multiplication of three transition matrices, i.e.,

$$\mathbf{W} = \mathbf{T}_S \times \mathbf{T}_C \times \mathbf{T}_A \tag{16}$$

where $\mathbf{T}_S$ denotes the transition of parental alleles to offspring, $\mathbf{T}_C$ denotes the transition of ancestral alleles to parents, and $\mathbf{T}_A$ denotes the assignment of QTL alleles to specific ancestral classes. For the example for Fig. 2b, these matrices are

Note that alleles A_4, A_5, and A_6 in matrix $\mathrm{T_C}$ are included for completeness; they were not transmitted in the example of Fig. 2.

References

Arbelbide M, Bernardo R (2006) Mixed-model QTL mapping for kernel hardness and dough strength in bread wheat. Theor Appl Genet 112:885–890

Beavis WD, Grant D, Albertsen M, Fincher R (1991) Quantitative trait loci for plant height in four maize populations and their associations with qualitative genetic loci. Theor Appl Genet 83: 141–145

Bernardo R, Yu JM (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082–1090

Bink MCAM, Meuwissen THE (2004) Fine mapping of quantitative trait loci using linkage disequilibrium in inbred plant populations. Euphytica 137:95–99

Bink MCAM, Uimari P, Sillanpaa J, Janss LLG, Jansen RC (2002) Multiple QTL mapping in related plant populations via a pedigree-analysis approach. Theor Appl Genet 104:751–762

Bink MCAM, Boer MP, ter Braak CJF, Jansen J, Voorrips RE, van de Weg WE (2008) Bayesian analysis of complex traits in pedigreed plant populations. Euphytica 161:85–96

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor Appl Genet 113:206–224

Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A Mixed-Model Quantitative Trait Loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics 177:1801–1813

Crepieux S, Lebreton C, Flament P, Charmet G (2005) Application of a new IBD-based QTL mapping method to common wheat breeding population: analysis of kernel hardness and dough strength. Theor Appl Genet 111:1409–1419

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theor Popul Biol 23:34–63

Falconer DS (1989) Introduction to quantitative genetics, 3rd edn. Longmans Green/John Wiley & Sons, Harlow, Essex

Fang M, Liu J, Sun D, Zhang Y, Zhang Q, Zhang S (2011) QTL mapping in outbred half-sib families using Bayesian model selection. Heredity 107:265–276

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman & Hall, London

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain monte carlo in practice. Chapman & Hall, London

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hayashi T, Awata T (2008) A Bayesian method for simultaneously detecting Mendelian and imprinted quantitative trait loci in experimental crosses of outbred species. Genetics 178:527–538

Hayashi T, Iwata H (2009) Bayesian QTL mapping for multiple families derived from crossing a set of inbred lines to a reference line. Heredity 102:497–505

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

Hill MO (1973) Diversity and evenness—unifying notation and its consequences. Ecology 54:427–432

Hoeschele I, Uimari P, Grignola FE, Zhang Q, Gage KM (1997) Advances in statistical methods to map quantitative trait loci in outbred populations. Genetics 147:1445–1457

Jannink JL, Wu XL (2003) Estimating allelic number and identity in state of QTLs in interconnected families. Genet Res 81:133–144

Kass RE (1993) Bayes factors in practice. Statistician 42:551–560

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795

Lander ES, Green P (1987) Construction of multilocus genetic-linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Maccluer JW, Vandeberg JL, Read B, Ryder OA (1986) Pedigree analysis by computer-simulation. Zoo Biology 5:147–160

Meuwissen TH, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421–430

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

NejatiJavaremi A, Smith C, Gibson JP (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci 75:1738–1745

Rosyara UR, Gonzalez-Hernandez JL, Glover KD, Gedye KR, Stein JM (2009) Family-based mapping of quantitative trait loci in plant breeding populations with resistance to fusarium head blight in wheat as an illustration. Theor Appl Genet 118: 1617–1631

Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388

ter Braak CJF, Kourmpetis Y, Kiers HAL, Bink MCAM (2009) Approximating a similarity matrix by a latent class model: a reappraisal of additive fuzzy clustering. Comput Stat Data Anal 53:3183–3194

ter Braak CJF, Boer MP, Totir LR, Winkler CR, Smith OS, Bink MCAM (2010) Identity-by-descent matrix decomposition using latent ancestral allele models. Genetics 185:1045–1057

van Eeuwijk FA, Boer M, Totir LR, Bink M, Wright D, Winkler CR, Podlich D, Boldman K, Baumgarten A, Smalley M, Arbelbide M, ter Braak CJF, Cooper M (2010) Mixed model approaches for the identification of QTLs within a maize hybrid breeding program. Theor Appl Genet 120:429–440