Mapping quantitative trait loci in plant breeding populations

Jansen, Ritsert; Jannink, Jean-Luc; Beavis, William D.

Published in:
Crop Science

DOI:
10.2135/cropsci2003.0829

Publication date:
2003

Mapping Quantitative Trait Loci in Plant Breeding Populations: Use of Parental Haplotype Sharing

Ritsert C. Jansen,* Jean-Luc Jannink, and William D. Beavis

ABSTRACT

Applied breeding programs evaluate large numbers of progeny derived from multiple related crosses for a wide range of agronomic traits and for tens to hundreds of molecular markers. This study was conducted to determine how these phenotypic and genetic data could be used for routinely mapping quantitative trait loci (QTLs). With dense maps, haplotype sharing of parents in a certain region is a good indicator for QTL-allele sharing, albeit not 100% perfect. With this in mind, an approximate and simple method has been developed where ancestral genome blocks in the parents of the crosses can be identified via haplotype analysis and where the effect of a putative QTL is then modeled and estimated per ancestral genome block. A simulation of an early-generation maize breeding scheme demonstrates the potential of the present approach for QTL detection in existing breeding programs. With this new QTL mapping strategy, the power, precision, and accuracy associated with large numbers of progeny may be attained, inferences about QTLs can be drawn across the breeding program rather than being limited to the sample of progeny from a single cross, and results may be much more valuable for marker-assisted breeding because the QTLs apply to agronomically challenging situations in the field.

In the practice of breeding for agronomically important crops such as maize (*Zea mays* L.) and soybean [*Glycine max* (L.) Merr.], the breeder annually generates many crosses. Typically, a few elite inbred lines or varieties are crossed with a wide range of new inbred lines or varieties to generate a large number of segregating crosses. For obvious reasons, the number of progeny per cross is small (often around 10, seldom more than 50), but the total number of progeny tested is relatively large. A commercial maize breeder, for example, may evaluate 1 000 to 10 000 F_3 topcrossed progeny derived from 100 to 200 crosses in replicated field trials. It is not unrealistic to assume that plant-breeding populations will be fingerprinted on a regular basis at 200 to 500 marker loci, and with chip technology soon at 1 000 or more loci. Beavis (1998) advocated integration of QTL mapping into existing breeding strategies. During the past years, human and animal geneticists have developed sophisticated methods for linkage and association analysis; see for instance Yi and Xu (2001) for an exact pedigree method and Meuwissen and Goddard (2000) for an advanced haplotype-based association method, and Jannink et al. (2001) for a recent review. These methods can be used after some modifications for analysis of plant breeding data. Recently, Jannink and Jansen (2001) developed multiple-QTL models (MQMs) for multiple related populations derived from the diallel of pairwise crosses among three inbred parents A, B, and C: A × B, A × C, and B × C. Three additive QTL allele substitution effects are estimable from populations segregating from these crosses, α_{AB}, α_{AC}, and α_{BC}, respectively. Jannink and Jansen (2001) showed that a reduction in the number of estimated parameters could be achieved if $\alpha_{AC} = \alpha_{AB} + \alpha_{BC}$. This parameter reduction defined REDUCED vs. FULL models. Differences in the likelihoods of these models provided evidence of epistatic interaction occurring between the locus analyzed and other loci in the genetic background (Jannink and Jansen, 2001). At the same time, assuming additive gene action, the REDUCED model detected QTL with higher power than the FULL model.

This method requires multiple crosses in a diallel structure, and its applicability is therefore restricted. In this paper we propose to broaden the applicability of reduced parameter models by focusing on short genome segments, determining the DNA-marker haplotype carried by each parent on such a segment, grouping parents that share common haplotypes, and formulating reduced models in terms of haplotype effects. We coin this new approach haploMQM. We apply the method to simulations with multiple related $F_{2:3}$ populations and highlight possible strengths and weaknesses of the new approach.

MATERIALS AND METHODS

Model Parametrization

Jannink and Jansen (2001) described MQMs and computational methods for populations derived from a simple diallel among three inbred parents A, B, and C. Here, we reparametrize the models of Jannink and Jansen (2001) to expand their models to arbitrary single-generation mating designs.We then show how these models apply when QTL alleles are not identified to specific parents but to putatively conserved DNA-marker haplotypes that may be carried by two or more parents. We subsequently propose an additional reduced model in which population means are no longer estimated independently but depend on the allelic values attributed to the haplotypes segregating within populations.

We now reparametrize these multipopulation models in terms of the effects of QTL alleles carried by inbred parents rather than in terms of allele substitution effects within each segregating population. We define two parents as *interconnected* if, in the system of inbred crosses studied, there exists a path of crosses joining them. For example, in the two-popula-

R.C. Jansen, Groningen Bioinformatics Centre, Inst. of Mathematics and Computing Sci., POB 800, NL-9700 AV, Groningen, The Netherlands; J.L. Jannink, Dep. of Agronomy, Iowa State Univ., Ames, IA 50011-1010, USA; W.D. Beavis, NCGR, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA. Received 17 Apr. 2002. *Corresponding author (r.c.jansen@cs.rug.nl).

Published in Crop Sci. 43:829–834 (2003).

Abbreviations: DH, doubled haploid; haploIM, haplotype interval mapping; haploMQM, haplotype multiple quantitative trait loci model; haploMQM⁻, reduced haplotype multiple quantitative trait loci model; IM, interval mapping; MQM, multiple quantitative trait loci model; QTL, quantitative trait loci.

tion system {$P_1 \times P_2$; $P_2 \times P_3$}, parents P_1 and P_3 are interconnected. In the system {$P_1 \times P_2$; $P_4 \times P_3$}, P_1 and P_3 are not interconnected. An interconnected system of populations, then, is one in which all parents are interconnected. An important property of such a system is that $N \geq P - 1$, where P is the number of inbred parents and N is the number of derived populations. The maximum number of estimable QTL effect parameters from a system of interconnected populations is N. A more parsimonious parametrization that is always possible for such a system is therefore to fix the effect of the allele carried by one arbitrary parent to zero and then to estimate the effects of the alleles carried by other parents as deviations from the fixed parent. From this point of view, the FULL model described above corresponds to a parametrization in which the interconnectedness of the diallel is ignored, and each cross is considered separately. The REDUCED model, on the other hand, recognizes the interconnectedness of the diallel between parents A, B and C, arbitrarily fixes the effect of the allele carried by one parent to zero, and estimates two further parameters: the effects of the alleles carried by the two remaining parents. From the perspective of parameter reduction between FULL and REDUCED models, we distinguish strongly and weakly interconnected systems of populations. In the former, $N >> P$, while in the later $N \approx P$. The diallel is a good example of strongly interconnected populations because it maximizes N relative to P, $N = P(P - 1)/2$, assuming no reciprocal crosses. Finally, among populations in a set available for analysis, not all will be interconnected. It will therefore be necessary to distinguish interconnected subsets and parameterize each independently.

In the development given above, we have identified QTL alleles according to the inbred parent from which they originate. Consider now two inbred parents that have retained the same genomic block from a common ancestor. Rather than identifying distinct QTL alleles for each parent, it would be more parsimonious to identify the single QTL allele carried by the common ancestor. To be able to benefit from this source of parsimony, we need to detect when inbred parents carry common genomic blocks. Assume that this task can be done so that, among a set of P parents, we now can identify H haplotypes ($H \leq P$) for a genomic block in a given region of the genome. Consequently, we are interested in the effects of the QTL alleles carried by the H haplotypes. In a process analogous to the one described above, we can determine interconnected sets of haplotypes as follows. If two haplotypes segregate in a population, they can said to be *crossed*. Two haplotypes are interconnected if there is a path of crossed haplotypes joining them. For haplotypes to be interconnected, it is sufficient that they be carried by interconnected parents. However, haplotypes can be interconnected even if they are not carried by interconnected parents. Consider the noninterconnected system of parents {$P_1 \times P_2$; $P_4 \times P_3$}. Assume that we have $P_1(H_1)$, $P_2(H_2)$, $P_4(H_2)$, and $P_3(H_3)$ where the haplotype carried by the parent is given in parentheses. Haplotypes H_1 and H_3 are then interconnected. Note that in evaluating haplotype interconnectedness, two new situations arise. First, if both crosses $P_1 \times P_2$ and $P_1 \times P_4$ were made, haplotypes H_1 and H_2 would be contrasted twice. Second, if the cross $P_2 \times P_4$ were made, haplotype H_2 would be contrasted to itself. Without loss of generality, haplotypes involved in such replicated or identity contrasts can be considered interconnected haplotypes.

Given these parametrizations, we distinguish the following models:

(i) Interval Mapping (IM)

$$y_{ij} = \mu_i + \alpha_i x_{ij} + e_{ij},$$

where y_{ij} is the phenotype for individual j in segregating population i, μ_i is the intercept for population i, α_i is effect of the QTL allele carried by an arbitrary parent of population i (the effect of the allele carried by the other parent is set to zero), and $e_{ij} \sim N(0, \sigma_e^2)$ is an error residual for individual j (the error variance, σ_e^2, is assumed equal across populations; it can also be taken unequal in cases of higher heritability). The independent variables x_{ij} depend on the QTL genotype carried by individuals ij. Table 1 shows parameterizations for various example populations. In practice, QTL genotypes remain unobserved, the x_{ij} are stochastic, and probabilities for the possible QTL configurations are calculated using flanking markers. For missing QTL or marker information, Jansen and Stam (1994) have shown that maximum likelihood estimates for the parameters μ_i, α_i, and e_{ij} can be obtained within each population by an expectation-maximization procedure using weighted regression.

(ii) Haplotype Interval Mapping (HaploIM)

The effect of the allele defined by one arbitrary haplotype is set to zero, the effects of the alleles of other haplotypes

Table 1. Parametrization of the family mean and quantitative trait loci (QTL) component of interval mapping (IM), haplotype interval mapping (haploIM), and reduced haplotype multiple-QTL models (haploMQM$^-$) as determined by population, parental haplotypes, and the genotype at a QTL.

Population†	QTL Genotype‡	IM	HaploIM§	HaploMQM$^-$
1. $P_1(H_1) \times P_2(H_2)$	**MM**	μ_1	μ_1	μ
	MP	$\mu_1 + \alpha_{12}$	$\mu_1 + \alpha_2$	$\mu + \alpha_2$
	PP	$\mu_1 + 2\alpha_{12}$	$\mu_1 + 2\alpha_2$	$\mu + 2\alpha_2$
2. $P_1(H_1) \times P_3(H_3)$	**MM**	μ_2	μ_2	μ
	MP	$\mu_2 + \alpha_{13}$	$\mu_2 + \alpha_3$	$\mu + \alpha_3$
	PP	$\mu_2 + 2\alpha_{13}$	$\mu_2 + 2\alpha_3$	$\mu + 2\alpha_3$
3. $P_2(H_2) \times P_3(H_3)$	**MM**	μ_3	$\mu_3 + 2\alpha_2$	$\mu + 2\alpha_2$
	MP	$\mu_3 + \alpha_{23}$	$\mu_3 + \alpha_2 + \alpha_3$	$\mu + \alpha_2 + \alpha_3$
	PP	$\mu_3 + 2\alpha_{23}$	$\mu_3 + 2\alpha_3$	$\mu + 2\alpha_3$
4. $P_5(H_1) \times P_6(H_2)$	**MM**	μ_4	μ_4	μ
	MP	$\mu_4 + \alpha_{56}$	$\mu_4 + \alpha_2$	$\mu + \alpha_2$
	PP	$\mu_4 + 2\alpha_{56}$	$\mu_4 + 2\alpha_2$	$\mu + 2\alpha_2$
5. $P_1(H_1) \times P_5(H_1)$	**MM**	μ_5	μ_5	μ
	MP	$\mu_5 + \alpha_{15}$	μ_5	μ
	PP	$\mu_5 + 2\alpha_{15}$	μ_5	μ
No. of parameters		**10**	**7**	**3**

† Maternal then paternal parent are given with the haplotype identity that they carry.
‡ Maternal (M) or paternal (P) derivation of the QTL is indicated.
§ The haploIM and haploMQM parametrizations arbitrarily fix the value of the QTL allele carried by H_1 to zero.

are estimated as deviations from the fixed one. Consider an individual in the segregating population obtained by crossing two parents, say P_1 and P_2. The model is

$$y_{ij} = \mu_i + \alpha_{h1(k)}x_{1ij} + \alpha_{h2(k)}x_{2ij} + e_{ij}.$$

The parameters are the same as in IM, save that allele effect $\alpha_{h1(k)}$ is the effect of the QTL allele defined by haplotype *h1* of P_1 in interconnected system *k,* and $\alpha_{h2(k)}$ is that of haplotype *h2* of P_2. Table 1 shows parametrizations for example populations derived from parents with identified haplotypes at a putative QTL locus; in this example $\alpha_{h1(k)}$ is set to zero, whereas $\alpha_{h2(k)}$ and $\alpha_{h3(k)}$ are free and written in short notation as α_2 and α_3. At the map location under study and within each population, probabilities for the three possible QTL configurations (MM, MP, PP) are calculated using flanking markers as in IM.

(iii) Haplotype Multiple-Quantitative-Trait-Loci Models (HaploMQM)

The exposition above is in terms of a single QTL only. Multiple-QTL models allow statistical control of genetic background noise due to QTL on other portions of the genome that are segregating in the population using multiple regression on marker cofactors (Jansen and Stam 1994; Jannink and Jansen 2001).

$$y_{ij} = \mu_i + \alpha_{h1(k)}x_{1ij} + \alpha_{h2(k)}x_{2ij} + \Sigma_c\alpha^c_{h1(kc)}x^c_{1ij} + \Sigma_c\alpha^c_{h2(kc)}x^c_{2ij} + e_{ij},$$

where the summation occurs over c marker cofactors. For each cofactor c, a set k_c of interconnected haplotypes have been identified. Note that inbred parents will have received genomic blocks from different ancestors in different portions of the genome. Consequently, the partitions of haplotypes into interconnected sets (k, k_1, k_2, and so on) will not be identical.

(iv) Reduced Haplotype Multiple-Quantitative-Trait-Loci Model (HaploMQM$^-$)

This model is identical to haploMQM, except that separate intercepts are no longer estimated for each population. Instead, differences in population means are assumed to derive from the different haplotypes segregating in each population. Thus, in this model, differences in population mean contribute to the estimate of haplotype QTL allele effects. Index i is dropped from the μ_i parameter in the models above and the model becomes:

$$y_{ij} = \mu + \alpha_{h1(k)}x_{1ij} + \alpha_{h2(k)}x_{2ij} + \Sigma_c\alpha^c_{h1(kc)}x^c_{1ij} + \Sigma_c\alpha^c_{h2(kc)}x^c_{2ij} + e_{ij}.$$

Implementation Issues

To fit a QTL at a certain map position under study, a *window* around this map position is defined. The different parental haplotypes in this window are identified. If two parents share the same haplotype, then we assume that they transmit the same QTL allele to their offspring. The window can be based either on a fixed map size, say 5 or 10 cM, or on a fixed number of markers, say four. In the latter case, the possible number of haplotypes can be very large (e.g., 2^4 for haplotypes of four biallelic markers and 5^4 for haplotypes of four five-allelic markers). However, if parents are derived from few ancestors, then the number of different haplotypes will be smaller than the number of parents, leading to the situation of strongly interconnected haplotypes that enables a large reduction in the number of estimated parameters. We here focus on such situations. To determine haplotypes, we use a window of four adjacent markers. Parents are grouped according to their haplotype and the allelic effects of the parental haplotypes are estimated from the data on the multiple populations using the models described above.

When analyzing multiple small populations, cofactor parametrization in MQMs can be problematic due to the large number of parameters involved (e.g., one cannot fit simultaneously 30 marker cofactors to a population with 10 offspring only; such simultaneous fitting would be possible in populations with 50 offspring, but if 60 such populations were analyzed, the computational burden would be excessive). Haplotype-based combining of cofactor parameters should reduce this problem substantially. Thus, haploIM can be extended to haploMQM by adding haplotype-based marker cofactors. For each marker cofactor, the local haplotypes form the basis for the reduction of parameters associated with that marker, in the same way as shown in Table 1. Since the clustering of parents is based on local haplotypes, clustering is likely to be different for each marker cofactor and different from the clustering at the focal QTL.

Procedure for Quantitative Trail Loci Analysis

We here briefly describe the haploMQM procedure (the approaches for haploIM and for haploMQM$^-$ are identical, except that in the former no cofactors are used and in the latter independent intercepts are not estimated for each population). Conceptually, the procedure for haploMQM is identical to the MQM procedure described in detail in Jansen (2001, p. 581–590). Three markers were chosen per chromosome to be candidate cofactors, leading to $3 \times 10 = 30$ candidate cofactors across the entire genome. Using all candidate cofactors, we calculated a bias-adjusted residual variance that was used for all further estimations on the population. We used backward elimination to retain in the model only those cofactors that explained a significant proportion of the variance. To determine whether to retain a cofactor in the backward elimination procedure, we used a threshold T such that $\text{Prob}(F_{df1,df2} > T) = 0.02$, with degrees of freedom df1 equal to the number of parameters of the cofactor and df2 equal to the residual degrees of freedom in the all-cofactor model. To locate a QTL, we then scanned the full genome in 5-cM steps. We first calculated the likelihood of the data in the presence of a QTL but without parameter reduction due to haplotyping (L_{Full}), and with parameter reduction (L_{Haplo}). Procedures without haplotyping used the likelihood ratio to quantify QTL likelihood:

$$\text{LR}_{\text{QTL}} = 2\log(L_{\text{Full}}/L_{\text{noQTL}}).$$

Procedures with haplotyping used the likelihood ratio:

$$\text{LR}_{\text{QTL}} = 2\log(L_{\text{Haplo}}/L_{\text{noQTL}}) \text{ and}$$
$$\text{LR}_{\text{QTL}} = 2\log(L_{\text{Haplo}^-}/L_{\text{noQTL}^-}).$$

Large values of the LR_{QTL} statistic indicate support for the presence of a QTL using the haplotype model. Note that the no-QTL likelihood is different in the haploMQM and the haploMQM$^-$ cases. In particular, since haploMQM$^-$ models the population means using QTL and cofactors, the likelihood L_{noQTL^-} includes no contribution of the putative QTL to the modeling of population means. In general, then, L_{noQTL^-} can be quite a bit smaller than L_{noQTL}.

To determine genome-wide significance thresholds for these statistics, one can perform simulation runs on populations generated without genetic variance or permutation. In the present discussion, an ad hoc approach was utilized for illustrative purposes just by taking a much more stringent

threshold per test: (approximate) thresholds for QTL detection at $\alpha = 0.001$ per test: $\chi^2(60; 0.001) \approx 102$ for IM, $\chi^2(15; 0.001) \approx 38$ and $\chi^2(37; 0.001) \approx 69$ for haploIM, and haploMQM can be used in Simulations 1 and 2. The degrees of freedom used to determine the χ^2 threshold derive from the number of parameters estimated by each model. The IM procedure estimates 60 QTL parameters, one for each family; in Simulation 1, an average of 15 haplotypes were distinguished for any given locus so that 15 allelic effects needed to be estimated; in Simulation 2, the average number of haplotypes distinguished was 37. In our simulations the number of QTL parameters varied little across the genome. In other simulations, where the degrees of freedom for the QTL (df_{QTL}) may vary more notably, it can be better to divide QTL likelihoods by the local df_{QTL} for graphical display and use local thresholds $\chi^2(df_{QTL}; 0.001)$ or determine such thresholds by simulation or permutation.

Two further likelihood ratios could be of interest:

$$LR_{Haplo} = 2 \log(L_{Full}/L_{Haplo}).$$

Large values of the LR_{Haplo} statistic indicate problems with the haplotype parametrization that could be due to failure of marker haplotypes to correctly group parents in terms of the QTL allele that they carry or due to strong QTL × genetic background interaction such that the same QTL allele would have divergent effects in the different families in which it segregates. One can also calculate

$$LR_{Means} = 2 \log(L_{Haplo}/L_{Haplo^-}).$$

Large values of LR_{Means} indicate failure of the QTL and selected cofactors to predict differences among population means. Such a failure could be due to the presence of QTL undetected by the mapping procedure or due to epistatic effects on family means not accounted for by the models presented. In the calculation of LR_{QTL}, the fact that L_{Haplo^-} is smaller than L_{Haplo} is compensated for by the fact that L_{noQTL^-} is smaller than L_{noQTL}.

Simulating Early Generation Progeny Tests in Maize Breeding

Early generation progeny tests in maize breeding are often referred to as first- and second-year topcross tests. We simulated marker and trait data for 60 related $F_{2:3}$ populations of size 10 each, with one QTL on each of chromosomes 1 to 5. Note that a basic property of the testcross design is that it eliminates dominance as a source of genetic variance. To mimic the genome of maize, the genome in our simulation consisted of 10 linkage groups, each containing 101 biallelic marker loci with 2-cM map distance between adjacent pairs (using Haldane's mapping function). The genotype and phenotype data were generated in a number of steps as follows.

Step 1: Generate Base Population of Candidate Parents

The following protocol was used for generating a base population of inbred lines with different (re)combinations of *ancestral* linkage blocks. First we crossed a hypothetical inbred parent homozygous for all markers (say, 11111 and so on) with a parent also homozygous for all markers but carrying different alleles (say, 22222 and so on), and generated a set of 400 doubled haploid (DH) lines. For Simulation 2, we used a higher recombination frequency of 0.2 between adjacent markers (instead of 0.02) to mimic more generations of recombination. The resulting genotype of a DH line consisted of linkage blocks of 1s and of 2s of different size, for example, 111211222, and so on. Next, we generated multiallelic marker genotypes (five alleles per marker) from these linkage blocks. This was accomplished by assigning marker alleles to the linkage blocks, linkage block by linkage block, from a multinomial distribution with frequencies of marker alleles: 0.55, 0.24, 0.12, 0.06, and 0.03, respectively. For example, the original genotype 111221 contains three linkage blocks, 111, 22, and 1. A marker allele is randomly assigned to each linkage block. Thus, after sampling marker alleles for linkage blocks, the original genotype 111221 could become 444113. The multiallelic state of marker loci is similar to the polymorphism index that has been observed in simple sequence repeat markers in maize (Senior et al., 1996). Finally, a QTL allele was placed in the middle of linkage groups 1 to 5. If the linkage block surrounding the middle of the linkage group carried marker allele 1, then a favorable allele (denoted +) was placed there, else an unfavorable allele (denoted −) was placed there. The phenotypic effect of a QTL was such that each QTL was expected to contribute 15% to the total phenotypic variation in a population where all five QTLs were segregating (heritability of 75%; populations where less than five QTLs segregate will have lower heritability).

Step 2: Select Sixty Parent Pairs

Pairs of parents for multiple crosses were selected from the base population. The 400 lines in the base population can be crossed amongst each other in various combinations. In Simulation 1, pairs of parents were selected in such a way that the two parents of a cross were ≈10% related, that is, ≈90% of the 1010 marker loci are polymorphic in each population. In Simulation 2, pairs of parents were selected so that the two parents of a cross were ≈45% related.

Step 3: Generate Sixty $F_{2:3}$ Populations

Parents were crossed to generate offspring. The marker profile and genetic value of each progeny was determined based on the markers and QTL segregating between its parents, using the laws of Mendelian segregation and recombination. Each population consisted of 10 offspring in Simulation 1, and 50 offspring in Simulation 2.

Step 4: Randomly Sample Two Hundred Markers Available for Analysis

As a last step in the simulation procedure, the set of 1010 markers was reduced: in each of the simulations, 200 loci were randomly sampled from the genome and only these marker data were available for analysis. This resembles currently available marker density.

RESULTS

Analyzing Simulated Early Generation Progeny Tests in Maize Breeding

Analysis of Chromosomes 6 to 10

In the haploIM and haploMQM, the number of parameters per QTL (or marker cofactor) is equal to the number of different parental haplotypes in the window under study. The haplotype-based models required, on average, 15 and 37 parameters in Simulations 1 and 2, respectively, with little variability. In contrast, IM takes 60 allele-substitution parameters per QTL (as many as there are populations). Under the null hypothesis, the likelihood-ratio test statistic at a fixed map position follows approximately a chi-squared (or *F*-like) distribu-

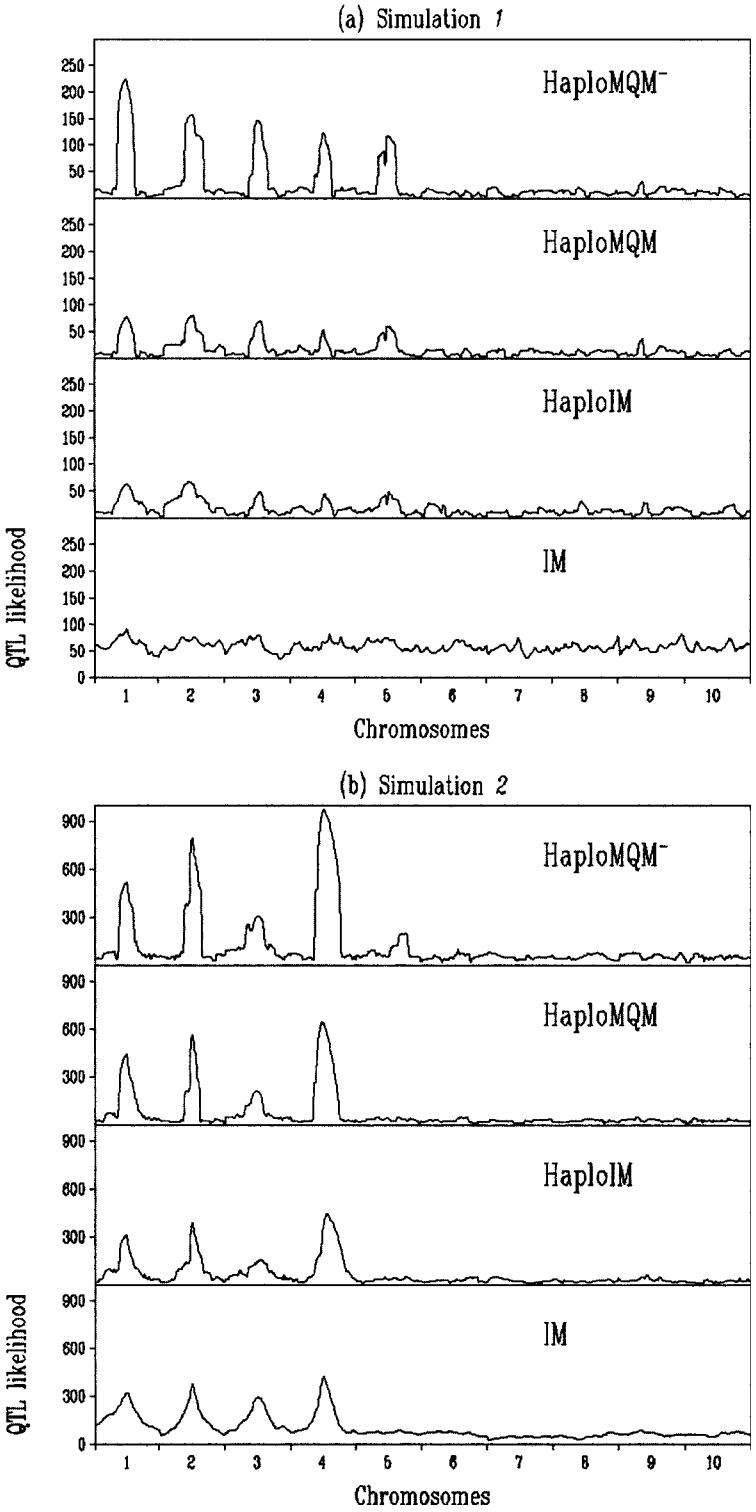

Fig. 1. Results from two simulation studies of 60 related plant breeding $F_{2:3}$ populations and 200 random markers on a 2000-cM genome and with unlinked quantitative trait loci (QTLs) on each of chromosomes 1 to 5. (a) Parent pairs of the crosses were 10% related, but originated from a base population with medium linkage disequilibrium level. Each population consisted of 10 offspring. (b) Parent pairs of the crosses were 45% related, but originated from a base population with low linkage disequilibrium level. Each population consisted of 50 offspring; see main text for a complete description of the simulated configurations and for the appropriate definitions of QTL likelihood. IM, interval mapping; haploIM, haplotype interval mapping; haploMQM, haplotype multiple quantitative trait loci model; haploMQM^{-}, reduced haplotype multiple quantitative trait loci model.

tion (Jansen, 1994) with degrees of freedom equal to the number of parameters involved in the test and the expected value of this test is equal to the degrees of freedom. On chromosomes 6 to 10, where there are actually no QTLs present, the expectation for LR_{QTL} is 60 for procedures without haplotyping (IM in Fig. 1) and 15 and 37 for procedures with haplotyping (haploIM, haploMQM, and haploMQM$^-$ in Fig. 1). These expectations were in fact observed (Fig. 1). An important consequence is that the threshold for genome-wide significance in haploIM and haploMQM is much lower than that in multipopulation IM.

Analysis of Chromosomes 1-5

Consistent with the analysis of chromosomes 6 to 10, the *valleys* out of which QTL likelihood ratio peaks emerge are lower in the haploIM case than the IM case (Fig. 1a). For the IM procedure, background likelihood ratio levels are so high that the peaks can scarcely be distinguished. The lack of distinct peaks occurs despite the fact that peaks using IM are generally slightly higher than using haploIM (Fig. 1a). The reason they are higher is because haplotype identity does not guarantee QTL identity. When haplotyping misidentifies a QTL, the resultant likelihood for the model will be lower. That decreased likelihood could be diagnosed using the LR_{Haplo} ratio. Under the simulation conditions presented, the LR_{Haplo} statistic never reached very high levels, indicating that identical haplotypes generally correctly identified common QTL alleles. Fig. 1 shows that the lowered threshold of haploIM over IM potentially increases the power of QTL detection using the haplotype-based methods. Fig. 1 also shows that the QTL-likelihood peaks can be higher for haploMQM than for IM. Modeling of multiple QTLs can be a second step to increase the power of QTL detection, in particular in situations of higher heritability as in our simulation. The haploMQM$^-$ approach provides a third possibility for further increasing the QTL power: the use of between-population information in addition to within-population information. Populations are usually not segregating for all QTLs involved. But that does not imply that the effect of those nonsegregating QTLs cannot be detected: they generate differences between the mean values of the populations. Therefore, the QTL-allele effects in our models should not only explain the within-population variation but they should also capture the between-population variation. Fig. 1b shows that the difference for QTL-likelihood between haploMQM and haploMQM$^-$ can be substantial. In other words, between-population information can be used to increase QTL-likelihood peaks. In fact, in Simulation 2 the QTL on chromosome 5 only becomes visible in the haploMQM$^-$ analysis.

DISCUSSION

In our approach we define a window of fixed number of markers (four) around the QTL position under study. Of course, a larger number of markers can be used, in particular if the marker map is dense. In our approach it is assumed that two parents with identical haplotype in the window under study are identical by descent and share the same QTL allele in this region. The probability that this is indeed true increases when the haplotype is based on more markers. However, with more markers the window can become large. In the extreme case, all markers are used simultaneously for haplotyping, and a 1:1 relation between haplotype and parent is established, in which case no reduction of QTL parameters is achieved. In general, there are also good reasons for using a window with few markers. Such a window tends to result in fewer haplotype classes, so that fewer QTL parameters are required; preferably much less than the number of populations to gain power over conventional IM. It will be expected that there is an optimal balance between the pros and cons of using more or less markers for haplotyping, and it is likely that the optimum can change, for example, when different marker densities are used, or when different types of marker are used. In breeding populations, biallelic markers (e.g., AFLP markers) are expected to be less informative in fingerprinting than multiallelic markers (e.g., microsatellite markers). Preferably, less informative marker types are available at a higher map density to achieve indirectly a high multilocus information content. Finally, the more generations passed by since the founding ancestral lines, the smaller the haplotypes shared and the denser the marker maps should be to tag ancestral blocks.

ACKNOWLEDGMENTS

Pioneer Hi-Bred supported this research in a collaborative project with Plant Research International B.V.

REFERENCES

Beavis, W.D. 1998. QTL analyses: Power, precision and accuracy. p. 145–161. *In* A.H. Paterson (ed.) Molecular analysis of complex traits. CRC Press, Boca Raton, FL.

Jannink, J.L., M.C.A.M. Bink, and R.C. Jansen. 2001. Using complex plant pedigrees to map valuable genes. Trends Plant Sci. 6(8): 337–342.

Jannink, J.L., and R.C. Jansen. 2001. Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157: 445–454.

Jansen, R.C. 1994. Controlling the Type 1 and Type 2 errors in mapping quantitative trait loci. Genetics 138:871–881.

Jansen, R.C. 2001. Quantitative trait loci in inbred lines. p. 567–597. *In* D.J. Balding et al. (ed.) Handbook of statistical genetics. Wiley, New York.

Jansen, R.C., and P. Stam. 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455.

Meuwissen, T.H.E., and M.E. Goddard. 2000. Fine mapping of quantitative trait loci using linkage disequilibrium with closely linked marker loci. Genetics 155:421–430.

Senior, M.L., E.C.L. Chin, M. Lee, J.S.C. Smith, and C.W. Stuber. 1996. Simple sequence repeat markers developed from maize sequences found in the genbank database: Map construction. Crop Sci. 36:1676–1683.

Yi, N., and S. Xu. 2001. Bayesian mapping of quantitative trait loci under complicated mating designs. Genetics 157:1759–1771.